Στατιστικές και Νευρωνικές Μέθοδοι για Προβλήματα Μηχανικής Μάθησης

Η ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Πληροφορικής Εξεταστική Επιτροπή

από τον

Κωνσταντίνο Κωνσταντινόπουλο

ως μέρος των υποχρεώσεων για τη λήψη του

ΔΙΔΑΚΤΟΡΙΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

Ιούνιος 2006

Η εργασία αυτή συγχρηματοδοτήθηκε από την Ευρωπαϊκή Ένωση, στα πλαίσια του προγράμματος με τίτλο "ΗΡΑΚΛΕΙΤΟΣ", του προγράμματος ΕΠΕΑΕΚ ΙΙ του 3ου Κοινοτικού Πλαισίου Στήριξης του ΥΠ.Ε.Π.Θ., με χρηματοδότηση κατά 25% από εθνικούς πόρους και κατά 75% από το Ευρωπαϊκό Κοινωνικό Ταμείο (ΕΚΤ).

This research was co-funded by the European Union in the framework of the program "HRAKLEITOS" of the "Operational Program for Education and Initial Vocational Training" of the 3rd Community Support Framework of the Hellenic Ministry of Education, funded by 25% from national sources and by 75% from the European Social Fund (ESF).

ΥΠΟΥΡΓΕΙΟ ΕΘΝΙΚΗΣ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΕΑΕΚ

ΕΥΡΩΠΑΪΚΗ ΕΝΩΣΗ
ΣΥΓΧΡΗΜΑΤΟΔΟΤΗΣΗ
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Η ΠΑΙΔΕΙΑ ΣΤΗΝ ΚΟΡΥΦΗ
Επιχειρησιακό Πρόγραμμα
Εκπαίδευσης και Αρχικής
Επαγγελματικής Κατάρτισης

ΕΥΧΑΡΙΣΤΙΕΣ

Η διατριβή αυτή εκπονήθηκε στο τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων, με επιβλέποντα τον Αναπληρωτή Καθηγητή κ. Αριστείδη Λύκα, και μέλη της τριμελούς συμβουλευτικής επιτροπής τους κ.κ. Ισαάκ Λαγαρή Καθηγητή του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων και Δημήτρη Ι. Φωτιάδη Αναπληρωτή Καθηγητή του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων.

Θα ήθελα να εκφράσω τις θερμές και ειλικρινείς μου ευχαριστίες στον κ. Αριστείδη Λύκα, επιβλέποντα της διατριβής μου και δάσκαλό μου στα χρόνια των μεταπτυχιακών μου σπουδών, για την διορατική επιλογή του θέματος, την αδιάκοπη και εμπνευσμένη επιστημονική του καθοδήγηση και την άοκνη φροντίδα του σε όλα τα στάδια της διατριβής. Τον ευχαριστώ ακόμα για τις ενέργειές του που εξασφάλισαν την χρηματοδότηση των σπουδών μου. Ιδιαίτερα τον ευχαριστώ για την εμπιστοσύνη που μου έδειξε, την ηθική του υποστήριξη και την υπομονή του.

Θα ήθελα ακόμα να ευχαριστήσω για την άριστη ερευνητική συνεργασία τους κ.κ. Νικόλαο Γαλατσάνο, Καθηγητή του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων, Κωνσταντίνο Μπλέκα, Λέκτορα του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων, Μιχάλη Τίτσια, Διδάκτορα του School of Informatics, University of Edinburgh και Δημήτρη Τζίκα, υποψήφιο διδάκτορα του Τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων.

Τέλος θα ήθελα να ευχαριστήσω μέσα από την καρδιά μου τους γονείς μου Γιώργο και Πηνελόπη και τα αδέλφια μου Δημήτρη και Ελένη για την ανεκτίμητη αγάπη τους, την κατανόηση και την συμπαράσταση που μου πρόσφεραν. Ως ενθύμιο τόσων χρόνων απουσίας μου τους αφιερώνω αυτή την διατριβή.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

ΠΕΡΙΛΗΨΗ

Κωνσταντίνος Κωνσταντινόπουλος του Γεωργίου και της Πηνελόπης.
PhD, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων. Ιούνιος, 2006.
Τίτλος: Στατιστικές και Νευρωνικές Μέθοδοι για Προβλήματα Μηχανικής Μάθησης.
Επιβλέποντας: Αριστείδης Λύκας.

Στην παρούσα διατριβή μελετάμε μέσα στα πλαίσια της Μηχανικής Μάθησης προβλήματα που παρουσιάζονται στην ταξινόμηση και την ομαδοποίηση δεδομένων, με ιδιαίτερη έμφαση στην επιλογή της πολυπλοκότητας του μοντέλου και την επιλογή των χαρακτηριστικών των δεδομένων. Αντιμετωπίζουμε τα προβλήματα από την οπτική της εκτίμησης κατανομών, χρησιμοποιώντας την μίξη κανονικών κατανομών και το πιθανοτικό δίκτυο ακτινικών συναρτήσεων βάσης (PRBF). Οι αλγόριθμοι που προτείνουμε απορρέουν από τον αλγόριθμο Expectation-Maximization (EM) και από την variational Bayesian (VB) μεθοδολογία. Ιδιαίτερη προσπάθεια έγινε για να προτείνουμε αυξητικούς αλγόριθμους εκπαίδευσης, οι οποίοι αυξάνουν σταδιακά την πολυπλοκότητα του εκάστοτε μοντέλου.

Το πρώτο πρόβλημα που εξετάζουμε είναι αυτό της ομαδοποίησης. Το μοντέλο που χρησιμοποιούμε είναι μια μίξη κανονικών κατανομών, με την οποία εκτιμούμε την κατανομή του συνόλου δεδομένων, και στην συνέχεια ορίζουμε τις ομάδες όπως υποδεικνύονται από τις συνιστώσες της μίξης. Ταυτόχρονα δίνουμε και μια απάντηση στην κρίσιμη επιλογή του αριθμού των συνιστωσών της μίξης. Προτείνουμε μια αυξητική μέθοδο για την επιλογή μοντέλου και την εκπαίδευση μιας μίξης, που βασίζεται στη variational Bayesian μέθοδο. Η μέθοδος προσθέτει συνιστώσες στην μίξη χρησιμοποιώντας έναν έλεγχο Bayesian διάσπασης: μια συνιστώσα διασπάται σε δύο νέες συνιστώσες και στη συνέχεια εφαρμόζονται variational Bayesian εξισώσεις ενημέρωσης μόνο στις παραμέτρους των δυο νέων συνιστω-

σών. Ως αποτέλεσμα, είτε και οι δύο συνιστώσες διατηρούνται στη μίξη, είτε μια από της δύο αποδεικνύεται περιττή και απαλείφεται από την μίξη. Στην προσέγγισή μας το πρόβλημα επιλογής μοντέλου αντιμετωπίζεται τοπικά σε μια περιοχή του χώρου δεδομένων, έτσι μπορούμε να θέσουμε πιο κατατοπιστικές εκ των προτέρων πιθανότητες βασιζόμενοι στην τοπική κατανομή των δεδομένων. Για την υλοποίηση αυτής της προσέγγισης παρουσιάζουμε μια βελτιωμένη Bayesian μίξη, καθώς και έναν αλγόριθμο εκπαίδευσης που εφαρμόζει επαναληπτικά έναν έλεγχο διάσπασης σε κάθε συνιστώσα της μίξης.

Το επόμενο πρόβλημα που εξετάζουμε είναι πάλι το πρόβλημα της ομαδοποίησης και επιλογής χαρακτηριστικών. Για την επιλογή του αριθμού των συνιστωσών της μίξης εκμεταλλευόμαστε την τυπική variational Bayesian μέθοδο, και επικεντρωνόμαστε στο πρόβλημα της επιλογής χαρακτηριστικών. Αναζητούμε χαρακτηριστικά τα οποία αναδεικνύουν τις διαφορές μεταξύ των ομάδων, και προτείνουμε μια variational Bayesian μέθοδο για την εκπαίδευση μιας μίξης που αντιμετωπίζει ταυτόχρονα το πρόβλημα της επιλογής χαρακτηριστικών και της επιλογής μοντέλου. Η επιλογή χαρακτηριστικών στηρίζεται σε μια Bayesian μίξη δύο επιπέδων, που για κάθε συνιστώσα της τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, και με μια πιθανότητα είναι και υπό συνθήκη ανεξάρτητα από την συνιστώσα.

Το τρίτο πρόβλημα που εξετάζουμε είναι το πρόβλημα της κατάτμησης μιας εικόνας, ως πρόβλημα ομαδοποίησης των εικονοστοιχείων της. Αφού εξάγουμε κατάλληλα χαρακτηριστικά από τα εικονοστοιχεία, η ομαδοποίησή τους μπορεί να γίνει με μια μίξη κανονικών κατανομών. Προτείνουμε την σταδιακή ομαδοποίηση των εικονοστοιχείων, και παρουσιάζουμε μια μέθοδο ενεργητικής μάθησης που αυξάνει σταδιακά το πλήθος των εικονοστοιχείων που χρησιμοποιούνται για την επαναλαμβανόμενη κατάτμηση της εικόνας. Σε κάθε επανάληψη της μεθόδου έχουμε ενσωματώσει έναν αλγόριθμο εκπαίδευσης της μίξης που βασίζεται στην variational Bayesian μέθοδο, και αντιμετωπίζει αποδοτικά το πρόβλημα επιλογής μοντέλου. Τα πειραματικά αποτελέσματα έδειξαν ότι η μέθοδος έχει ικανοποιητικό χρόνο εκτέλεσης και προσαρμόζει με επιτυχία των αριθμό των περιοχών που χρησιμοποιεί στην πολυπλοκότητα της εικόνας.

Το τέταρτο πρόβλημα που εξετάζουμε είναι το πρόβλημα της ταξινόμηση με το δίκτυο PRBF. Η τυπική μέθοδος εκπαίδευσης του PRBF για ταξινόμηση χρησιμοποιεί τον αλγόριθμο EM, και το τελικό δίκτυο εξαρτάται από την αρχικοποίηση των παραμέτρων του.

Προτείνουμε μια μέθοδο για την αυξητική εκπαίδευση του δικτύου PRBF, η οποία ξεκινά με μια μόνο συνιστώσα, και σταδιακά προσθέτει περισσότερες σε κατάλληλες θέσεις στον χώρο δεδομένων. Η προσθήκη μιας επιπλέον συνιστώσας βασίζεται σε ένα κριτήριο για τον εντοπισμό μιας περιοχής που είναι κρίσιμη για την ταξινόμηση, συγκεκριμένα στο όριο απόφασης. Μετά από την προσθήκη όλων των συνιστωσών, η μέθοδος διασπά κάθε συνιστώσα σε υπο-συνιστώσες που κάθε μια αντιστοιχεί σε διαφορετική κατηγορία.

Το τελευταίο πρόβλημα που μας απασχολεί είναι το πρόβλημα της ενεργητικής μάθησης PRBF για ταξινόμηση. Μια ενδιαφέρουσα παραλλαγή του προκύπτει αν το σύνολο εκπαίδευσης περιέχει ταξινομημένα και αταξινόμητα πρότυπα, τότε ο αλγόριθμος μπορεί να επιλέξει και να ρωτήσει την κατηγορία κάποιου από τα αταξινόμητα. Παρουσιάζουμε μια αυξητική μέθοδο για εκπαίδευση με ημι-επίβλεψη χρησιμοποιώντας τα ταξινομημένα και τα αταξινόμητα δεδομένα, που βασίζεται στον αλγόριθμο EM. Στην συνέχεια προτείνουμε μια μέθοδο ενεργητικής μάθησης που επαναληπτικά εφαρμόζει την διαδικασία μάθησης με ημι-επίβλεψη, και στην συνέχεια επιλέγει ένα αταξινόμητο δεδομένο και ζητά να μάθει την κατηγορία του. Αφού προστεθεί στο σύνολο εκπαίδευσης η κατηγορία αυτού του προτύπου, συνεχίζεται η αυξητική εκπαίδευση του δικτύου. Το κριτήριο επιλογής που προτείνουμε επιλέγει σημεία κοντά στο όριο απόφασης του τρέχοντος ταξινομητή, και διευκολύνει την αυξητική μέθοδο με ημι-επίβλεψη που επίσης εκμεταλλεύεται το όριο απόφασης.

Extended Abstract in English

Constantinos Constantinopoulos, G. P.
PhD, Computer Science Department, University of Ioannina, Greece. June, 2006.
Title: Statistical and Neural Methods for Machine Learning Problems.
Supervisor: Aristidis Likas.

This dissertation deals with Machine Learning, and more specifically with the most important issues that arise during classification and clustering, namely model and feature selection. We also examined certain aspects of active learning for classification and image segmentation. We treated learning as a density estimation problem, and exploited the Gaussian Mixture model (GMM) and the Probabilistic Radial Basis Function network (PRBF). The proposed algorithms are based on the Expectation-Maximization (EM) algorithm and the variational Bayesian (VB) approach. We concentrated our efforts on incremental training algorithms, which successively increase the complexity of the model that we train.

The GMM is a convex linear combination of Gaussian probability density functions, and can approximate arbitrary probability density functions in clustering and classification problems. The PRBF network constitutes a special case of the well known RBF network, and can estimate class conditional densities in classification problems. Moreover it extends the typical GMM approach to classification by allowing the sharing of mixture components among all classes. The proposed training algorithms for PRBF are based on the EM. The EM is an iterative method that estimates the parameters of a model maximizing the likelihood of a data set. The proposed training algorithms for GMM are based on the VB approach. The VB is an approximate Bayesian framework that yields EM-like algorithms.

Chapter 1 constitutes the introduction of this dissertation, and in Chapter 2 we present basic concepts of Machine Learning that we use throughout the dissertation. Namely the Maximum Likelihood (ML) criterion, the GMM and PRBF models, and the EM and VB methods for training these models.

In Chapter 3 we use a GMM for clustering data, and we try to estimate the true number of clusters simultaneously. We propose an incremental method that derives from VB. This method adds a new component using a Bayesian split test: a component is splitted in two, and VB update equations are applied to estimate the parameters of the new components. As a result, either both of them are retained, or one of them happens to be redundant and is rejected. In our approach we can attack model selection locally, in the space that the new components compete. So we can set more informative priors based on the local density of data. We present an appropriate Bayesian mixture, and a training algorithm that iteratively applies the split test on each component. Experimental results demonstrated that the proposed method effectively clusters artificial and real data.

In Chapter 4 we present a VB method for GMM training that simultaneously treats the feature selection and the model selection problem. The method is based on the integration of a mixture model formulation that takes into account the saliency of the features and a Bayesian approach to mixture learning that can be used to estimate the number of mixture components. The proposed learning algorithm follows the VB framework and can simultaneously optimize over the number of components, the saliency of the features and the parameters of the mixture model. Experimental results using high-dimensional artificial and real data illustrate the effectiveness of the method.

In Chapter 5 we present a method for image segmentation that uses a GMM to cluster the pixels of the image. Our method offers a solution to the problem of selecting the segments, and tackles the problem of high volume of data for an image with standard resolution (e.g. 256×256). We propose an iterative clustering of the pixels, using an active learning method that increases gradually the number of pixels used for segmentation. The method selects pixels that exhibit large square distances between the original and the segmented image. In each iteration we apply the VB method described in Chapter 3 to train the model and simultaneously estimate the number of clusters. Experimental

results demonstrated that the method can be fast, and successfully adapt the number of segments to the complexity of the target image.

In Chapter 6 we present an incremental algorithm for PRBF training. The typical learning method of PRBF in a classification context employs the EM algorithm, and depends strongly on the initial parameter values of the model. The proposed incremental algorithm starts with a single component, and incrementally adds more components at appropriate positions in the data space. The addition of the new component is based on a criterion that detects the region in the data space that is crucial for the classification task, namely the classification boundary. This stage of the algorithm concludes when a maximum number of components have been added. In the following stage, the algorithm splits every component of the network into sub-components, where each one corresponds to a different class. Experimental results using several well-known classification data sets indicate that the proposed incremental methodology provides superior solutions as compared to standard hierarchical PRBF training. Comparative experiments with Support Vector Machines (SVM) were also conducted, and the obtained results along with a qualitative comparison of the two approaches is presented.

In Chapter 7 we present an active learning methodology for training the PRBF network. We propose an incremental method for semi-supervised learning of labeled and unlabeled data simultaneously, based on the algorithm proposed in Chapter 6. We then present an active learning method that iteratively applies semi-supervised learning, and then employs a suitable criterion to select an unlabeled observation and query its label. The new label is added in the training set, and the incremental semi-supervised learning continues. The proposed criterion selects points near the decision boundary, and facilitates the incremental semi-supervised learning that also exploits the decision boundary. The proposed algorithm were experimentally tested using well-known data sets, and the results were promising.

This dissertation concludes in Chapter 8 with a short review, and some interesting issues for future work.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η μηχανική μάθηση είναι ο τομέας που ασχολείται με την δημιουργία προγραμμάτων υπολογιστών που μπορούν να βελτιώνουν την συμπεριφορά τους καθώς αποκτούν εμπειρία. Η εμπειρία αυτή μπορεί να έχει την μορφή παραδειγμάτων που έχουν ήδη συλλεχθεί ή συλλέγονται κατά την διάρκεια της μάθησης. Η ανάγκη για την ανάπτυξη της μηχανικής μάθησης προκύπτει από την αδυναμία μας να δώσουμε απευθείας λύση σε κάποια προβλήματα. Τέτοια είναι τα προβλήματα για τα οποία δεν υπάρχουν εμπειρογνώμονες για να περιγράψουν μια λύση ή η αποδοτική περιγραφή της είναι αδύνατη, όπως για παράδειγμα ο μηχανισμός για την αναγνώριση χειρόγραφων χαρακτήρων ή φωνημάτων ομιλίας. Την ίδια δυσκολία έχουμε όταν πρέπει να δώσουμε λύσεις που να προσαρμόζονται σε ένα περιβάλλον που αλλάζει δυναμικά, όπως για παράδειγμα στην περίπτωση ενός ρομπότ που ρυθμίζει την συμπεριφορά του ανάλογα με τις συνήθειες ή την ψυχολογική διάθεση του ανθρώπου που έχει μπροστά του. Σε τέτοιες περιπτώσεις καταφεύγουμε στη σταδιακή ανάπτυξη μιας λύσης χρησιμοποιώντας διαθέσιμα παραδείγματα ή συλλέγοντας καινούρια. Το σύνολο αυτών των παραδειγμάτων ονομάζεται *σύνολο εκπαίδευσης*, και το χρησιμοποιούμε για να αποτιμήσουμε την επίδοση της λύσης μας βάση ενός *κριτηρίου μάθησης* και στη συνέχεια να τη διαμορφώσουμε ανάλογα χρησιμοποιώντας έναν *αλγόριθμο μάθησης*.

1.1 Προβλήματα Μηχανικής Μάθησης

Τα προβλήματα που αντιμετωπίζουμε στην περιοχή της μηχανικής μάθησης μπορούμε να τα διακρίνουμε σε τρεις μεγάλες κατηγορίες ανάλογα με το τι είδους ανατροφοδότηση για την επίδοσή μας έχουμε, δηλαδή πώς διαπιστώνουμε αν το σύστημά μας λειτουργεί σωστά. Συγκεκριμένα μπορούμε να διακρίνουμε την κατηγορία της *μάθησης με επίβλεψη* (supervised learning), την κατηγορία της *μάθησης χωρίς επίβλεψη* (unsupervised learning) και την κατηγορία της *ενισχυτικής μάθησης* (reinforcement learning) [60, 2]. Στη μάθηση με επίβλεψη κάθε παράδειγμα στο σύνολο εκπαίδευσης αποτελείται από ένα ζεύγος εισόδου και επιθυμητής εξόδου, έτσι μπορούμε άμεσα να διαπιστώσουμε αν το σύστημα αποδίδει σωστά για δεδομένη είσοδο. Στην μάθηση με επίβλεψη εμπίπτουν τα προβλήματα της *ταξινόμησης* (classification) και της *παλινδρόμησης* (regression). Η μάθηση χωρίς επίβλεψη περιλαμβάνει μεταξύ άλλων τα προβλήματα της *εκτίμησης κατανομής* (density estimation) και της *ομαδοποίησης* (clustering). Σε αυτά δεν γνωρίζουμε την σωστή απόφαση, ωστόσο η απόφαση που παίρνουμε αφορά το σύνολο των παραδειγμάτων όχι κάθε παράδειγμα ξεχωριστά, και εκμεταλλευόμαστε των πλεονασμό (redudancy) πληροφορίας. Τέλος στα προβλήματα ενισχυτικής μάθησης παίρνουμε μια επιβράβευση (ή τιμωρία) μετά από μια σειρά αποφάσεων, η οποία δεν αναφέρεται σε κάθε απόφαση χωριστά αλλά στο κατά πόσο η αλληλουχία τους μας οδήγησε στο στόχο.

Στα προβλήματα ταξινόμησης στόχος μας είναι να διακρίνουμε την κατηγορία στην οποία ανήκουν οι παρατηρήσεις μας. Έχουμε για παράδειγμα φωτογραφίες από τα ψηφία (0–9) που εμφανίζονται στον ταχυδρομικό κωδικό ενός φακέλου και θέλουμε να αναγνωρίσουμε τους αντίστοιχους αριθμούς. Γενικά κάθε διαθέσιμο παράδειγμα στο σύνολο εκπαίδευσης αποτελείται από ένα ζεύγος (πρότυπο, κατηγορία), όπου το *πρότυπο* αποτελεί τα παρατηρούμενα χαρακτηριστικά της οντότητας που θέλουμε να ταξινομήσουμε. Στόχος μας είναι να εκμεταλλευτούμε το σύνολο εκπαίδευσης, και να δημιουργήσουμε ένα σύστημα που θα μπορεί να ταξινομεί σωστά νέα πρότυπα που δεν υπάρχουν στο σύνολο εκπαίδευσης. Η ικανότητα της σωστής ταξινόμησης καινούριων προτύπων ονομάζεται ικανότητα *γενίκευσης*. Για να κατασκευάσουμε έναν ταξινομητή μπορούμε να ακολουθήσουμε δύο προσεγγίσεις. Η πρώτη είναι να προσπαθήσουμε να μάθουμε την συνάρτηση που απεικονίζει ένα πρότυπο σε μια κατηγορία, όπως αυτή περιγράφεται από το σύνολο εκπαίδευσης. Με όρους πιθα-

νοτήτων, βρίσκουμε την πιθανότητα μιας κατηγορίας δοθέντος ενός προτύπου. Αυτή είναι η discriminative προσέγγιση και περιλαμβάνει μεθόδους όπως νευρωνικά δίκτυα, Support Vector Machines SVM, Relevance Vector Machines RVM και δένδρα απόφασης. Η δεύτερη προσέγγιση είναι να προσπαθήσουμε να μάθουμε τι είδους πρότυπα παράγει κάθε κατηγορία. Δηλαδή να υπολογίσουμε την πιθανότητα ενός προτύπου δοθείσας μιας κατηγορίας, και στη συνέχεια μπορούμε χρησιμοποιώντας το θεώρημα του Bayes να υπολογίσουμε εύκολα την πιθαγότητα της κατηγορίας δοθέντος του προτύπου. Αυτή είναι generative προσέγγιση, και βασίζεται σε μεθόδους εκτίμησης κατανομής, που θα αναφέρουμε στην συνέχεια. Εκ πρώτης όψεως φαίνεται ότι με αυτό τον τρόπο δυσκολεύουμε το πρόβλημα της ταξινόμησης. Ωστόσο το πλεονέκτημα είναι ότι μαθαίνοντας την κατανομή αποκτάμε μια πιο γενική περιγραφή του χώρου, από αυτή που προσφέρουν τα πρότυπα εκπαίδευσης. Η πληροφορία των προτύπων γίνεται πληροφορία πυκνότητας πιθανότητας, και απεικονίζουμε μάζα πιθανότητας σε κατηγορία. Αυτό μπορεί να μας βοηθήσει για παράδειγμα όταν για κάποια πρότυπα λείπει η κατηγορία τους ή κάποια από τα χαρακτηριστικά τους. Τέλος αξίζει να αναφέρουμε μια τρίτη προσέγγιση, αυτή των k-κοντινότερων γειτόνων (k-nearest neighbours) που αποφεύγει να μάθει οποιαδήποτε απεικόνιση (lazy learning). Απλά αποθηκεύει τα πρότυπα, και με βάση τα πιο κοντινά ταξινομεί κάποιο άγνωστο πρότυπο.

Στα προβλήματα παλινδρόμησης στόχος μας είναι να μάθουμε μια συνάρτηση στο χώρο των πραγματικών αριθμών. Για παράδειγμα θέλουμε να προβλέψουμε την τιμή πώλησης ενός σπιτιού κρίνοντας από την γειτονιά που βρίσκεται, την απόσταση από το κέντρο, αν βρίσκεται κοντά σε ποτάμι κ.λ.π. Μας δίνεται λοιπόν ένα σύνολο παραδειγμάτων της μορφής (πρότυπο, τιμή συνάρτησης), και θέλουμε για νέα πρότυπα που δεν υπάρχουν στο σύνολο να προβλέψουμε την τιμή της συνάρτησης. Η υπόθεση που πρέπει να λάβουμε υπόψη μας κατά την μάθηση είναι ότι οι επιθυμητές έξοδοι που έχουμε στο σύνολο εκπαίδευσης περιέχουν θόρυβο. Επομένως δεν θέλουμε η συνάρτησή μας να απεικονίζει ακριβώς τα ζεύγη του συνόλου εκπαίδευσης, αλλά αναζητούμε μια προσέγγιση που θα συμβιβάζει την ακρίβεια (μικρή απόκλιση από την δοθείσα έξοδο) με την απλότητα της συναρτησιακής μορφής. Για τη λύση τους χρησιμοποιούμε μοντέλα που είναι αρκετά ευέλικτα ώστε να προσεγγίζουν μεγάλο πλήθος συναρτήσεων, όπως πολυώνυμα, νευρωνικά δίκτυα, Support Vector Machines, και Relevance Vector Machines, ενώ τα δένδρα απόφασης προσφέρουν μια μη παραμετρική

αντιμετώπιση.

Στα προβλήματα εκτίμησης κατανομής μας δίνεται ένα σύνολο προτύπων, και θέλουμε να βρούμε μια κατανομή η οποία με μεγάλη πιθανότητα θα μπορούσε να τα έχει παράγει. Η γνώση της κατανομής μας βοηθάει να υπολογίσουμε υπό συνθήκη πιθανότητες. Επιπλέον όπως αναφέραμε ήδη, ένα πρόβλημα ταξινόμησης ακολουθώντας την generative προσέγγιση ανάγεται σε πρόβλημα εκτίμησης της κατανομής των προτύπων κάθε κατηγορίας ξεχωριστά. Οι παραμετρικές μέθοδοι που χρησιμοποιούνται για την εκτίμηση κατανομής υποθέτουν μια συναρτησιακή μορφή για την κατανομή, π.χ. κανονική. Το πλεονέκτημα σε αυτή την περίπτωση είναι ότι η συνάρτηση καθορίζεται από ένα μικρό αριθμό παραμέτρων, που αν τις εκτιμήσουμε τότε η κατανομή είναι γνωστή. Το μειονέκτημα είναι ότι δεν μπορούμε να εκτιμήσουμε έτσι αυθαίρετα πολύπλοκες κατανομές, π.χ. με πολλές κορυφές. Για να το ξεπεράσουμε μπορούμε να καταφύγουμε στην χρήση της *μίξης κατανομών* (mixture model), που είναι ένας γραμμικός συνδυασμός από κατανομές απλής συναρτησιακής μορφής, συνήθως κανονικές.

Στα προβλήματα ομαδοποίησης δίνεται ένα σύνολο προτύπων και θέλουμε να βρούμε τις ομάδες που σχηματίζουν, χωρίς ωστόσο να έχουμε καμιά πληροφορία για τα χαρακτηριστικά των ομάδων, όπως ο αριθμός τους ή τη μορφή τους. Για παράδειγμα μια εταιρία συλλέγει στοιχεία για τις αγορές που κάνουν οι πελάτες της καθώς και δημογραφικά στοιχεία για αυτούς όπως ηλικία, φύλλο, ή μορφωτικό επίπεδο, και θέλει να δει τις προτιμήσεις τους, δηλαδή αν υπάρχουν ομάδες πελατών που έχουν την τάση να αγοράζουν ή να μην αγοράζουν κάποιο προϊόν. Για την επίλυση τέτοιων προβλημάτων μπορούμε χρησιμοποιήσουμε μεθόδους εκτίμησης κατανομής, όπως είναι η μίξη κατανομών και να αντιστοιχήσουμε μια ομάδα σε κάθε συνιστώσα της μίξης. Παρόμοια προσέγγιση ακολουθεί η γνωστή μέθοδος k-μέσων (k-means), χωρίς ωστόσο να κάνει εκτίμηση κατανομής. Μια διαφορετική μέθοδος είναι η ιεραρχική ομαδοποίηση που βασίζεται στην δημιουργία ομάδων, τέτοιων ώστε η απόσταση μεταξύ των ατόμων μιας ομάδας να είναι η ελάχιστη. Ο agglomerative αλγόριθμος ξεκινά με μία ομάδα για κάθε πρότυπο, και φτιάχνει μια ιεραρχία από ομάδες ενώνοντας αυτές που είναι πιο κοντά, μέχρι να καταλήξει σε μια ομάδα που περιλαμβάνει όλα τα πρότυπα. Μια προσέγγιση που βασίζεται στην θεωρία γράφων είναι η *φασματική ομαδοποίηση* (spectral clustering). Μπορούμε να σχηματίσουμε ένα γράφο όπου η απόσταση μεταξύ των προτύ-

πων αποτελεί τα βάρη των ακμών ενός γράφου, και να πάρουμε αποφάσεις για τις ομάδες χρησιμοποιώντας ιδιότητες του πίνακα γειτνίασης (adjacency matrix).

Στα προβλήματα ενισχυτικής μάθησης στόχος μας δεν είναι να δημιουργήσουμε ένα σύστημα που παίρνει μια απόφαση, αλλά μια σειρά από αποφάσεις ακολουθώντας μια στρατηγική. Έτσι έχουμε έναν πράκτορα που αλληλεπιδρά με το περιβάλλον και παίρνει αποφάσεις ανάλογα με την κατάστασή του. Η απόφαση που παίρνει κάθε φορά καθορίζεται από την στρατηγική του, την οποία επιδιώκει να βελτιώσει. Ωστόσο δεν ξέρει κάθε στιγμή αν πήρε την σωστή απόφαση, μόνο μετά από μια αλληλουχία αποφάσεων παίρνει μια επιβράβευση (ή τιμωρία) ανάλογα με το βαθμό της επιτυχίας του, την οποία χρησιμοποιεί για την εκμάθηση της βέλτιστης στρατηγικής.

Ολοκληρώνοντας αυτή την ενότητα αξίζει να αναφερθούμε σε δύο νέα ενδιαφέροντα προβλήματα που έχουν συγκεντρώσει την προσοχή της ερευνητικής κοινότητας. Το πρώτο πρόβλημα είναι η *μάθηση με ημι-επίβλεψη* (semi-supervised learning), η οποία εκμεταλλεύεται τα κοινά στοιχεία της μάθησης με επίβλεψη και της μάθησης χωρίς επίβλεψη. Σε αυτή την περίπτωση το σύνολο εκπαίδευσης περιέχει πρότυπα ταξινομημένα και αταξινόμητα (άγνωστης κατηγορίας). Στόχος είναι η δημιουργία ενός ταξινομητή χρησιμοποιώντας όλα τα διαθέσιμα πρότυπα, επιπλέον η επίδοσή του καθορίζεται από την βέλτιστη χρήση των αταξινόμητων προτύπων τα οποία αποτελούν το μεγαλύτερο μέρος του συνόλου ενώ τα ταξινομημένα είναι πολύ λίγα για να δώσουν μια καλή λύση. Το δεύτερο είναι η *ενεργή μάθηση* (active learning), η οποία συνήθως χρησιμοποιείται σε προβλήματα μάθησης με επίβλεψη αλλά μπορεί να χρησιμοποιηθεί και σε προβλήματα μάθησης χωρίς επίβλεψη. Η βασική ιδέα είναι ότι το σύστημα που εκπαιδεύουμε μπορεί να ζητάει καινούρια πρότυπα εκπαίδευσης για να βελτιώσει την επίδοσή του. Το ενδιαφέρον ζήτημα είναι από ποια περιοχή του χώρου προτύπων θα προέρχονται, έτσι απαιτούνται κριτήρια επιλογής προτύπων που αν τα μάθει το σύστημα θα βελτιώσει την επίδοσή του.

1.2 Συμβολή της Διατριβής

Στην συνέχεια θα παρουσιάσουμε μια περίληψη της διατριβής, και θα επισημάνουμε την συμβολή του πρωτότυπου μέρους της. Το Κεφάλαιο 2 αποτελεί μια σύντομη παρουσίαση των βασικών εννοιών και μεθόδων της μηχανικής μάθησης πάνω στις οποίες αναπτύχθηκε η διατριβή. Γίνεται αναφορά σε θεμελιώδη κριτήρια μάθησης, μοντέλα και τους αλγόριθμους μάθησης. Έτσι κατά σειρά παρουσιάζουμε τα κριτήρια Maximum Likelihood, Maximum a Posteriori και Penalized Likelihood. Στην συνέχεια αναφερόμαστε στα graphical models τα οποία αποτελούν έναν γενικό τρόπο αναπαράστασης πιθανοτικών μοντέλων. Ακολουθεί η περιγραφή της μίξης κανονικών κατανομών (Gaussian mixture model) και του πιθανοτικού δικτύου ακτινικών συναρτήσεων βάσης (Probabilistic RBF), με τα οποία σε επόμενα κεφάλαια προτείνουμε λύσεις σε προβλήματα μάθησης χωρίς επίβλεψη και μάθησης με επίβλεψη αντίστοιχα. Το κεφάλαιο ολοκληρώνεται με μια αναφορά στον αλγόριθμο EM, στην εφαρμογή του για εκπαίδευση της μίξης κανονικών κατανομών, και στη variational Bayesian μέθοδο.

Στο Κεφάλαιο 3 εξετάζουμε το πρόβλημα της ομαδοποίησης. Το μοντέλο που χρησιμοποιούμε είναι μια μίξη κανονικών κατανομών, με την οποία εκτιμούμε την κατανομή του συνόλου δεδομένων, και στην συνέχεια ορίζουμε τις ομάδες όπως υποδεικνύονται από τις συνιστώσες της μίξης. Ταυτόχρονα δίνουμε και μια απάντηση στην κρίσιμη επιλογή του αριθμού των συνιστωσών της μίξης. Προτείνουμε μια αυξητική μέθοδο για την επιλογή μοντέλου και την εκπαίδευση μιας μίξης, που βασίζεται στη variational Bayesian μέθοδο. Η μέθοδος προσθέτει συνιστώσες στην μίξη χρησιμοποιώντας έναν έλεγχο Bayesian διάσπασης: μια συνιστώσα διασπάται σε δύο νέες συνιστώσες και στη συνέχεια εφαρμόζονται variational Bayesian εξισώσεις ενημέρωσης μόνο στις παραμέτρους των δυο νέων συνιστωσών. Ως αποτέλεσμα, είτε και οι δύο συνιστώσες διατηρούνται στη μίξη, είτε μια από της δύο αποδεικνύεται περιττή και απαλείφεται από την μίξη. Στην προσέγγισή μας το πρόβλημα επιλογής μοντέλου αντιμετωπίζεται τοπικά σε μια περιοχή του χώρου δεδομένων, έτσι μπορούμε να θέσουμε πιο κατατοπιστικές εκ των προτέρων πιθανότητες βασιζόμενοι στην τοπική κατανομή των δεδομένων. Για την υλοποίηση αυτής της προσέγγισης παρουσιάζουμε μια βελτιωμένη Bayesian μίξη, καθώς και έναν αλγόριθμο εκπαίδευσης που εφαρμόζει επαναληπτικά έναν έλεγχο διάσπασης σε κάθε συνιστώσα της μίξης.

Στο Κεφάλαιο 4 εξετάζουμε πάλι το πρόβλημα της ομαδοποίησης, στο οποίο δίνουμε λύση εκτιμώντας την κατανομή του συνόλου δεδομένων με μια μίξη κανονικών κατανομών. Για την επιλογή του αριθμού των συνιστωσών της μίξης εκμεταλλευόμαστε την τυπική variational Bayesian μέθοδο, και επικεντρωνόμαστε στο πρόβλημα της επιλογής χαρακτηριστικών. Αναζητούμε χαρακτηριστικά τα οποία αναδεικνύουν τις διαφορές μεταξύ των ομάδων, και προτείνουμε μια variational Bayesian μέθοδο για την εκπαίδευση μιας μίξης που αντιμετωπίζει ταυτόχρονα το πρόβλημα της επιλογής χαρακτηριστικών και της επιλογής μοντέλου. Η μέθοδος στηρίζεται σε μια Bayesian μίξη δύο επιπέδων, που για κάθε συνιστώσα της τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους, και με μια πιθανότητα είναι και υπό συνθήκη ανεξάρτητα από την συνιστώσα. Η προτεινόμενη μέθοδος μπορεί να εκτιμήσει ταυτόχρονα τις παραμέτρους της μίξης, τον αριθμό των συνιστωσών της και τα εξέχοντα χαρακτηριστικά των δεδομένων, και έχει παρουσιαστεί στην [23].

Στο Κεφάλαιο 5 εξετάζουμε το πρόβλημα της κατάτμησης μιας εικόνας, ως πρόβλημα ομαδοποίησης των εικονοστοιχείων (pixel) της. Αφού εξάγουμε κατάλληλα χαρακτηριστικά από τα εικονοστοιχεία, η ομαδοποίηση των προτύπων μπορεί να γίνει με μια μίξη κανονικών κατανομών. Ωστόσο στη μέχρι τώρα βιβλιογραφία πολλά από τα ζητήματα που αφορούν την εκτίμηση των παραμέτρων της μίξης για την μοντελοποίηση εικόνων δεν έχουν αντιμετωπιστεί επαρκώς. Τέτοια ζητήματα είναι η επιλογή του αριθμού των συνιστωσών της μίξης, και ο μεγάλος όγκος δεδομένων για εικόνες με τυπική ανάλυση (π.χ. 256×256). Προτείνουμε την σταδιακή ομαδοποίηση των εικονοστοιχείων, και παρουσιάζουμε μια μέθοδο ενεργητικής μάθησης που αυξάνει σταδιακά το πλήθος των εικονοστοιχείων που χρησιμοποιούνται για την επαναλαμβανόμενη κατάτμηση της εικόνας. Σε κάθε επανάληψη της μεθόδου έχουμε ενσωματώσει έναν αλγόριθμο εκπαίδευσης της μίξης που βασίζεται στην variational Bayesian προσέγγιση που παρουσιάσαμε στο Κεφάλαιο 3, και αντιμετωπίζει αποδοτικά το πρόβλημα επιλογής μοντέλου. Τα πειραματικά αποτελέσματα έδειξαν ότι η μέθοδος έχει ικανοποιητικό χρόνο εκτέλεσης και προσαρμόζει με επιτυχία των αριθμό των περιοχών που χρησιμοποιεί στην πολυπλοκότητα της εικόνας. Η προτεινόμενη μέθοδος έχει παρουσιαστεί στην [19].

Στο Κεφάλαιο 6 εξετάζουμε το πρόβλημα της ταξινόμησης, και χρησιμοποιούμε το πιθανοτικό δίκτυο ακτινικών συναρτήσεων βάσης (Probabilistic RBF) που αποτελεί μια πιθανοτική παραλλαγή του νευρωνικού δικτύου RBF για προβλήματα ταξινόμησης, και επιπλέον

επεκτείνει την τυπική μίξη κανονικών κατανομών επιτρέποντας σε όλες της κατηγορίες να μοιράζονται τις ίδιες συνιστώσες. Η τυπική μέθοδος εκπαίδευσης του PRBF για ταξινόμηση χρησιμοποιεί τον αλγόριθμο EM, και το τελικό δίκτυο εξαρτάται από την αρχικοποίηση των παραμέτρων του. Προτείνουμε μια μέθοδο για την αυξητική εκπαίδευση του δικτύου PRBF, η οποία ξεκινά με μια μόνο συνιστώσα, και σταδιακά προσθέτει περισσότερες σε κατάλληλες θέσεις στον χώρο δεδομένων. Η προσθήκη μιας επιπλέον συνιστώσας βασίζεται σε ένα κριτήριο για τον εντοπισμό μιας περιοχής που είναι κρίσιμη για την ταξινόμηση, συγκεκριμένα στο όριο απόφασης. Μετά από την προσθήκη όλων των συνιστωσών, η μέθοδος διασπά κάθε συνιστώσα σε υπο-συνιστώσες που κάθε μια αντιστοιχεί σε διαφορετική κατηγορία. Η προτεινόμενη μέθοδος έχει παρουσιαστεί στις [18, 22].

Στο Κεφάλαιο 7 εξετάζουμε το πρόβλημα της ενεργητικής μάθησης για ταξινόμηση. Μια ενδιαφέρουσα παραλλαγή του προκύπτει αν το σύνολο εκπαίδευσης περιέχει ταξινομημένα και αταξινόμητα πρότυπα (pool-based active learning), τότε ο αλγόριθμος μπορεί να επιλέξει και να ρωτήσει την κατηγορία κάποιου από τα αταξινόμητα. Προτείνουμε μια μέθοδο ενεργητικής μάθησης για την εκπαίδευση του PRBF. Αρχικά προτείνουμε μια αυξητική μέθοδο για την εκπαίδευση του δικτύου με ημι-επίβλεψη που βασίζεται στον αλγόριθμο EM. Στην συνέχεια προτείνουμε μια μέθοδο ενεργητικής μάθησης που επαναληπτικά εφαρμόζει την διαδικασία μάθησης με ημι-επίβλεψη στα ταξινομημένα και τα αταξινόμητα δεδομένα, και στην συνέχεια επιλέγει ένα αταξινόμητο δεδομένο και ζητά να μάθει την κατηγορία του. Το κριτήριο επιλογής που προτείνουμε επιλέγει σημεία κοντά στο όριο απόφασης του τρέχοντος ταξινομητή, και διευκολύνει την αυξητική μέθοδο με ημι-επίβλεψη που επίσης εκμεταλλεύεται το όριο απόφασης. Η προτεινόμενη μέθοδος έχει παρουσιαστεί στις [21, 20].

Η διατριβή ολοκληρώνεται με το Κεφάλαιο 8, όπου γίνεται μια ανασκόπηση των μεθόδων που παρουσιάστηκαν και παρατίθενται ζητήματα για περαιτέρω έρευνα.

ΚΕΦΑΛΑΙΟ 2

ΣΤΟΙΧΕΙΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Κάθε φορά που ερχόμαστε αντιμέτωποι με ένα πρόβλημα μηχανικής μάθησης πρέπει να κάνουμε τρεις επιλογές, οι οποίες θα καθορίσουν και την λύση μας. Η πρώτη επιλογή αφορά το μοντέλο που θα χρησιμοποιήσουμε για να περιγράψουμε την λύση, η δεύτερη αφορά το κριτήριο βελτιστοποίησης με το οποίο θα αποτιμήσουμε την συμπεριφορά του μοντέλου, και τέλος είναι η επιλογή του αλγορίθμου που θα χρησιμοποιήσουμε για να ρυθμίσουμε τις παραμέτρους του μοντέλου. Οι επιλογές μας αυτές οδηγούνται από την φύση του προβλήματος και το είδος της μάθησης που απαιτείται. Εξαρτάται επίσης από την φύση των διαθέσιμων δεδομένων που περιγράφουν το πρόβλημα, και ποια χρήσιμη πληροφορία μπορούμε να εξάγουμε από αυτά. Στις επόμενες ενότητες θα παρουσιάσουμε τα κριτήρια μάθησης, τα μοντέλα και τους αλγόριθμους μάθησης που μας απασχόλησαν.

2.1 Κριτήρια Μάθησης

Στην περίπτωση των νευρωνικών δικτύων, το πιο συχνά χρησιμοποιούμενο κριτήριο για μάθηση είναι το *τετραγωνικό σφάλμα*, που ορίζεται ως το άθροισμα των τετραγωνικών αποστάσεων της επιθυμητής εξόδου από την έξοδο του μοντέλου για κάθε είσοδο. Σε ένα πρόβλημα μάθησης με επίβλεψη που περιγράφεται από ένα σύνολο προτύπων $D = \{(x_n, y_n) | n = 1, \ldots, N\}$, όπου το $y_n \in \Re^m$ αποτελεί την επιθυμητή έξοδο για την είσο-

δο $x_n \in \Re^d$, το τετραγωνικό σφάλμα E ενός νευρωνικού δικτύου f ορίζεται ως

$$E(\theta) = \sum_{n=1}^{N} ||y_n - f(x_n; \theta)||^2 \tag{2.1}$$

όπου θ οι παράμετροι του δικτύου. Στόχος ενός αλγόριθμου μάθησης είναι να βρει τις παραμέτρους θ^* του μοντέλου που ελαχιστοποιούν το τετραγωνικό σφάλμα

$$\theta^* = \arg\max_{\theta} \sum_{n=1}^{N} ||y_n - f(x_n; \theta)||^2 \tag{2.2}$$

Μια συχνή τροποποίηση του τετραγωνικού σφάλματος είναι η προσθήκη ενός όρου ποινής, σύμφωνα με τα παρακάτω

$$\theta^* = \arg\max_{\theta} \{E(\theta) - \lambda g(\theta)\} \tag{2.3}$$

όπου το λ είναι μια παράμετρος που καθορίζει το πόσο σημαντικός είναι ο όρος ποινής. Η $g(\theta)$ είναι μια μη-αρνητική συνάρτηση, που μετρά την πολυπλοκότητα του μοντέλου όπως αυτή καθορίζεται από τις παραμέτρους. Η μορφή της εξαρτάται κάθε φορά από το πρόβλημα και το μοντέλο που χρησιμοποιούμε, και γενικά είναι μικρή για απλά μοντέλα ("ομαλές" συναρτήσεις) και μεγάλη για πολύπλοκα μοντέλα ("απότομες" συναρτήσεις). Έτσι κατά την μεγιστοποίηση πρέπει να γίνει ένας συμβιβασμός ανάμεσα σε μικρό τετραγωνικό σφάλμα και μικρή πολυπλοκότητα. Μια συνηθισμένη μορφή του όρου ποινής για ένα νευρωνικό δίκτυο είναι το άθροισμα των τετραγώνων των βαρών του, και έχει στόχο να περιορίσει τις τιμές τους κοντά στο μηδέν.

Η *μέγιστη πιθανοφάνεια* (Maximum Likelihood ML) είναι μια διαφορετική αντικειμενική συνάρτηση, που προέρχεται από την στατιστική βιβλιογραφία. Τη χρησιμοποιούμε όταν μπορούμε να ανάγουμε ένα πρόβλημα μάθησης σε ένα πρόβλημα εκτίμησης μιας κατανομής $p(x|\theta)$, όπου το x είναι η παρατηρούμενη τυχαία μεταβλητή και θ είναι οι παράμετροι της κατανομής της. Για ένα σύνολο παρατηρήσεων $X = \{x_1, \ldots, x_N\}$, η πιθανοφάνεια είναι η κατανομή $p(X|\theta)$ σαν συνάρτηση των παραμέτρων θ. Διαισθητικά το κριτήριο της μέγιστης πιθανοφάνειας στοχεύει στο να βρει τις παραμέτρους θ_{ML} που μεγιστοποιούν την πιθανότητα των παρατηρούμενων δεδομένων. Στην πράξη για λόγους απλότητας, συνήθως μεγιστοποιείται ο λογάριθμος της πιθανοφάνειας $\log p(X|\theta)$, αφού ο λογάριθμος είναι μονότονη συνάρτηση και δεν επηρεάζει την θέση των ακρότατων της συνάρτησης πιθανοφάνειας:

$$\theta_{ML} = \arg\max_{\theta} \log p(X|\theta) \tag{2.4}$$

Μια άλλη συνηθισμένη υπόθεση είναι ότι οι παρατηρήσεις μας είναι ανεξάρτητες και όμοια κατανεμημένες, οπότε για την λογαριθμική πιθανοφάνεια ισχύει

$$\log p(X|\theta) = \log p(x_1, \ldots, x_N|\theta) = \log \prod_{n=1}^{N} p(x_n|\theta) = \sum_{n=1}^{N} \log p(x_n|\theta) \tag{2.5}$$

Έχει ενδιαφέρον το ότι το κριτήριο μέγιστης πιθανοφάνειας σχετίζεται με το τετραγωνικό σφάλμα. Ειδικότερα, αν υποθέσουμε ότι οι παρατηρήσεις μας προέρχονται από μια ντετερμινιστική σύνάρτηση και στην συνέχεια τους προστίθεται θόρυβος, ο οποίος είναι ανεξάρτητος και όμοια κατανεμημένος για κάθε παρατήρηση και ακολουθεί κανονική κατανομή με μέση τιμή μηδέν, τότε τα δυο κριτήρια ταυτίζονται, μια απόδειξη υπάρχει στην [10].

Ένα άλλο κριτήριο που σχετίζεται στενά με την μέγιστη πιθανοφάνεια είναι η *μέγιστη εκ των υστέρων πιθανότητα* (Maximum a Posteriori MAP), όπου αναζητάμε τις παραμέτρους θ που μεγιστοποιούν την $p(\theta|X)$. Η πιθανότητα των παραμέτρων δοθέντων των παρατηρήσεων $p(\theta|X)$ ονομάζεται *εκ των υστέρων* (a posteriori) πιθανότητα για να δηλώσει ότι ποσοτικοποιεί την πεποίθησή μας για τις τιμές των παραμέτρων αφού έχουμε δει ένα σύνολο δεδομένων. Σε αντιδιαστολή με την *εκ των προτέρων* (a priori) πιθανότητα $p(\theta)$, που αφορά της πεποίθησή μας για τις τιμές που μπορεί να έχουν οι παράμετροι πριν συλλέξουμε ένα σύνολο δεδομένων. Από το θεώρημα του Bayes μπορούμε να υπολογίσουμε την εκ των υστέρων πιθανότητα σαν συνάρτηση της πιθανοφάνειας

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \tag{2.6}$$

όπου $p(X) = \sum_\theta p(X|\theta)p(\theta)$ είναι η κατανομή περιθωρίου τον δεδομένων, και αποτελεί την πεποίθησή μας για το πόσο είναι πιθανό να παρατηρήσουμε το συγκεκριμένο σύνολο δεδομένων, ανεξάρτητα από τις παραμέτρους που το παρήγαγαν. Μέχρι τώρα υποθέσαμε σιωπηλά ότι οι παράμετροι θ είναι διακριτές, ωστόσο τα ίδια ισχύουν και αν είναι συνεχείς κάνοντας τις απαραίτητες αλλαγές, το άθροισμα δηλαδή στην περιθωριοποίηση γίνεται ολοκλήρωμα. Οπότε οι παράμετροι θ_{MAP} που ψάχνουμε προκύπτουν από την παρακάτω μεγιστοποίηση

$$\theta_{MAP} = \arg\max_\theta \log p(\theta|X) = \arg\max_\theta \log \{p(X|\theta)p(\theta)\} \tag{2.7}$$

όπου η δεύτερη ισότητα προκύπτει εφαρμόζοντας το θεώρημα του Bayes, και απαλείφοντας τον όρο που δεν εξαρτάται από τις παραμέτρους. Δηλαδή το κριτήριο που μεγιστοποιούμε

είναι η πιθανοφάνεια συν μια ποσότητα που εξαρτάται μόνο από τις παραμέτρους και περιορίζει τις τιμές που μπορούν να πάρουν. Αν δεν εκφράσουμε κάποια προτίμηση για τις τιμές των παραμέτρων και επιλέξουμε ομοιόμορφη $p(\theta)$ τότε προκύπτει ακριβώς το κριτήριο της μέγιστης πιθανοφάνειας.

Ένα κριτήριο με πιο γενική μορφή περιορισμών είναι η πιθανοφάνεια με ποινή (Penalized Likelihood PL), όπου κατά την μεγιστοποίηση προσθέτουμε στην συνάρτηση της λογαριθμικής πιθανοφάνειας έναν όρο ποινής που εξαρτάται και πάλι από τις παραμέτρους, *χωρίς να είναι απαραίτητα κατανομή*. Η γενική μορφή του κριτηρίου είναι

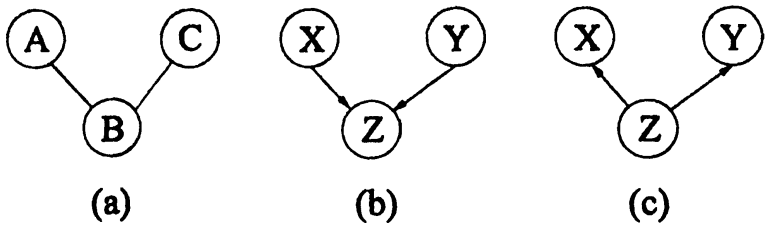
$$\theta_{MP} = \arg\max_{\theta} \{\log p(X|\theta) - \lambda g(\theta)\} \tag{2.8}$$

όπου η παράμετρος λ και η συνάρτηση $g(\theta)$ παίζουν τον ίδιο ρόλο όπως και στον όρο ποινής του τετραγωνικού σφάλματος (2.3).

Στα προβλήματα που μας απασχόλησαν στην παρούσα διατριβή η πιθανοφάνεια παίζει κεντρικό ρόλο. Στην συνέχεια του κεφαλαίου θα παρουσιάσουμε τα μοντέλα που χρησιμοποιήσαμε και τις βασικές αλγοριθμικές προσεγγίσεις που εκμεταλλευτήκαμε.

2.2 Γραφικά Μοντέλα

Στην παρούσα ενότητα θα μας απασχολήσουν τα γραφικά μοντέλα [12, 40] για την περιγραφή κατανομών. Ένα γραφικό μοντέλο αποτελεί την γραφική αναπαράσταση ενός συνόλου από κοινές κατανομές, ορισμένες επί ενός συνόλου πολυδιάστατων τυχαίων μεταβλητών. Στο Σχήμα 2.1 εμφανίζονται κάποια τυπικά παραδείγματα. Η οπτική περιγραφή των εξαρτήσεων μεταξύ των τυχαίων μεταβλητών που μας προσφέρει το γραφικό μοντέλο μας βοηθάει στο να κατανοήσουμε διαισθητικά, και να εμβαθύνουμε την μελέτη του εκάστοτε προβλήματος. Προσφέρει επίσης ένα κοινό πλαίσιο για την περιγραφή και την ενοποίηση μοντέλων από διαφορετικές επιστημονικές περιοχές. Έτσι έμμεσα διευκολύνει την μεταφορά ιδεών μεταξύ επιστημονικών περιοχών, και διευκολύνει τον σχεδιασμό νέων μοντέλων. Το άμεσο όφελος από την χρήση τους είναι η αποδοτική εφαρμογή του θεωρήματος του Bayes και άλλων βασικών θεωρημάτων των πιθανοτήτων, καθώς και η δυνατότητα για εφαρμογή αποτελεσμάτων της θεωρίας γράφων. Οι δύο πιο σημαντικές κατηγορίες γραφικών μοντέλων εί-



Σχήμα 2.1: Τρία τυπικά γραφικά μοντέλα.

ναι τα μη-κατευθυνόμενα (undirected) γραφήματα, και τα κατευθυνόμενα άκυκλα (directed acyclic) γραφήματα, δηλαδή κατευθυνόμενα γραφήματα χωρίς κατευθυνόμενους κύκλους. Στο Σχήμα 2.1 εμφανίζεται ένα μη-κατευθυνόμενο γραφικό μοντέλο για το σύνολο τυχαίων μεταβλητών $\{A, B, C\}$, και δύο διαφορετικά κατευθυνόμενα άκυκλα γραφικά μοντέλα για το σύνολο τυχαίων μεταβλητών $\{X, Y, Z\}$. Τα κατευθυνόμενα άκυκλα γραφήματα, τα οποία και θα μας απασχολήσουν στην συνέχεια, αναφέρονται στην βιβλιογραφία και ως Bayesian δίκτυα.

Ας θεωρήσουμε ένα οποιοδήποτε σύστημα το οποίο περιγράφεται από τρεις τυχαίες μεταβλητές X, Y και Z, και έστω ότι για την από κοινού τους κατανομή ισχύει $p(X, Y, Z) = p(Z|Y, X)p(Y, X) = p(Z|Y, X)p(Y)p(X)$, δηλαδή η κατανομή του Z εξαρτάται από τα X και Y, και τα X και Y είναι ανεξάρτητα. Αυτή η σχέση μεταξύ των τυχαίων μεταβλητών περιγράφεται από το γραφικό μοντέλο στο Σχήμα 2.1(b). Κάθε κόμβος του γραφικού μοντέλου αντιστοιχεί σε μια τυχαία μεταβλητή, και κάθε κατευθυνόμενη ακμή του δηλώνει την εξάρτηση του κόμβου τερματισμού από τον κόμβο έναρξης. Στο Σχήμα 2.1(b) οι δύο ακμές δηλώνουν ότι το Z εξαρτάται από τα X και Y, και η εξάρτησή του περιγράφεται από την υπό συνθήκη κατανομή $p(Z|Y, X)$. Η απουσία ακμής μεταξύ δύο κόμβων δηλώνει την ανεξαρτησία τους, δηλαδή στο Σχήμα 2.1(b) η απουσία ακμής μεταξύ X και Y δηλώνει ότι $p(Y, X) = p(Y)p(X)$. Ωστόσο πιο γενικά η απουσία ακμής δηλώνει την υπό συνθήκη ανεξαρτησία δυο τυχαίων μεταβλητών. Έτσι στο Σχήμα 2.1(c) περιγράφεται ότι $p(X, Y|Z) = p(X|Z)p(Y|Z)$, και για την από κοινού κατανομή ισχύει $p(X, Y, Z) = p(X|Z)p(Y|Z)p(Z)$.

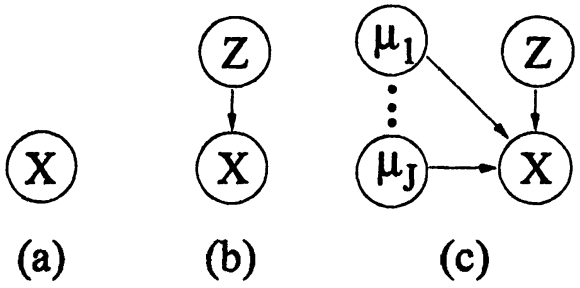
Ένα γραφικό μοντέλο αναπαριστά μια διάσπαση (decomposition) της από κοινού κατανομής όλων των τυχαίων μεταβλητών ενός συστήματος. Αποτελεί έτσι έναν τρόπο για να περιγράψουμε ποιες τυχαίες μεταβλητές είναι υπό συνθήκη ανεξάρτητες, και ποιες υπό συν-

θήκη κατανομές μας είναι απαραίτητες για να υπολογίσουμε την από κοινού κατανομή όλων των τυχαίων μεταβλητών. Χρησιμοποιώντας λοιπόν το γραφικό μοντέλο για ένα σύστημα τυχαίων μεταβλητών $U = \{X_1, X_2, \ldots, X_n\}$ μπορούμε να γράψουμε την από κοινού κατανομή τους ως $p(X_1, X_2, \ldots, X_n) = \prod_i p(X_i|\text{pa}(X_i))$, όπου $\text{pa}(\cdot)$ δηλώνει τους προγόνους ενός κόμβου στο γραφικό μοντέλο. Όσο αφορά τις υπό συνθήκη ανεξαρτησίες, ισχύει για έναν κόμβο ότι δοθέντων των προγόνων του, των απογόνων του και των προγόνων τους είναι ανεξάρτητος από κάθε άλλο κόμβο στο δίκτυο. Έτσι για να απομονώσουμε όλες τις επιδράσεις που δέχεται ο κόμβος πρέπει να λάβουμε υπόψη μας εκτός από το προφανές σύνολο των προγόνων και των απογόνων του, και το σύνολο των προγόνων των απογόνων του γιατί η επίδρασή τους συμψηφίζεται με την επίδραση του εν λόγω κόμβου για να καθορίσουν τους κοινούς τους απογόνους.

Στην συνέχεια θα δούμε πώς περιγράφονται κάποια ευρέως χρησιμοποιούμενα μοντέλα χρησιμοποιώντας κατευθυνόμενα άκυκλα γραφήματα. Αν θεωρήσουμε ότι έχουμε μια μονοδιάστατη παρατήρηση X που ακολουθεί κανονική κατανομή $\mathcal{N}(X|\mu, \sigma^2)$ τότε το γραφικό μοντέλο αποτελείται από έναν μόνο κόμβο, και περιγράφει ότι $p(X) = \mathcal{N}(X|\mu, \sigma^2)$ όπως φαίνεται στο Σχήμα 2.2(a). Ένα πιο σύνθετο σύστημα προκύπτει αν υποθέσουμε ότι έχουμε στην διάθεσή μας ένα σύνολο από J κανονικές κατανομές $\{\mathcal{N}(X|\mu_j, \sigma_j^2)|j = 1, \ldots, J\}$, και μια διακριτή τυχαία μεταβλητή Z που παίρνει τιμές στο $\{1, \ldots, J\}$. Το Z καθορίζει την παραγωγή του X, με την έννοια ότι αν $Z = j$ τότε $p(X|Z = j) = \mathcal{N}(X|\mu_j, \sigma_j^2)$ και η πιθανότητα να συμβεί αυτό είναι $p(Z = j) = \pi_j$. Χρειαζόμαστε ένα γραφικό μοντέλο με δύο κόμβους για να περιγράψουμε την από κοινού κατανομή $p(X, Z) = p(X|Z)p(Z)$ όπως φαίνεται στο Σχήμα 2.2(b). Η κατανομή περιθωρίου (marginal distribution) του X είναι

$$p(X) = \sum_Z p(X|Z)p(Z) = \sum_{j=1}^{J} \pi_j \mathcal{N}(X|\mu_j, \sigma_j^2),$$

δηλαδή μια μίξη κανονικών κατανομών. Επεκτείνοντας ακόμα περισσότερο το σύστημα παραγωγής του X μπορούμε να ορίσουμε κατανομές επί των παραμέτρων της μίξης. Ορίζοντας λοιπόν για τα κέντρα των κανονικών συνιστωσών της μίξης τις κατανομές $p(\mu_1), \ldots, p(\mu_J)$, το X ακολουθεί κανονική κατανομή δοθέντων του $Z = j$ και των $\mu_1, \ldots, \mu_J$, δηλαδή



Σχήμα 2.2: (a) Το γραφικό μοντέλο για μια κανονική κατανομή. (b) Το γραφικό μοντέλο για μια μίξη κανονικών κατανομών. (c) Το γραφικό μοντέλο για μια Bayesian μίξη κανονικών κατανομών.

$p(X|Z=j,\mu_1,\ldots,\mu_J)=\mathcal{N}(X|\mu_j,\sigma_j^2)$. Η από κοινού κατανομή γίνεται

$$p(X,Z,\mu_1,\ldots,\mu_J)=p(X,Z|\mu_1,\ldots,\mu_J)\prod_{j=1}^{J}p(\mu_j)=p(X|Z,\mu_1,\ldots,\mu_J)p(Z)\prod_{j=1}^{J}p(\mu_j),$$

και το αντίστοιχο γραφικό μοντέλο φαίνεται στο Σχήμα 2.2(c). Περιθωριοποιώντας το Z, η κατανομή περιθωρίου του X δοθέντων των $\mu_1,\ldots,\mu_J$ είναι πάλι μια μίξη

$$p(X|\mu_1,\ldots,\mu_J)=\sum_{Z}p(X|Z,\mu_1,\ldots,\mu_J)p(Z)=\sum_{j=1}^{J}\pi_j\mathcal{N}(X|\mu_j,\sigma_j^2).$$

Συνήθως μια τέτοια μίξη, για την οποία έχουν οριστεί εκ των προτέρων κατανομές επί των παραμέτρων της, αναφέρεται στην βιβλιογραφία σαν Bayesian μίξη. Άλλα παραδείγματα γραφικών μοντέλων αποτελούν τα Hidden Markov Models [33], Independent Factor Analysis [7], Mixture of Factor Analyzers [34] και Mixture of Probabilistic Principal Component Analyzers [75].

2.3 Τοπικά Μοντέλα

Αλλάζοντας οπτική γωνία, στην συνέχεια θα μελετήσουμε τα μοντέλα με βάση την πληροφορία που περιγράφουν οι παράμετροί τους. Μας ενδιαφέρει η περιγραφή του χώρου των δεδομένων, και θα επικεντρωθούμε σε μοντέλα τα οποία περιγράφουν τον χώρο χρησιμοποιώντας ένα σύνολο από μονάδες (συνιστώσες). Κάθε μια από αυτές τις συνιστώσες περιγράφει τα τοπικά χαρακτηριστικά μιας ομάδας δεδομένων. Τα δύο μοντέλα που θα μας

απασχολήσουν είναι η μίξη κανονικών κατανομών και το νευρωνικό δίκτυο ακτινικών συναρτήσεων βάσης. Τα μοντέλα αυτά προέρχονται από διαφορετικές επιστημονικές περιοχές, το πρώτο από την στατιστική και το δεύτερο από τα νευρωνικά δίκτυα, ωστόσο θα δούμε ότι μοιράζονται κοινά χαρακτηριστικά, τα οποία θα αναδείξουμε, γεγονός που μας επιτρέπει να τα χειριστούμε με έναν ενιαίο και συνεπή τρόπο.

2.3.1 Η μίξη κανονικών κατανομών

Η μίξη κανονικών κατανομών αποτελεί το πιο ευέλικτο μοντέλο της στατιστικής για την εκτίμηση κατανομών, και μπορεί να χρησιμοποιηθεί τόσο για την επίλυση προβλημάτων μάθησης χωρίς επίβλεψη όσο και μάθησης με επίβλεψη. Αυτές οι δυνατότητές του το καθιστούν ένα πολύ ενδιαφέρον ερευνητικό αντικείμενο και στόχο εκτεταμένης μελέτης, μια λεπτομερής ανασκόπηση της περιοχής γίνεται στις [28, 58]. Η μίξη είναι ένας κυρτός γραμμικός συνδυασμός κατανομών, και χρησιμοποιείται για να περιγράψει την κατανομή συνόλων δεδομένων σε περιπτώσεις που οι απλές γνωστές κατανομές δεν επαρκούν, π.χ. δεν επαρκεί μόνο μια κανονική, ή μια εκθετική κατανομή. Το βασικό χαρακτηριστικό τέτοιων συνόλων δεδομένων είναι ότι η κατανομή τους παρουσιάζει πολλές κορυφές. Στο Σχήμα 2.3 δίνεται ένα παράδειγμα μιας μονοδιάστατης μίξης με δύο κανονικές συνιστώσες. Συνήθως ερμηνεύουμε κάθε κορυφή της κατανομής με μια συνιστώσα της μίξης, ωστόσο δεν υπάρχει πάντα ένα-προς-ένα αντιστοιχία μεταξύ συνιστωσών και κορυφών, μια ανάλυση του φαινομένου γίνεται στις [28, 14].

Η υπόθεση ότι ένα σύνολο παρατηρήσεων X ακολουθεί μια μίξη κανονικών κατανομών δηλώνει ότι οι παρατηρήσεις έχουν παραχθεί σύμφωνα με το παρακάτω μοντέλο παραγωγής:

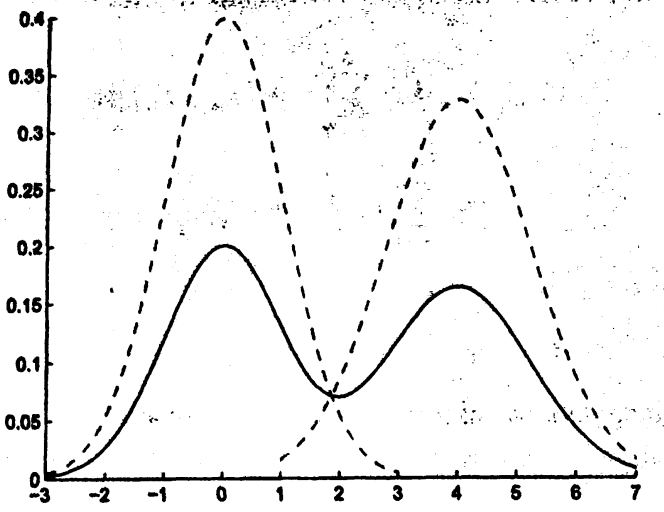
Μας δίνεται ένα σύνολο από J κανονικές κατανομές $\{p_1, \ldots, p_J\}$.

1. Επιλέγουμε τυχαία μια από αυτές με πιθανότητα $\pi_1, \ldots, \pi_J$ αντίστοιχα, έστω την p_j.

2. Δειγματοληπτούμε μια παρατήρηση από την επιλεγμένη κατανομή p_j.

Το γραφικό μοντέλο που αντιστοιχεί σε αυτή την διαδικασία παρουσιάζεται στο Σχήμα 2.2(b), και η τιμή της τυχαίας μεταβλητής Z καθορίζει την κατανομή p_j που επιλέγεται στο πρώτο βήμα. Για την από κοινού κατανομή των τυχαίων μεταβλητών X και Z ισχύει

$$p(X, Z) = p(X|Z)p(Z) \tag{2.9}$$



Σχήμα 2.3: Μια μονοδιάστατη μίξη κανονικών κατανομών (συνεχής γραμμή) με δυο συνιστώσες (διακεκομμένη γραμμή).

Περιθωριοποιώντας την Z βρίσκουμε ότι η κατανομή της X είναι μια μίξη

$$p(X=x) = \sum_{Z} p(X=x|Z)p(Z) = \sum_{j=1}^{J} p(X=x|Z=j)p(Z=j) = \sum_{j=1}^{J} \pi_j p_j(x) \quad (2.10)$$

Αν οι συνιστώσες της μίξης είναι πολυδιάστατες κανονικές κατανομές $p_j(x) = \mathcal{N}(x|\mu_j, \Sigma_j)$ με μέση τιμή $\mu_j \in \Re^d$ και πίνακα συνδιακύμανσης $\Sigma_j \in \Re^{d,d}$, τότε το x ακολουθεί την κατανομή

$$p(x) = \sum_{j=1}^{J} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j) \quad (2.11)$$

και έχει μέση τιμή $\langle x \rangle_p = \sum_{j=1}^{J} \pi_j \mu_j$ και πίνακα συνδιακύμανσης

$$\mathrm{cov}\{x\}_p = \sum_{j=1}^{J} \pi_j \Sigma_j + \sum_{j=1}^{J} \pi_j \mu_j \mu_j^T - \left(\sum_{j=1}^{J} \pi_j \mu_j\right)\left(\sum_{j=1}^{J} \pi_j \mu_j\right)^T$$

Η πρώτη σχέση αποδεικνύεται εύκολα, από τον ορισμό της μέσης τιμής έχουμε

$$\begin{aligned}\langle x\rangle_p &= \int x \sum_{j=1}^{J} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)\, \mathrm{d}x \\ &= \sum_{j=1}^{J} \pi_j \int x\, \mathcal{N}(x|\mu_j, \Sigma_j)\, \mathrm{d}x \\ &= \sum_{j=1}^{J} \pi_j\, \mu_j \end{aligned} \tag{2.12}$$

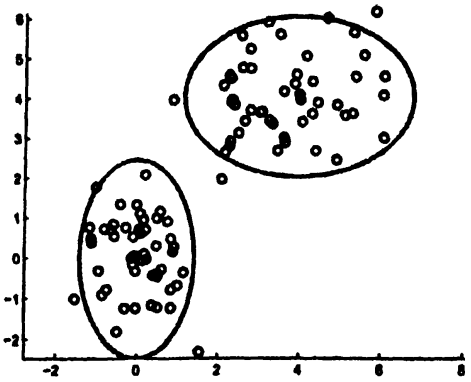
Με τον ίδιο τρόπο μπορούμε να υπολογίσουμε την αυτοσυσχέτιση του x. Από τον ορισμό της ισχύει

$$\begin{aligned}\langle xx^T\rangle_p &= \int xx^T \sum_{j=1}^{J} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)\, \mathrm{d}x \\ &= \sum_{j=1}^{J} \pi_j \int xx^T\, \mathcal{N}(x|\mu_j, \Sigma_j)\, \mathrm{d}x \\ &= \sum_{j=1}^{J} \pi_j\, \left(\Sigma_j + \mu_j\mu_j^T\right) \end{aligned} \tag{2.13}$$

Η τελευταία ισότητα προκύπτει επειδή η αυτοσυσχέτιση μιας τυχαίας μεταβλητής ισούται με το άθροισμα του πίνακα συνδιακύμανσης συν το εξωτερικό γινόμενο της μέσης τιμής της. Για τον ίδιο λόγο ισχύει ότι

$$\begin{aligned}\mathrm{cov}\left\{x\right\}_p &= \langle xx^T\rangle_p - \langle x\rangle_p\langle x\rangle_p^T \\ &= \sum_{j=1}^{J} \pi_j\, \left(\Sigma_j + \mu_j\mu_j^T\right) - \left(\sum_{j=1}^{J} \pi_j\, \mu_j\right)\left(\sum_{j=1}^{J} \pi_j\, \mu_j\right)^T \end{aligned} \tag{2.14}$$

Όπως έχει φανεί μέχρι τώρα, η μίξη κανονικών κατανομών αποτελεί μια προσπάθεια να περιγραφεί η κατανομή των παρατηρήσεων χρησιμοποιώντας τοπικά χαρακτηριστικά, ουσιαστικά χρησιμοποιώντας μια τοπική μέση τιμή και έναν τοπικό πίνακα συνδιακύμανσης. Εκμεταλλευόμενοι την ύπαρξη αυτών των διαφορετικών χαρακτηριστικών μπορούμε να οδηγηθούμε σε μια διαμέριση του χώρου. Από το μοντέλο παραγωγής δεδομένων που περιγράψαμε, μπορούμε να συμπεράνουμε ότι παρατηρήσεις που έχουν προέλθει από την ίδια κανονική συνιστώσα θα βρίσκονται κοντά στον χώρο. Πιο συγκεκριμένα θα σχηματίζουν ένα νέφος, που το κέντρο του και η διάχυση γύρω από αυτό θα καθορίζονται από την



Σχήμα 2.4: Μια διδιάστατη μίξη κανονικών κατανομών με δυο συνιστώσες, και ένα σύνολο παρατηρήσεων που παρήγαγε.

μέση τιμή και τον πίνακα συνδιακύμανσης της υπεύθυνης συνιστώσας. Στο Σχήμα 2.4 παρουσιάζεται μια διδιάστατη μίξη με δύο συνιστώσες, και ένα σύνολο δεδομένων που δειγματοληπτήθηκαν από αυτή. Εδώ και όπου αλλού σχεδιάζουμε μια μίξη, δεν σχεδιάζουμε τις ισοϋψείς καμπύλες της μίξης αλλά μια συγκεκριμένη ισοϋψή για κάθε κανονική συνιστώσα, η οποία δείχνει την περιοχή επιρροής της συνιστώσας. Συγκεκριμένα για μια συνιστώσα $p_j(x) = \mathcal{N}(x|\mu_j, \Sigma_j)$ σχεδιάζουμε την ισοϋψή που τέμνει την διεύθυνση κάθε ιδιοδιανύσματος του Σ_j σε δύο σημεία που απέχουν από το μ_j απόσταση διπλάσια από την τετραγωνική ρίζα της αντίστοιχης ιδιοτιμής. Η ισοϋψής αυτή περικλείει το 95.4% της μάζας πιθανότητας της p_j.

Όταν χρησιμοποιούμε την μίξη σε προβλήματα μάθησης χωρίς επίβλεψη για την ομαδοποίηση δεδομένων, αντιστρέφουμε το μοντέλο παραγωγής. Από τα υπάρχοντα δεδομένα μαθαίνουμε τις παραμέτρους μιας μίξης, και στην συνέχεια αντιστοιχίζοντας τα δεδομένα σε συνιστώσες της μίξης αντιστοιχίζουμε μια ομάδα δεδομένων σε κάθε συνιστώσα. Η τυπική διαδικασία μάθησης των παραμέτρων είναι η μεγιστοποίηση της πιθανοφάνειας των δεδομένων χρησιμοποιώντας τον αλγόριθμο EM, και θα την περιγράψουμε αναλυτικά στην επόμενη ενότητα. Αφού μάθουμε τις παραμέτρους της μίξης, για να αναθέσουμε ένα x σε μια συνιστώσα p_j υπολογίζουμε την εκ των υστέρων πιθανότητα $p(Z = j|x)$ να το παρήγαγε η συνιστώσα p_j, δοθέντος του x. Από το θεώρημα του Bayes ισχύει:

$$p(Z = j|x) = \frac{p(x|Z = j)p(Z = j)}{p(x)} = \frac{\pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}{\sum_{j=1}^{J} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)} \tag{2.15}$$

αυτή η πιθανότητα αναφέρεται και ως *υπευθυνότητά* (responsibility) της συνιστώσας, επειδή μετρά τον βαθμό με τον οποίο είναι υπεύθυνη η συνιστώσα για το δεδομένο. Αφού υπολογιστούν όλες οι υπευθυνότητες, το x αντιστοιχίζεται στην ομάδα $\hat{\jmath}$ για την οποία ισχύει

$$\hat{\jmath} = \arg\max_j p(Z = j|x) \tag{2.16}$$

Με αυτό τον τρόπο πετυχαίνουμε να χωρίσουμε τα δεδομένα σε J ομάδες. Μέχρι τώρα κάναμε σιωπηρά την υπόθεση ότι ο αριθμός των ομάδων που ψάχνουμε είναι γνωστός, αν δεν είναι γνωστός τότε το πρόβλημα γίνεται πολύ πιο δύσκολο. Για να το αντιμετωπίσουμε η πιο τυπική προσέγγιση είναι να κάνουμε πολλές ομαδοποιήσεις με διαφορετικό αριθμό ομάδων, και στην συνέχεια να διαλέξουμε την καλύτερη ομαδοποίηση με κάποιο κατάλληλο κριτήριο. Η άλλη προσέγγιση είναι να θεωρήσουμε το J σαν μια παράμετρο που πρέπει να εκτιμηθεί κατά την μάθηση, και να χρησιμοποιήσουμε κάποιον κατάλληλο αλγόριθμο. Σε επόμενο κεφάλαιο θα εμβαθύνουμε στο σημαντικό αυτό πρόβλημα.

Η μίξη κανονικών κατανομών μπορεί να χρησιμοποιηθεί και στη μάθηση με επίβλεψη, για την ταξινόμηση δεδομένων. Σε ένα πρόβλημα που οι παρατηρήσεις ανήκουν σε μια από K γνωστές κατηγορίες μπορούμε να χρησιμοποιήσουμε μια μίξη για να εκτιμήσουμε την κατανομή κάθε κατηγορίας. Το μοντέλο παραγωγής για την προσέγγιση είναι το ακόλουθο:

> Μας δίνονται K σύνολα από J κανονικές κατανομές $M_k = \{p_{k1}, \ldots, p_{kJ}\}$, και ένα σύνολο από δείκτες κατηγορίας $\{1, \ldots, K\}$.

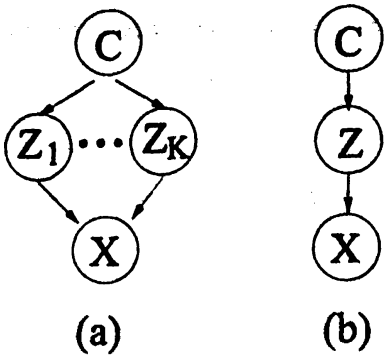
1. Επιλέγουμε τυχαία έναν δείκτη κατηγορίας με πιθανότητα $w_1, \ldots, w_K$ αντίστοιχα, έστω τον k.

2. Επιλέγουμε τυχαία μια από τις κατανομές του συνόλου M_k με πιθανότητα $\pi_1^{(k)}, \ldots, \pi_J^{(k)}$ αντίστοιχα, έστω την p_{kj}.

3. Δειγματοληπτούμε μια παρατήρηση από την επιλεγμένη κατανομή p_{kj}.

Στο Σχήμα 2.5(a) παρουσιάζεται το κατάλληλο γραφικό μοντέλο, όπου η τιμή της C καθορίζει την κατηγορία του προτύπου που παράγεται. Για την από κοινού κατανομή των τυχαίων μεταβλητών ισχύει:

$$p(X, Z_1, \ldots, Z_K, C) = p(X|Z_1, \ldots, Z_K)p(Z_1, \ldots, Z_K|C)p(C) \tag{2.17}$$



Σχήμα 2.5: (a) Το γραφικό μοντέλο μιας μίξης για ταξινόμηση. (b) Το γραφικό μοντέλο ενός δικτύου PRBF.

Αν γνωρίζουμε τις πιθανότητες του μοντέλου παραγωγής $p(C = k) = w_k$, $p(Z_k = j|C = k) = \pi_j^{(k)}$ και την κατανομή $p(X = x|Z_k = j) = p_{kj}(x) = \mathcal{N}(x|\mu_j^{(k)}, \Sigma_j^{(k)})$, τότε μπορούμε να υπολογίσουμε την κατανομή των προτύπων κάθε κατηγορίας περιθωριοποιώντας την Z

$$p(X = x|C = k) = \sum_{Z_k} p(X = x|Z_k)p(Z_k|C = k) = \sum_{j=1}^{J} \pi_j^{(k)} \mathcal{N}(x|\mu_j^{(k)}, \Sigma_j^{(k)}). \quad (2.18)$$

Όπως στην περίπτωση της ομαδοποίησης χρησιμοποιούμε τον αλγόριθμο EM για να μάθουμε τους εκτιμητές μέγιστης πιθανοφάνειας για τις παραμέτρους κάθε μίξης, και ο εκτιμητής μέγιστης πιθανοφάνειας για τις πιθανότητες κάθε κατηγορίας w_k ισούται με τον λόγο του αριθμού των προτύπων της κατηγορίας προς τον αριθμό όλων των προτύπων. Οπότε μπορούμε να χρησιμοποιήσουμε τον κανόνα απόφασης του Bayes για να ταξινομήσουμε ένα νέο πρότυπο άγνωστης κατηγορίας. Στην συνέχεια θα δούμε πως μπορούμε να τροποποιήσουμε κατάλληλα ένα δίκτυο ακτινικών συναρτήσεων βάσης για να καταλήξουμε σε ένα πιο γενικό μοντέλο ταξινόμησης.

2.3.2 Το δίκτυο ακτινικών συναρτήσεων βάσης

Τα δίκτυα ακτινικών συναρτήσεων βάσης (Radial Basis Function, RBF) είναι μια πολύ σημαντική κατηγορία νευρωνικών δικτύων. Μελετήθηκαν για πρώτη φορά στις [66, 73], και στην [10] γίνεται μια ανασκόπηση της εκτενούς ερευνητικής δραστηριότητας που ακολούθησε. Σε ένα δίκτυο RBF για κάθε κρυμμένο κόμβο υπάρχει ένα διάνυσμα που ορίζει την θέση του κόμβου στο χώρο των δεδομένων, και η ενεργοποίηση του κόμβου εξαρτάται από

την απόσταση αυτού του διανύσματος από το διάνυσμα εισόδου. Με αυτό τον τρόπο οι κρυμμένοι κόμβοι ενεργοποιούνται μόνο για πρότυπα που βρίσκονται στην περιοχή επιρροής τους. Το βιολογικό κίνητρο για την ανάπτυξή των RBF ήταν η διαπίστωση παρόμοιας τοπικής δράσης σε νευρώνες του εγκεφάλου. Αποτελούν μια ενεργή ερευνητική περιοχή, και χρησιμοποιούνται εκτενώς για προβλήματα ταξινόμησης και παλινδρόμησης. Παρουσιάζουν μεγάλη ευελιξία κατά την προσέγγιση συναρτήσεων, και έχουν την ιδιότητα της *παγκόσμιας προσέγγισης* (universal approximation) [64] που εγγυάται ότι κάτω από γενικές συνθήκες ένα δίκτυο RBF μπορεί να προσεγγίσει οποιαδήποτε συνάρτηση με αυθαίρετη ακρίβεια χρησιμοποιώντας ικανό αριθμό κόμβων.

Ένα δίκτυο RBF υλοποιεί μια μη-γραμμική απεικόνιση $f : x \in \Re^d \rightarrow f(x) \in \Re^M$. Έχει ένα κρυμμένο επίπεδο με J κόμβους (συναρτήσεις βάσης) και ένα επίπεδο εξόδου με M κόμβους. Η συνάρτηση ενεργοποίησης ϕ_j κάθε κρυμμένου κόμβου είναι μια ακτινική συνάρτηση βάσης $\phi_j(||x - \mu_j||^2)$. Η m-οστή έξοδος του δικτύου δίνεται από την

$$f_m(x) = \sum_{j=1}^{J} a_{mj}\phi_j(||x - \mu_j||^2) \tag{2.19}$$

Η πιο συνηθισμένη μορφή της ακτινικής συνάρτησης βάσης είναι μια Gaussian, $\phi_j(||x-\mu_j||^2) = \mathcal{N}(x|\mu_j, \sigma_j^2\mathrm{I})$. Το μέγεθος της ακτίνας σ_j^2 καθορίζει πόσο μεγάλη είναι η περιοχή του χώρου προτύπων για την οποία ενεργοποιείται ο κόμβος. Προκειμένου να γίνει πιο ευέλικτη η περιγραφή του χώρου ενεργοποίησης, πολλές φορές χρησιμοποιείται η γενική μορφή του πίνακα συνδιακύμανσης Σ_j αντί της περιορισμένης $\sigma_j^2\mathrm{I}$. Το κέντρο μ_j του κόμβου καθορίζει την θέση του στον χώρο των δεδομένων, και αποτελεί μια καθοριστική παράμετρο του μοντέλου. Στις πρώιμες μελέτες του δικτύου χρησιμοποιούνταν ένας κόμβος για κάθε πρότυπο, με κέντρο το πρότυπο. Αυτό είχε σαν αποτέλεσμα την δημιουργία μοντέλων μεγάλης πολυπλοκότητας, και την απαίτηση για απλοποιημένες και γρήγορες μεθόδους εκπαίδευσης. Στην πορεία προτάθηκαν δίκτυα που τα κέντρα τους ήταν ομοιόμορφα κατανεμημένα στον χώρο ή αντιπροσώπευαν ένα υποσύνολο των προτύπων εκπαίδευσης. Για να οδηγηθούμε σε δίκτυα που εκμεταλλεύονται αλγόριθμους ομαδοποίησης, και ο αριθμός των κόμβων τους εξαρτάται από την πολυπλοκότητα της απεικόνισης που επιχειρείται και όχι από το πλήθος των προτύπων.

Οι παράμετροι που πρέπει να εκτιμηθούν κατά την εκπαίδευση του δικτύου είναι τα κέν-

τρα μ_j των συναρτήσεων βάσης, οι παράμετροι που καθορίζουν την ακτίνα τους, π.χ οι διακυμάνσεις σ_j^2 των Gaussian συναρτήσεων, και τα βάρη a_{mj}. Μια γενική κατεύθυνση που ακολουθείται είναι η εκπαίδευση να γίνεται πρώτα για τις παραμέτρους των συναρτήσεων βάσης, και σε δεύτερη φάση να εκτιμούνται τα βάρη. Με αυτό τον τρόπο στην πρώτη φάση μπορεί να εφαρμοστεί μάθηση χωρίς επίβλεψη για την κατάτμηση του χώρου προτύπων. Για παράδειγμα χρησιμοποιώντας τον αλγόριθμο k-κέντρων μπορούν να ομαδοποιηθούν τα πρότυπα, και τα χαρακτηριστικά των ομάδων να καθορίσουν τις παραμέτρους των συναρτήσεων βάσης. Στην συνέχεια εφαρμόζεται μάθηση με επίβλεψη για την εκτίμηση των βαρών. Η τυπική προσέγγιση είναι η ελαχιστοποίηση του τετραγωνικού σφάλματος, δηλαδή το άθροισμα των τετραγώνων των διαφορών μεταξύ της επιθυμητής εξόδου και της εξόδου του δικτύου για κάθε πρότυπο εισόδου.

Για την επίλυση προβλημάτων ταξινόμησης με K κατηγορίες, στη [10] περιγράφεται η χρήση ενός δικτύου RBF που σε κάθε του έξοδο υπολογίζει την πιθανότητας $p(k|x)$ της k-οστής κατηγορίας δοθέντος ενός προτύπου x. Η μέθοδος είναι ανάλογη με την χρήση K μικτών κατανομών, που περιγράψαμε στην προηγούμενη υπο-ενότητα. Ωστόσο δεν χρησιμοποιεί ξεχωριστές συνιστώσες για κάθε μίξη, αλλά θεωρεί ότι για όλες της κατηγορίες χρησιμοποιούνται οι ίδιες κανονικές συνιστώσες. Η εκπαίδευση γίνεται σε δύο στάδια, και στο πρώτο εφαρμόζει τον αλγόριθμο EM σε μια μίξη J κανονικών κατανομών $p(x) = \sum_j \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)$ για την εκτίμηση της κατανομής όλων των προτύπων χωρίς επίβλεψη. Στην συνέχεια χρησιμοποιεί αυτή την μίξη για να ορίσει τις συναρτήσεις βάσης του δικτύου RBF. Το δίκτυο αυτό έχει J κανονικοποιημένες συναρτήσεις βάσης $g_j(x)$, που κάθε μια εκτιμά την πιθανότητα να είναι υπεύθυνη για ένα πρότυπο x δοθέντος του προτύπου, δηλαδή

$$g_j(x) = p(j|x) = \frac{p(j)p(x|j)}{p(x)} \tag{2.20}$$

όπου $p(j) = \pi_j$ και $p(x|j) = \mathcal{N}(x|\mu_j, \Sigma_j)$. Έτσι δημιουργείται ένα δίκτυο όπου η k-οστή του έξοδος είναι

$$f_k(x) = \sum_{j=1}^{J} a_{jk} \frac{\pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}{p(x)} \tag{2.21}$$

και στο δεύτερο στάδιο της εκπαίδευσής του εκτιμούνται τα βάρη a_{kj} χρησιμοποιώντας μια μέθοδο εκπαίδευσης με επίβλεψη.

Για να δείξουμε ότι το δίκτυο υπολογίζει σε κάθε του έξοδο την εκ των υστέρων πιθα-

νότητα μιας κατηγορίας $p(k|x)$ αρκεί να θεωρήσουμε ότι τα βάρη a_{kj} εκτιμούν την εκ των προτέρων πιθανότητα να παραχθούν πρότυπα της κατηγορίας k από την g_j, δηλαδή

$$a_{kj} = p(k|j) = \frac{p(j|k)p(k)}{p(j)} \tag{2.22}$$

όπου $p(k)$ η εκ των προτέρων πιθανότητα της k-οστής κατηγορίας, και $p(j|k)$ η εκ των προτέρων πιθανότητα η j-οστή συνάρτηση βάσης να είναι υπεύθυνη για πρότυπα της k-οστής κατηγορίας δοθείσας της κατηγορίας. Οπότε για την k-οστή έξοδο του δικτύου ισχύει

$$f_k(x) = \sum_j a_{kj} g_j(x) = \sum_j \frac{p(x|j)p(j|k)p(k)}{p(x)} = \frac{p(k)}{p(x)} \sum_j p(x|j)p(j|k). \tag{2.23}$$

Λόγω της υπόθεσης ότι κάθε συνάρτηση βάσης παράγει πρότυπα κάθε κατηγορίας, η κατανομή ενός προτύπου είναι ανεξάρτητη από την κατηγορία του δοθείσας της συνάρτησης βάσης που το παρήγαγε, δηλαδή $p(x|j,k) = p(x|j)$. Αυτό συνεπάγεται ότι $\sum_j p(x|j)p(j|k) = p(x|k)$, και από την (2.23) προκύπτει

$$f_k(x) = \frac{p(x|k)p(k)}{p(x)} = p(k|x). \tag{2.24}$$

Στην συνέχεια της ενότητας θα περιγράψουμε το πιθανοτικό δίκτυο ακτινικών συναρτήσεων βάσης (Probabilistic Radial Basis Function PRBF) [77, 78, 79] το οποίο αποτέλεσε αντικείμενο έρευνας και της παρούσας διατριβής. Το δίκτυο PRBF είναι ένα δίκτυο RBF για ταξινόμηση, που σε κάθε του έξοδο υπολογίζει την κατανομή των προτύπων μιας κατηγορίας $p(x|k)$. Όπως και το δίκτυο στη [10], γίνεται η υπόθεση ότι η κατανομή ενός προτύπου είναι ανεξάρτητη από την κατηγορία του δοθείσας της συνάρτησης βάσης που το παρήγαγε. Επιπλέον επιβάλουμε τον περιορισμό ότι τα βάρη που καταλήγουν σε μια έξοδο είναι μη-αρνητικά και αθροίζουν στην μονάδα. Το PRBF έχει συναρτήσεις βάσης Gaussian, και η k-οστή έξοδος ενός δικτύου με J συναρτήσεις βάσης είναι

$$f_k(x) = \sum_{j=1}^{J} a_{kj} \mathcal{N}(x|\mu_j, \Sigma_j) \tag{2.25}$$

και επιβάλλεται ο περιορισμός

$$\sum_{j=1}^{J} a_{kj} = 1, a_{kj} \geq 0, \text{ για κάθε } j. \tag{2.26}$$

Λόγω της υπό συνθήκης ανεξαρτησίας που υποθέσαμε ισχύει $p(x|j,k) = p(x|j) = \mathcal{N}(x|\mu_j, \Sigma_j)$, και θεωρώντας ότι το βάρος a_{kj} περιγράφει την εκ των προτέρων κατανομή της συνάρτησης βάσης δοθείσας της κατηγορίας, δηλαδή $a_{kj} = p(j|k)$, τότε ισχύει

$$f_k(x) = \sum_{j=1}^{J} p(j|k)p(x|j) = p(x|k). \tag{2.27}$$

Αν διαμερίσουμε το σύνολο $\{p(x|j)|\, j = 1, \ldots, J\}$ των συναρτήσεων βάσης του σε k μη-κενά σύνολα B_k ξένα μεταξύ τους, και περιορίσουμε τα βάρη ώστε $p(j|k) = 0$ αν $p(x|j) \notin B_k$, τότε κάθε συνάρτηση βάσης συνδέεται με μια μόνο έξοδο. Αυτό έχει ως συνέπεια το δίκτυο να μετατρέπεται σε ένα σύστημα k ξεχωριστών μίξεων, όπου κάθε σύνολο B_k περιέχει τις συνιστώσες μιας μίξης.

Το δίκτυο PRBF υποδηλώνει ένα μοντέλο παραγωγής δεδομένων αφού υπολογίζει την κατανομή των προτύπων κάθε κατηγορίας. Πιο συγκεκριμένα τα πρότυπα παράγονται ως εξής:

Μας δίνεται ένα σύνολο από J κανονικές κατανομές $\{p_1, \ldots, p_J\}$, και ένα σύνολο από δείκτες κατηγορίας $\{1, \ldots, K\}$.

1. Επιλέγουμε τυχαία ένα δείκτη κατηγορίας με πιθανότητα $w_1, \ldots, w_K$ αντίστοιχα, έστω τον k.
2. Επιλέγουμε τυχαία μια από τις κανονικές κατανομές με πιθανότητα $\pi_{1k}, \ldots, \pi_{Jk}$ αντίστοιχα, έστω την p_j.
3. Δειγματοληπτούμε μια παρατήρηση από την επιλεγμένη κατανομή p_j.

Το κατάλληλο γραφικό μοντέλο παρουσιάζεται στο Σχήμα 2.5(b), και για την από κοινού κατανομή των τυχαίων μεταβλητών ισχύει

$$p(X, Z, C) = p(X|Z)p(Z|C)p(C). \tag{2.28}$$

Αν γνωρίζουμε τις πιθανότητες του μοντέλου παραγωγής $p(C = k) = w_k$, $p(Z = j|C = k) = \pi_{jk}$ και την κατανομή $p(X = x|Z = j) = p_j(x) = \mathcal{N}(x|\mu_j, \Sigma_j)$, τότε μπορούμε να υπολογίσουμε την κατανομή των προτύπων κάθε κατηγορίας περιθωριοποιώντας την Z

$$p(X = x|C = k) = \sum_{Z} p(X = x|Z)p(Z|C = k) = \sum_{j=1}^{J} \pi_{jk}\mathcal{N}(x|\mu_j, \Sigma_j) \tag{2.29}$$

που αποτελεί την k-οστή έξοδο του δικτύου. Αν σε ένα πρόβλημα ταξινόμησης μας δοθεί ένα σύνολο προτύπων $\{(x_n, y_n) | \, n = 1, \ldots, N\}$, όπου y_n είναι ο δείκτης κατηγορίας του x_n, τότε μπορούμε να υπολογίσουμε τις παραμέτρους του δικτύου μεγιστοποιώντας την λογαριθμική πιθανοφάνεια των προτύπων

$$\log \prod_{n=1}^{N} p(X = x_n, C = y_n) = \sum_{n=1}^{N} \log p(X = x_n | C = y_n) p(C = y_n) \tag{2.30}$$

με τον αλγόριθμο ΕΜ.

2.4 Ο Αλγόριθμος ΕΜ

Έστω ένα σύνολο δεδομένων $X = \{x_n | \, x_n \in \Re^d, n = 1, \ldots, N\}$, την κατανομή του οποίου θέλουμε να εκτιμήσουμε. Η παραμετρική εκτίμησή της προϋποθέτει την επιλογή μιας συναρτησιακής μορφής για την κατανομή, και στη συνέχεια εκτίμηση των παραμέτρων της. Η τυπική επιλογή σε αυτή την περίπτωση είναι η κανονική κατανομή $\mathcal{N}(x|\mu, \Sigma)$, όπου

$$\mathcal{N}(x|\mu, \Sigma) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{d/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\} \tag{2.31}$$

με παραμέτρους το κέντρο $\mu \in \Re^d$ και τον $d \times d$ πίνακα συνδιακύμανσης Σ. Μια τυπική επιλογή για τις παραμέτρους είναι η χρήση της μέσης τιμής και του πίνακα συνδιακύμανσης των δεδομένων. Πρόκειται στην ουσία για την λύση της μέγιστης πιθανοφάνειας, που προκύπτει από την μεγιστοποίηση της πιθανοφάνειας των δεδομένων ως προς τα μ και Σ, για μια απόδειξη δες [3]. Αν υποθέσουμε ότι τα δεδομένα μας είναι ανεξάρτητα και όμοια κατανεμημένα, τότε η λογαριθμική πιθανοφάνειά τους είναι

$$\mathcal{L} = \log \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \Sigma) = \sum_{n=1}^{N} \log \mathcal{N}(x_n|\mu, \Sigma). \tag{2.32}$$

Μεγιστοποιώντας αυτή την ποσότητα, βρίσκουμε ότι οι εκτιμητές μέγιστης πιθανοφάνειας είναι ο δειγματικός μέσος και ο δειγματικός πίνακας συνδιακύμανσης

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{2.33}$$

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)(x_n - \mu)^T. \tag{2.34}$$

Μέχρι τώρα υποθέσαμε ότι οι παρατηρήσεις που βρίσκονται στο σύνολο X έχουν προέλθει από την ίδια κανονική κατανομή. Σε πολλές όμως περιπτώσεις μπορεί να έχουν προέλθει από δύο ή περισσότερες κανονικές κατανομές. Τότε κάθε μια κατανομή λειτουργεί σαν μια ξεχωριστή "πηγή" δεδομένων, και παράγει ένα διαφορετικό υποσύνολο των παρατηρήσεων. Έστω ότι στο X έχουμε δεδομένα που παράχθηκαν από J "πηγές", με κατανομές $\mathcal{N}(x|\mu_1, \Sigma_1), \ldots, \mathcal{N}(x|\mu_J, \Sigma_J)$ αντίστοιχα. Η κατανομή αυτών των δεδομένων είναι μια μίξη κανονικών κατανομών $p(x) = \sum_{j=1}^{J} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)$, όπου κάθε συνιστώσα της μίξης αποτελεί μια διαφορετική "πηγή" δεδομένων, και ο συντελεστής π_j είναι η εκ των προτέρων πιθανότητα ένα οποιοδήποτε δεδομένο να παραχθεί από την j-οστή συνιστώσα. Αν ξέραμε από ποια συνιστώσα έχει παραχθεί κάθε παρατήρηση, τότε θα ήταν πολύ εύκολο να εκτιμήσουμε όλα τα μ_j, Σ_j και π_j $(j = 1, \ldots, J)$. Θα αρκούσε να διαμερίσουμε το X σε J υποσύνολα έτσι ώστε κάθε ένα να αντιστοιχεί σε μια "πηγή" δεδομένων, και στην συνέχεια να υπολογίσουμε τα στατιστικά μεγέθη τους. Ας υποθέσουμε λοιπόν ότι εκτός από το σύνολο X μας δίνεται και η διαμέρισή του με τη μορφή ενός συνόλου $Z = \left\{z_n |\ z_n \in \{0,1\}^J, n = 1, \ldots, N\right\}$, όπου κάθε δυαδικό διάνυσμα z_n αντιστοιχίζει μια παρατήρηση x_n σε μια από τις J "πηγές". Δηλαδή αν για κάποιο z_n ισχύει $z_{jn} = 1$ και $z_{in} = 0$ για κάθε $i \neq j$, τότε το x_n έχει προέλθει από την "πηγή" με κατανομή $\mathcal{N}(x|\mu_j, \Sigma_j)$. Οπότε ανάλογα με τις (2.33) και (2.34) μπορούμε να υπολογίσουμε τις παραμέτρους της μίξης σύμφωνα με τις παρακάτω εξισώσεις, όπου για κάθε j ισχύει

$$\mu_j = \frac{\sum_{n=1}^{N} z_{jn}\, x_n}{\sum_{n=1}^{N} z_{jn}} \tag{2.35}$$

$$\Sigma_j = \frac{\sum_{n=1}^{N} z_{jn}\, (x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^{N} z_{jn}} \tag{2.36}$$

$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} z_{jn}. \tag{2.37}$$

Με αυτό τον τρόπο το κέντρο κάθε συνιστώσας εκτιμάται από τον δειγματικό μέσο των δεδομένων που παρήγαγε, και αντίστοιχα εκτιμάται και ο πίνακας συνδιακύμανσής της. Όσο για τους συντελεστές της μίξης, αφού αποτελούν την πιθανότητα με την οποία η συνιστώσα παράγει κάποιο από τα δεδομένα του X, τους εκτιμούμε ως το ποσοστό του αριθμού των δεδομένων που παρήγαγε η συνιστώσα σε σχέση με το συνολικό πλήθος των δεδομένων στο X.

Στην πράξη όμως το σύνολο Z δεν είναι γνωστό, για αυτό και αναφέρεται σαν το σύνολο των *κρυμμένων* (hidden) μεταβλητών. Αυτό που μπορούμε να κάνουμε είναι να βρούμε τις αναμενόμενες τιμές $\langle z_n \rangle$ των κρυμμένων μεταβλητών. Αν με πιθανότητα $p(j|x_n)$ ισχύει $z_{jn} = 1$, τότε από τον ορισμό της μέσης τιμής προκύπτει ότι $\langle z_{jn} \rangle = p(j|x_n)$. Η εκ των υστέρων πιθανότητα $p(j|x_n)$ με την οποία το x_n έχει παραχθεί από την κατανομή $\mathcal{N}(x|\mu_j, \Sigma_j)$ αναφέρεται και σαν "υπευθυνότητα" (responsibility), αφού όταν ισούται με μονάδα δηλώνει ότι η συγκεκριμένη κατανομή έχει παράγει την παρατήρηση. Χρησιμοποιώντας τις αναμενόμενες τιμές των κρυμμένων μεταβλητών μπορούμε να τροποποιήσουμε τις (2.35), (2.36) και (2.37), και να εκτιμήσουμε τις άγνωστες παραμέτρους σύμφωνα με τις:

$$\mu_j = \frac{\sum_{n=1}^{N} \langle z_{jn} \rangle \, x_n}{\sum_{n=1}^{N} \langle z_{jn} \rangle} \tag{2.38}$$

$$\Sigma_j = \frac{\sum_{n=1}^{N} \langle z_{jn} \rangle \, (x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^{N} \langle z_{jn} \rangle} \tag{2.39}$$

$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} \langle z_{jn} \rangle. \tag{2.40}$$

Τις αναμενόμενες τιμές των κρυμμένων μεταβλητών μπορούμε να τις υπολογίσουμε από το θεώρημα του Bayes, οπότε

$$\langle z_{jn} \rangle = \frac{\pi_j \, \mathcal{N}(x_n|\mu_j, \Sigma_j)}{\sum_{j=1}^{J} \pi_j \, \mathcal{N}(x_n|\mu_j, \Sigma_j)}. \tag{2.41}$$

Αυτές όμως εξαρτώνται από τις παραμέτρους της μίξης που θέλουμε να εκτιμήσουμε. Μπορούμε λοιπόν να υιοθετήσουμε μια επαναληπτική προσέγγιση, όπου εκτιμούμε εναλλάξ τις παραμέτρους της μίξης με τις (2.38–2.40) και τις αναμενόμενες τιμές των κρυμμένων μεταβλητών με την (2.41). Ξεκινώντας από κάποιες αρχικές τιμές, μπορούμε να εγγυηθούμε ότι αυτή η διαδικασία θα συγκλίνει στους εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων. Πρόκειται στην ουσία για εφαρμογή του αλγορίθμου EM, τον οποίο θα παρουσιάσουμε πιο διεξοδικά στην συνέχεια.

Μέχρι τώρα είδαμε πώς μπορεί να προκύψει διαισθητικά ο EM για την εκτίμηση των παραμέτρων μιας μίξης. Γενικά όμως ο αλγόριθμος EM εφαρμόζεται για την εύρεση εκτιμητών μέγιστης πιθανοφάνειας, όταν εκτός από το σύνολο των *παρατηρούμενων* τυχαίων μεταβλητών X υπάρχει και ένα σύνολο από *κρυμμένες* τυχαίες μεταβλητές Z που συμμετέχουν στην εκτίμηση των παραμέτρων θ της πιθανοφάνειας. Η εκτίμηση των παραμέτρων

μιας μίξης είναι το πιο χαρακτηριστικό πρόβλημα που επιλύεται με τον ΕΜ αλλά όχι το μόνο. Μια άλλη ενδιαφέρουσα εφαρμογή του είναι για την συμπλήρωση των ελλιπών τιμών ενός συνόλου παρατηρήσεων. Σε αυτή την περίπτωση θεωρείται ότι τα παρατηρούμενα διανύσματα ακολουθούν μια κανονική κατανομή, αλλά κάποιες από τις τιμές των διανυσμάτων λείπουν. Χρησιμοποιώντας τον ΕΜ μπορούμε να εκτιμήσουμε τις παραμέτρους της κατανομής καθώς και τις τιμές που λείπουν. Ο ΕΜ παρουσιάστηκε και μελετήθηκε διεξοδικά στην [26], αν και σαν ιδέα προϋπήρχε και είχε χρησιμοποιηθεί διαισθητικά και σε άλλες εργασίες. Ωστόσο στην [26] ορίστηκε χρησιμοποιώντας την έννοια της κρυμμένης μεταβλητής, και αποδείχθηκε η πιο σημαντική ιδιότητά του, ότι σε κάθε του επανάληψη η λογαριθμική πιθανοφάνεια δεν μειώνεται. Στην [67] γίνεται μια σύντομη ανασκόπηση της προσέγγισης της μέγιστης πιθανοφάνειας για μίξη κατανομών, και αναπτύσσονται οι θεωρητικές και πρακτικές ιδιότητες του ΕΜ για μίξη κατανομών. Στην [57] γίνεται μια εκτεταμένη παρουσίαση και μελέτη των ιδιοτήτων του ΕΜ, και των παραλλαγών του που έχουν κατά καιρούς παρουσιαστεί στην βιβλιογραφία.

Το βασικό πρόβλημα που αντιμετωπίζουμε είναι η μεγιστοποίηση της λογαριθμικής πιθανοφάνειας των παρατηρούμενων δεδομένων $\mathcal{L} = \log p(X|\theta)$ ως προς το διάνυσμα παραμέτρων θ της κατανομής τους. Στην [67] παρουσιάζεται ένα θεώρημα το οποίο ισχυρίζεται ότι, κάτω από λογικές υποθέσεις, αν θέσουμε τις μερικές παραγώγους της λογαριθμικής πιθανοφάνειας ίσες με το μηδέν, και λύσουμε τις εξισώσεις που προκύπτουν, τότε οι εξισώσεις $\nabla_\theta \mathcal{L} = 0$ έχουν μοναδική λύση, και αυτή μεγιστοποιεί τοπικά την λογαριθμική πιθανοφάνεια. Όταν μας ενδιαφέρει και η εκτίμηση των κρυμμένων Z ή όταν αυτά καθορίζουν την λύση, ακόμα και αν δεν μας ενδιαφέρουν άμεσα, τότε καταφεύγουμε στον ΕΜ. Σε αυτή την περίπτωση απαιτείται ο ορισμός της από κοινού κατανομής $p(X, Z|\theta)$ των παρατηρήσεων X και των κρυμμένων μεταβλητών Z, η οποία συνήθως αναφέρεται ως η *πλήρης πιθανοφάνεια* (complete likelihood) των δεδομένων.

Ξεκινώντας από μια αρχική εκτίμηση θ_0, στην t-οστή επανάληψη ο αλγόριθμος ορίζει μια μετάβαση από το τρέχον διάνυσμα παραμέτρων θ_t σε ένα νέο διάνυσμα θ_{t+1}. Σε κάθε επανάληψή του εφαρμόζουμε δυο βήματα

- *Expectation-βήμα*: Ορίζουμε μια κατανομή επί των κρυμμένων μεταβλητών $q(Z) = p(Z|X, \theta_t)$ δοθέντων των παρατηρήσεων και της τρέχουσας εκτίμησης των παραμέ-

τρων, και υπολογίζουμε την αναμενόμενη τιμή ως προς q της πλήρους πιθανοφάνειας

$$\mathcal{Q}(\theta, \theta_t) = \langle \log p(X, Z|\theta) \rangle_q = \int q(Z) \log p(X, Z|\theta) \, \mathrm{d}Z \tag{2.42}$$

- *Maximization-βήμα*: Θέτουμε το νέο διάνυσμα παραμέτρων $\theta_{t+1} = \arg\max_\theta \mathcal{Q}(\theta, \theta_t)$.

Αν επιστρέψουμε στην εκτίμηση των παραμέτρων της μίξης, όπως την εισαγάγαμε διαισθητικά στην αρχή της ενότητας, μπορούμε να διαπιστώσουμε ότι το Ε-βήμα αντιστοιχεί στην εκτίμηση της μέσης τιμής των κρυμμένων μεταβλητών μέσω της (2.41), και το Μ-βήμα στην εκτίμηση των παραμέτρων τις μίξης μέσω των (2.38), (2.39) και (2.40).

Στην συνέχεια θα δώσουμε μια απλή απόδειξη για την θεμελιώδη ιδιότητα του αλγορίθμου να μην μειώνει την λογαριθμική πιθανοφάνεια, έτσι ώστε σε κάθε επανάληψή του να ισχύει $\mathcal{L}(\theta_{t+1}) \geq \mathcal{L}(\theta_t)$. Από το θεώρημα του Bayes, για την υπό συνθήκη κατανομή έχουμε

$$p(Z|X, \theta_{t+1}) = \frac{p(X, Z|\theta_{t+1})}{p(X|\theta_{t+1})} \tag{2.43}$$

$$\log p(Z|X, \theta_{t+1}) = \log p(X, Z|\theta_{t+1}) - \mathcal{L}(\theta_{t+1}). \tag{2.44}$$

Μπορούμε στην συνέχεια να υπολογίσουμε την μέση τιμή των δυο μελών τις εξίσωσης ως προς την κατανομή $q(Z) = p(Z|X, \theta_t)$

$$\langle \log p(Z|X, \theta_{t+1}) \rangle_q = \langle \log p(X, Z|\theta_{t+1}) \rangle_q - \mathcal{L}(\theta_{t+1}). \tag{2.45}$$

Παρατηρήστε ότι $\langle \mathcal{L}(\theta_{t+1}) \rangle_q = \mathcal{L}(\theta_{t+1})$ καθώς η $\mathcal{L}(\theta_{t+1}) = \log p(X|\theta_{t+1})$ δεν εξαρτάται από τις Z. Ομοίως μπορούμε να αποδείξουμε ότι

$$\langle \log p(Z|X, \theta_t) \rangle_q = \langle \log p(X, Z|\theta_t) \rangle_q - \mathcal{L}(\theta_t). \tag{2.46}$$

Αφαιρώντας κατά μέλη τις δύο προηγούμενες εξισώσεις και αναδιατάσσοντας τους όρους τους παίρνουμε

$$\begin{aligned} \mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) &= \langle \log p(X, Z|\theta_{t+1}) \rangle_q - \langle \log p(X, Z|\theta_t) \rangle_q \\ &- \langle \log p(Z|X, \theta_{t+1}) \rangle_q + \langle \log p(Z|X, \theta_t) \rangle_q. \end{aligned} \tag{2.47}$$

Από τον ορισμό του Μ-βήματος ισχύει ότι $\langle \log p(X, Z|\theta_{t+1}) \rangle_q - \langle \log p(X, Z|\theta_t) \rangle_q \geq 0$. Όσο αφορά την δεύτερη διαφορά στο δεξιό μέλος, παρατηρούμε ότι πρόκειται για την απόσταση

Kullback-Liebler μεταξύ των κατανομών $p(Z|X,\theta_{t+1})$ και $p(Z|X,\theta_t)$. Γενικά η απόσταση Kullback-Liebler μεταξύ δυο κατανομών $p_1(x)$ και $p_2(x)$ ορίζεται ως

$$\mathrm{KL}(p_1||p_2) = -\int p_1(x)\log\frac{p_2(x)}{p_1(x)}\,\mathrm{d}x. \tag{2.48}$$

Εφαρμόζοντας την ανισότητα του Jensen μπορούμε να δείξουμε ότι $\mathrm{KL}(p_1||p_2) \geq 0$, και με αντικατάσταση στην (2.48) βλέπουμε ότι η ισότητα ισχύει αν $p_1(x) = p_2(x)$. Ομοίως $-\langle\log p(Z|X,\theta_{t+1})\rangle_q + \langle\log p(Z|X,\theta_t)\rangle_q \geq 0$, και έτσι καταλήγουμε στο συμπέρασμα ότι ισχύει $\mathcal{L}(\theta_{t+1}) \geq \mathcal{L}(\theta_t)$ σε κάθε επανάληψη του EM. Επιπλέον έχει αποδειχθεί ότι κάτω από πολύ γενικές συνθήκες η λογαριθμική πιθανοφάνεια αυξάνεται σε κάθε επανάληψη, μέχρι να συγκλίνει σε κάποιο τοπικό μέγιστο. Πιο συγκεκριμένα στην [83] αποδεικνύεται ότι αν η $\mathcal{Q}(\theta',\theta)$ είναι συνεχής ως προς θ' και θ, τότε όλα τα σημεία σύγκλισης ενός αλγόριθμου EM αποτελούν στάσιμα σημεία της $\mathcal{L}$, και η ακολουθία των $\mathcal{L}$ που παράγει ο EM συγκλίνει μονότονα στο $\mathcal{L}^* = \mathcal{L}(\theta^*)$ για κάποιο στάσιμο σημείο θ^* της λογαριθμικής πιθανοφάνειας. Επιπλέον, αν κανένα σαγματικό σημείο της $\mathcal{L}$ δεν είναι είναι ολικό μέγιστο της $\mathcal{Q}$, τότε η $\mathcal{L}$ θα συγκλίνει σε κάποιο σημείο μεγίστου. Στην πράξη είναι πολύ δύσκολο να προβλέψουμε την σύγκλιση σε σαγματικό σημείο, ωστόσο αυτή εμφανίζεται σπάνια και σε εκφυλισμένες περιπτώσεις. Μια πιο εκτενής περιγραφή των ιδιοτήτων σύγκλισης του EM περιέχεται στην [57].

Όσο αφορά την ταχύτητα σύγκλισης ο EM θεωρείται γενικά αργός, και στην πράξη η σύγκλισή του στις τελευταίες επαναλήψεις καθυστερεί. Η πρώτη μελέτη της ταχύτητάς του έχει γίνει την [26], ενώ στην [41] η μελέτη έχει επεκταθεί σε μια ευρύτερη κλάση αλγορίθμων που περιλαμβάνει και προσεγγιστικές παραλλαγές του EM, και έχει αποδειχθεί ότι η νόρμα $||\theta_{t+1}-\theta^*||$ συγκλίνει μονότονα στο μηδέν με τουλάχιστον γραμμική ταχύτητα σύγκλισης, δηλαδή

$$||\theta_{t+1} - \theta^*|| \leq \alpha\,||\theta_t - \theta^*||, \text{ με } 0 \leq \alpha < 1$$

όπου θ^* το σημείο σύγκλισης των θ_t. Για να επιταχυνθεί η σύγκλισή του έχουν προταθεί πολλές μέθοδοι. Ιδιαίτερο ενδιαφέρον έχει η μέθοδος που έχει προταθεί στην [16], και εκμεταλλεύεται την σχέση μεταξύ του EM και των Proximal Point (PP) αλγορίθμων. Το βασικό κίνητρο για τους PP είναι ότι προσθέτοντας στην αντικειμενική συνάρτηση μια ακολουθία από ποινές που εξαρτώνται από την τρέχουσα επανάληψη (καλούνται proximal penalties)

παίρνουμε ευσταθείς επαναληπτικούς αλγορίθμους, οι οποίοι συχνά έχουν επιδόσεις που ξεπερνούν τις τυπικές μεθόδους βελτιστοποίησης χωρίς τις ποινές. Έτσι στην [16] έχει αποδειχθεί ότι επιλέγοντας έναν κατάλληλο όρο ποινής που προστίθεται στο Μ-βήμα του ΕΜ, ο τροποποιημένος αλγόριθμος έχει ταχύτητα σύγκλισης υπερ-γραμμική, με την έννοια ότι

$$\lim_{t\to\infty} \frac{||\theta_{t+1} - \theta^*||}{||\theta_t - \theta^*||} = 0.$$

Επιπλέον ο τροποποιημένος αλγόριθμος διατηρεί την ιδιότητα της μονότονης αύξησης της πιθανοφάνειας.

Μετά την ολοκλήρωση της περιγραφής του ΕΜ και των ιδιοτήτων του θα επιχειρήσουμε να αναδείξουμε την μεγάλη του χρησιμότητα για την εύρεση εκτιμητών μέγιστης πιθανοφάνειας, εκτιμητών μέγιστης εκ των υστέρων πιθανότητας ή εκτιμητών μέγιστης πιθανοφάνειας με ποινή. Καταρχήν η διατύπωση του ΕΜ είναι πολύ γενική, ώστε τα βήματά του να αποτελούν στην ουσία μια κατευθυντήρια γραμμή για την επίλυση του εκάστοτε προβλήματος μεγιστοποίησης. Αυτό του επιτρέπει να βρίσκει εφαρμογή σε πολλά προβλήματα διαφόρων ερευνητικών περιοχών, όπως της στατιστικής, της μηχανικής μάθησης, της επεξεργασίας σήματος και εικόνας, και της επεξεργασίας φυσικής γλώσσας. Το χαρακτηριστικό αυτών των προβλημάτων είναι η ύπαρξη των κρυμμένων μεταβλητών, οι οποίες καθορίζουν την λύση του προβλήματος είτε αποτελούν το ζητούμενο είτε όχι. Για παράδειγμα όταν εφαρμόζουμε τον ΕΜ για να εκτιμήσουμε την κατανομή ενός συνόλου δεδομένων με μια μίξη κανονικών κατανομών δεν μας ενδιαφέρουν άμεσα οι κρυμμένες μεταβλητές, αλλά επικεντρωνόμαστε στην προσέγγιση της συνάρτησης. Αντίθετα αν με την μίξη σκοπεύουμε να ομαδοποιήσουμε ένα σύνολο δεδομένων, τότε μας ενδιαφέρουν κυρίως οι κρυμμένες μεταβλητές οι οποίες και θα καθορίσουν τον διαχωρισμό των δεδομένων. Όμως πάντα η ύπαρξη των κρυμμένων μεταβλητών είναι καθοριστική, αφού απλοποιούν την επίλυση των εξισώσεων και προσφέρουν μια καλή διαισθητική εξήγηση της διαδικασίας βελτιστοποίησης. Αν αντί του ΕΜ ακολουθούσαμε μια εναλλακτική προσέγγιση, θα μπορούσαμε να ορίσουμε την συνάρτηση πιθανοφάνειας των δεδομένων μας και να λύσουμε τις εξισώσεις που προκύπτουν μηδενίζοντας τις μερικές παραγώγους της πιθανοφάνειας. Έτσι όμως είτε θα βρίσκαμε εκτιμητές μόνο για τις παραμέτρους του μοντέλου και όχι για τις κρυμμένες μεταβλητές, είτε θα καταλήγαμε και πάλι στις εξισώσεις του ΕΜ κάνοντας κάποιες υποθέσεις και αλγεβρικούς

χειρισμούς που στον ΕΜ γίνονται αυθόρμητα. Εναλλακτικά θα μπορούσαμε να ορίσουμε την πλήρη πιθανοφάνεια των δεδομένων και των χρυμμένων μεταβλητών, και να λύσουμε πάλι της εξισώσεις που προκύπτουν από τον μηδενισμό των μερικών παραγώγων. Έτσι όμως δεν θα μπορούσαμε να εγγυηθούμε εύκολα την μονότονη αύξηση της πιθανοφάνειας, και οι εκτιμητές που θα παίρναμε δεν θα ήταν εκτιμητές μέγιστης πιθανοφάνειας. Οπότε είναι πιο βολικό να μεγιστοποιούμε την πλήρη πιθανοφάνεια και να εκτιμούμε την μέση τιμή των χρυμμένων μεταβλητών επαναληπτικά, δηλαδή να εφαρμόσουμε τον ΕΜ.

2.5 Ο Αλγόριθμος ΕΜ για Μίξη Κανονικών Κατανομών

Στην παρούσα ενότητα θα εκτιμήσουμε και πάλι τις παραμέτρους μιας μίξης κανονικών κατανομών δοθέντος ενός συνόλου δεδομένων $X = \{x_n | n = 1, \ldots, N\}$ όπου $x_n \in \Re^d$, αυτή την φορά ακολουθώντας πιστά τον αλγόριθμο ΕΜ. Το διάνυσμα των παραμέτρων θ που θα εκτιμήσουμε αποτελείται από τους συντελεστές της μίξης και της παραμέτρους των συνιστωσών της: $\theta = \{\pi_j, \mu_j, \Sigma_j | j = 1, \ldots, J\}$. Ξεκινώντας από το Ε-βήμα θα ορίσουμε την κατανομή q των χρυμμένων μεταβλητών και θα υπολογίσουμε την αναμενόμενη τιμή της λογαριθμικής πλήρους πιθανοφάνειας $\langle \log p(X, Z|\theta) \rangle_q$. Όπως ήδη αναφέραμε, σε κάθε $x_n \in X$ αντιστοιχεί μια χρυμμένη μεταβλητή $z_n \in Z$ που καθορίζει ποια συνιστώσα έχει παράγει την παρατήρηση x_n. Πρόκειται για ένα δυαδικό διάνυσμα με J στοιχεία z_{jn}, για τα οποία είναι ισχύει $\sum_{j=1}^{J} z_{jn} = 1$ για κάθε n. Αφού το z_{jn} είναι μονάδα με πιθανότητα π_j και $\sum_{j=1}^{J} \pi_j = 1$, τότε τα στοιχεία του z_n ακολουθούν πολυωνυμική κατανομή, και η $p(Z|\theta)$ είναι το γινόμενο N πολυωνυμικών κατανομών

$$p(Z|\theta) = \prod_{n=1}^{N} \prod_{j=1}^{J} \pi_j^{z_{jn}} \tag{2.49}$$

υποθέτοντας ανεξάρτητες και όμοια κατανεμημένες χρυμμένες μεταβλητές. Επιπλέον δοθέντος του z_n, γνωρίζουμε ότι το x_n έχει παραχθεί από κάποια συγκεκριμένη συνιστώσα και ακολουθεί κανονική κατανομή, οπότε ισχύει

$$p(X|Z, \theta) = \prod_{n=1}^{N} \prod_{j=1}^{J} \mathcal{N}(x_n|\mu_j, \Sigma_j)^{z_{jn}} \tag{2.50}$$

υποθέτοντας ανεξάρτητες και όμοια κατανεμημένες παρατηρήσεις. Οπότε για την αναμενόμενη τιμή της λογαριθμικής πλήρους πιθανοφάνειας ισχύει

$$\begin{aligned}\mathcal{Q}(\theta,\theta_t) &= \langle \log p(X,Z|\theta)\rangle_q \\ &= \langle \log p(X|Z,\theta)\rangle_q + \langle \log p(Z|\theta)\rangle_q \\ &= \sum_{n=1}^{N}\sum_{j=1}^{J}\langle z_{jn}\rangle_q \log \pi_j + \sum_{n=1}^{N}\sum_{j=1}^{J}\langle z_{jn}\rangle_q \log \mathcal{N}(x_n|\mu_j,\Sigma_j). \end{aligned} \tag{2.51}$$

Για να συνεχίσουμε πρέπει να ορίσουμε την κατανομή $q = p(Z|X,\theta_t)$ και να υπολογίσουμε τις αναμενόμενες τιμές ως προς αυτή. Για την q ισχύει από το θεώρημα του Bayes

$$\begin{aligned} q(Z) &= \frac{p(X|Z,\theta_t)\,p(Z|\theta_t)}{p(X|\theta_t)} \\ &= \prod_{n=1}^{N}\frac{\prod_{j=1}^{J}\left[\pi_j^{(t)}\,\mathcal{N}(x_n|\mu_j^{(t)},\Sigma_j^{(t)})\right]^{z_{jn}}}{\sum_{j=1}^{J}\pi_j^{(t)}\,\mathcal{N}(x_n|\mu_j^{(t)},\Sigma_j^{(t)})} \\ &= \prod_{n=1}^{N}\prod_{j=1}^{J}\left[\frac{\pi_j^{(t)}\,\mathcal{N}(x_n|\mu_j^{(t)},\Sigma_j^{(t)})}{\sum_{j=1}^{J}\pi_j^{(t)}\,\mathcal{N}(x_n|\mu_j^{(t)},\Sigma_j^{(t)})}\right]^{z_{jn}} \end{aligned} \tag{2.52}$$

όπου η τελευταία ισότητα ισχύει λόγω του περιορισμού $\sum_{j=1}^{J} z_{jn} = 1$. Επομένως η $q(Z)$ είναι το γινόμενο N ανεξάρτητων πολυωνυμικών κατανομών, και για τις αναμενόμενες τιμές των χρυμμένων μεταβλητών ισχύει

$$\langle z_{jn}\rangle_q = \frac{\pi_j\,\mathcal{N}(x_n|\mu_j,\Sigma_j)}{\sum_{j=1}^{J}\pi_j\,\mathcal{N}(x_n|\mu_j,\Sigma_j)} \tag{2.53}$$

Έτσι ολοκληρώσαμε του υπολογισμούς που απαιτούνται στο Ε-βήμα του αλγορίθμου.

Στο Μ-βήμα πρέπει να μεγιστοποιήσουμε την αναμενόμενη τιμή της λογαριθμικής πλήρους πιθανοφάνειας ως προς τις παραμέτρους της μίξης. Από την (2.51) βλέπουμε ότι μπορούμε να μεγιστοποιήσουμε την $\mathcal{Q}$ ως προς τους συντελεστές της μίξης ανεξάρτητα από τις παραμέτρους των Gaussian. Αρχίζουμε παραγωγίζοντας την $\mathcal{Q}(\theta,\theta_t)$ ως προς τους συντελεστές της μίξης π_j, και στην συνέχεια θα μηδενίσουμε την μερική παράγωγο. Όμως για τους συντελεστές πρέπει να ισχύει $\sum_{j=1}^{J}\pi_j = 1$, και για να ικανοποιηθεί ο περιορισμός χρησιμοποιούμε έναν πολλαπλασιαστή Lagrange λ, οπότε λόγω της (2.51) παίρνουμε

$$\begin{aligned}\frac{\partial}{\partial \pi_j}\left[\mathcal{Q}(\theta,\theta_t) + \lambda\left(\sum_{j=1}^{J}\pi_j - 1\right)\right] &= \sum_{n=1}^{N}\langle z_{jn}\rangle_q \frac{\partial}{\partial \pi_j}\log \pi_j + \lambda \\ &= \sum_{n=1}^{N}\langle z_{jn}\rangle_q \frac{1}{\pi_j} + \lambda \end{aligned} \tag{2.54}$$

Θέτοντας το αποτέλεσμα της παραγώγισης ίσο με το μηδέν έχουμε

$$\begin{aligned}\sum_{n=1}^{N}\langle z_{jn}\rangle_q \frac{1}{\pi_j} + \lambda &= 0 \\ \sum_{n=1}^{N}\langle z_{jn}\rangle_q + \lambda\,\pi_j &= 0 \end{aligned} \tag{2.55}$$

Επαναλαμβάνοντας την διαδικασία για κάθε συντελεστή π_j καταλήγουμε στην (2.55) για κάθε τιμή του $j = 1, \ldots, J$. Αθροίζοντας λοιπόν την (2.55) ως προς j παίρνουμε

$$\sum_{j=1}^{J}\sum_{n=1}^{N}\langle z_{jn}\rangle_q + \lambda\sum_{j=1}^{J}\pi_j = 0 \tag{2.56}$$

και λόγω του περιορισμού στους συντελεστές ισχύει

$$\lambda = -\sum_{j=1}^{J}\sum_{n=1}^{N}\langle z_{jn}\rangle_q = -N \tag{2.57}$$

λόγω της (2.53). Αντικαθιστώντας το λ στην (2.55) παίρνουμε την εξίσωση ανανέωσης για τους συντελεστές της μίξης

$$\pi_j^{(t+1)} = \frac{1}{N}\sum_{n=1}^{N}\langle z_{jn}\rangle_q \tag{2.58}$$

Συνεχίζουμε παραγωγίζοντας την $\mathcal{Q}(\theta, \theta_t)$ ως προς το κέντρο μ_j της j-οστής συνιστώσας, και μηδενίζοντας την μερική παράγωγο. Από την (2.51) ισχύει

$$\frac{\partial}{\partial \mu_j}\mathcal{Q}(\theta, \theta_t) = \sum_{n=1}^{N}\langle z_{jn}\rangle_q \frac{\partial}{\partial \mu_j}\log\mathcal{N}(x_n|\mu_j, \Sigma_j) \tag{2.59}$$

Η μερική παράγωγος του λογαρίθμου της Gaussian είναι

$$\begin{aligned}\frac{\partial}{\partial \mu_j}\log\mathcal{N}(x_n|\mu_j, \Sigma_j) &= \frac{\partial}{\partial \mu_j}\log\frac{|\Sigma_j|^{-1/2}}{(2\pi)^{d/2}} - \frac{1}{2}\frac{\partial}{\partial \mu_j}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j) \\ &= \Sigma_j^{-1}(x-\mu_j)\end{aligned} \tag{2.60}$$

Η νέα εκτίμηση για το κέντρο $\mu_j^{(t+1)}$ της j-οστής συνιστώσας βρίσκεται μηδενίζοντας την

αντίστοιχη μερική παράγωγο τις $\mathcal{Q}$, οπότε έχουμε

$$\begin{aligned}
\frac{\partial}{\partial \mu_j}\mathcal{Q}(\theta,\theta_t)\Big|_{\mu_j^{(t+1)}} &= 0 \\
\sum_{n=1}^{N}\langle z_{jn}\rangle_q \Sigma_j^{-1}(x-\mu_j^{(t+1)}) &= 0 \\
\sum_{n=1}^{N}\langle z_{jn}\rangle_q \mu_j^{(t+1)} &= \sum_{n=1}^{N}\langle z_{jn}\rangle_q x \\
\mu_j^{(t+1)} &= \frac{\sum_{n=1}^{N}\langle z_{jn}\rangle_q x}{\sum_{n=1}^{N}\langle z_{jn}\rangle_q}
\end{aligned} \tag{2.61}$$

όπου ο πίνακας Σ_j^{-1} απαλείφθηκε από την δεύτερη ισότητα πολλαπλασιάζοντας και τα δύο μέλη της από αριστερά με τον Σ_j.

Ολοκληρώνουμε το M-βήμα του EM παραγωγίζοντας την $\mathcal{Q}(\theta,\theta_t)$ ως προς τον πίνακα συνδιακύμανσης Σ_j της j-οστής συνιστώσας, και μηδενίζοντας την μερική παράγωγο. Από την (2.51) ισχύει

$$\frac{\partial}{\partial \Sigma_j}\mathcal{Q}(\theta,\theta_t) = \sum_{n=1}^{N}\langle z_{jn}\rangle_q \frac{\partial}{\partial \Sigma_j}\log \mathcal{N}(x_n|\mu_j,\Sigma_j) \tag{2.62}$$

Για να υπολογίσουμε την μερική παράγωγο του λογαρίθμου της Gaussian πρέπει να λάβουμε υπόψη μας τον περιορισμό για συμμετρικό και θετικά ορισμένο πίνακα συνδιακύμανσης. Για αυτό γράφουμε $\Sigma_j^{-1} = S^T S$, χρησιμοποιώντας τον τετραγωνικό πίνακα S, και χρησιμοποιούμε δυο γνωστά αποτελέσματα της παραγώγισης ως προς πίνακα:

$$\frac{\partial}{\partial S}\, a^T S^T S\, b = S(ab^T + ba^T) \tag{2.63}$$

$$\frac{\partial}{\partial S}\, log|S^T S| = 2S(S^T S)^{-1} \tag{2.64}$$

όπου ο S είναι ένας $d\times d$ πίνακας και τα a, b είναι $d\times 1$ διανύσματα. Οπότε για την παράγωγο της Gaussian ισχύει

$$\begin{aligned}
\frac{\partial}{\partial S}\log \mathcal{N}(x_n|\mu_j,(S^T S)^{-1}) &= \frac{\partial}{\partial \Sigma_j}\log\frac{|S^T S|^{1/2}}{(2\pi)^{d/2}} - \frac{1}{2}\frac{\partial}{\partial S}(x-\mu_j)^T S^T S(x-\mu_j) \\
&= S(S^T S)^{-1} - S(x-\mu_j)(x-\mu_j)^T
\end{aligned} \tag{2.65}$$

Στην συνέχεια μηδενίζουμε την μερική παράγωγο της $Q(\theta, \theta_t)$ ως προς S, και παίρνουμε

$$\begin{aligned} \frac{\partial}{\partial S} Q(\theta, \theta_t)|_{\hat{S}} &= 0 \\ \sum_{n=1}^{N} \langle z_{jn} \rangle_q \hat{S} (\hat{S}^T \hat{S})^{-1} &= \sum_{n=1}^{N} \langle z_{jn} \rangle_q \hat{S} (x - \mu_j)(x - \mu_j)^T \\ (\hat{S}^T \hat{S})^{-1} &= \frac{\sum_{n=1}^{N} \langle z_{jn} \rangle_q (x - \mu_j)(x - \mu_j)^T}{\sum_{n=1}^{N} \langle z_{jn} \rangle_q} \end{aligned} \tag{2.66}$$

Επομένως από τον ορισμό του S προκύπτει ότι η νέα εκτίμηση για τον πίνακα συνδιακύμανσης $\Sigma_j^{(t+1)}$ της j-οστής συνιστώσας είναι

$$\Sigma_j^{(t+1)} = \frac{\sum_{n=1}^{N} \langle z_{jn} \rangle_q (x - \mu_j)(x - \mu_j)^T}{\sum_{n=1}^{N} \langle z_{jn} \rangle_q} \tag{2.67}$$

Αν και κατά την βελτιστοποίηση λάβαμε υπόψη μας την απαίτηση για συμμετρικό και θετικά ορισμένο πίνακα συνδιακύμανσης, στην πράξη όταν εφαρμόζουμε τον αλγόριθμο εξαρτόμαστε και από τα δεδομένα στο X και την αρχικοποίηση των παραμέτρων για να την ικανοποιήσουμε. Γενικά πρέπει μια συνιστώσα να είναι υπεύθυνη για πολλά δεδομένα προκειμένου να εξασφαλίσουμε ότι η (2.67) δίνει ένα θετικά ορισμένο πίνακα. Τυπικά για δεδομένα σε ένα d-διάστατο χώρο, πρέπει η μέση τιμή των χρυμμένων μεταβλητών να είναι μεγαλύτερη του μηδενός για τουλάχιστον $d+1$ δεδομένα τα οποία δεν θα ανήκουν στο ίδιο υπερ-επίπεδο διάστασης $d-1$. Για παράδειγμα στον διδιάστατο χώρο χρειαζόμαστε τουλάχιστον τρία μη συνευθειακά σημεία για να ορίσουμε έναν θετικά ορισμένο πίνακα, διαφορετικά η μια ιδιοτιμή του θα είναι μηδενική. Η πιο συνηθισμένη μορφή του προβλήματος είναι όταν μια συνιστώσα j έχει ως κέντρο ένα δεδομένο $x_{n'}$ και είναι ταυτόχρονα υπεύθυνη μόνο για αυτό, δηλαδή $\mu_j = x_{n'}$ και το $\langle z_{jn} \rangle_q$ τείνει στο μηδέν για κάθε $n \neq n'$. Τότε ο πίνακας συνδιακύμανσης της συνιστώσας συρρικνώνεται συνεχώς, καθώς όλες του οι ιδιοτιμές τείνουν στο μηδέν. Αυτό αποτελεί και μια απόδειξη για το ότι η πιθανοφάνεια των δεδομένων δεν είναι άνω φραγμένη όσο αφορά τους πίνακες συνδιακύμανσης της μίξης, αφού η πιθανοφάνεια του συγκεκριμένου δεδομένου αυξάνεται απεριόριστα καθώς ο πίνακας συνδιακύμανσης συρρικνώνεται. Ο πιο απλός τρόπος για να αντιμετωπίσουμε το πρόβλημα είναι να ξεκινήσουμε πάλι τον αλγόριθμο από διαφορετικές αρχικές τιμές, ελπίζοντας ότι θα μας οδηγήσει σε διαφορετικό τοπικό μέγιστο. Μία πιο πρακτική λύση είναι να θέσουμε ένα κάτω φράγμα για τις ιδιοτιμές των πινάκων, και αν σε κάποια επανάληψη του αλγορίθμου η νέα εκτίμηση κάποιου πίνακα παραβιάζει το φράγμα να παραλείπουμε την ανανέωση των τιμών του πίνακα.

Ολοκληρώνοντας την περιγραφή του ΕΜ θα κάνουμε μια μικρή αναφορά στο θέμα της αρχικοποίησής του. Είναι προφανές ότι το τοπικό μέγιστο στο οποίο θα συγκλίνει ο αλγόριθμος εξαρτάται από την αρχικοποίησή του, για αυτό έχει αξία η αναζήτηση μιας "καλής" αρχικοποίησης. Η πιο απλή μέθοδος αναζήτησης είναι η επαναληπτική εκκίνηση του αλγορίθμου από τυχαίες θέσεις. Επίσης συχνά έχει χρησιμοποιηθεί και ο αλγόριθμος των k-μέσων για να προσδιορίσει αρχικές τιμές για μια μίξη κανονικών κατανομών, έτσι όμως το πρόβλημα μετατίθεται στην αρχικοποίηση των k-μέσων που είναι και πάλι τυχαία. Επιπλέον η τυχαιότητα που εισάγουν αυτές οι μέθοδοι δεν είναι επιθυμητή. Η ιδανική προσέγγιση στην αρχικοποίηση πρέπει να μας οδηγεί κοντά στο ολικό μέγιστο, και να το κάνει αυτό με συνέπεια. Με αυτό το στόχο, στην [82] έχει προταθεί ένας άπληστος ΕΜ αλγόριθμος για την εκπαίδευση μιας μίξης. Συγκεκριμένα ο αλγόριθμος ξεκινά με μια συνιστώσα και επαναληπτικά προσθέτει μια καινούρια, μέχρι έναν προκαθορισμένο μέγιστο αριθμό. Το κίνητρο για αυτή την προσέγγιση είναι το ότι αν έχουμε μια "καλή" λύση για την μίξη με j συνιστώσες, τότε με προσεκτική αναζήτηση μπορούμε να αρχικοποιήσουμε την $j+1$ συνιστώσα με "βέλτιστο" τρόπο και να οδηγηθούμε σε μια "καλή" λύση με $j+1$ συνιστώσες. Σε επόμενο κεφάλαιο θα εφαρμόσουμε αυτή την ιδέα για να καταλήξουμε σε έναν αλγόριθμο εκπαίδευσης του PRBF.

2.6 Η Variational Bayesian Μέθοδος

Στην παρούσα ενότητα θα μελετήσουμε την variational Bayesian μέθοδο έχοντας σαν αφετηρία τον αλγόριθμο ΕΜ. Η variational Bayesian μέθοδος έχει γνωρίσει τα τελευταία μεγάλη αποδοχή στην ερευνητική κοινότητα της μηχανικής μάθησης, και στην βιβλιογραφία υπάρχουν πολλές εισαγωγικές μελέτες, ενδεικτικά αναφέρουμε τις [48, 46]. Στην προσπάθεια να αναδειχθούν διαφορετικές πλευρές του ΕΜ έχουν προταθεί πολλές ερμηνείες που τον συγκρίνουν με άλλες μεθόδους βελτιστοποίησης. Με το ίδιο στόχο έχουν παρουσιαστεί ερμηνείες του που μελετάνε αντικειμενικές συναρτήσεις που διαφέρουν από την πιθανοφάνεια, όπως στις [39, 61] που στην πλήρη πιθανοφάνεια έχει προστεθεί ένας όρος εντροπίας. Στην [61] έχει παρουσιαστεί μια διαφορετική εξήγηση του ΕΜ, σύμφωνα με την οποία και στα δυο

του βήματα γίνεται μεγιστοποίηση μιας κατάλληλα ορισμένης αντικειμενικής συνάρτησης F. Χρησιμοποιώντας μια κατανομή q επί των κρυμμένων μεταβλητών Z ορίζουν την συνάρτηση

$$F(q,\theta) = \langle \log p(X,Z|\theta) \rangle_q - \langle \log q(Z) \rangle_q \tag{2.68}$$

όπου το $\langle \cdot \rangle_q$ συμβολή την μέση τιμή ως προς την κατανομή $q(Z)$. Οπότε η F προκύπτει αν στην αναμενόμενη τιμή της πλήρους πιθανοφάνειας $\langle \log p(X,Z|\theta) \rangle_q$ προστεθεί η εντροπία $-\langle \log q(Z) \rangle_q$ της κατανομής q. Ο αλγόριθμος που έχει προταθεί (Variational EM) μεγιστοποιεί επαναληπτικά την F ως προς q και ως προς θ, και αποτελεί μια ισοδύναμη διατύπωση του EM:

- *Expectation-βήμα*: Θέτουμε την νέα κατανομή $q_{t+1} = \arg\max_q F(q,\theta_t)$.
- *Maximization-βήμα*: Θέτουμε το νέο διάνυσμα παραμέτρων $\theta_{t+1} = \arg\max_\theta F(q_{t+1},\theta)$.

Όπου στο Ε-βήμα η μεγιστοποίηση γίνεται στο χώρο των συναρτήσεων q. Όπως αναφέρεται στην [61], εκτός από μια αλλαγή στο πρόσημο, η συνάρτηση F είναι ανάλογη με την "free energy" της στατιστικής φυσικής. Βασισμένοι σε ανάλογα συμπεράσματα της στατιστικής φυσικής, έχουν αποδείξει ότι για δεδομένη τιμή του θ, υπάρχει μια μοναδική κατανομή q_θ η οποία μεγιστοποιεί την $F(q,\theta)$ και δίνεται από την $q_\theta(Z) = p(Z|X,\theta)$. Οπότε η ισοδυναμία τού Ε-βήματος με το αντίστοιχο βήμα του EM είναι προφανής. Όσο αφορά την ισοδυναμία των δυο Μ-βημάτων, αρκεί να παρατηρήσουμε ότι ο όρος της εντροπίας στην (2.68) δεν εξαρτάται από τις παραμέτρους θ.

Η επαναληπτική εφαρμογή των βημάτων συγκλίνει σε τιμές q^* και θ^* που μεγιστοποιούν τοπικά την $F(q,\theta)$. Αυτό όμως που έχει μεγάλο ενδιαφέρον είναι ότι στην [61] έχει αποδειχθεί ότι αν η $F(q,\theta)$ έχει ένα τοπικό μέγιστο (q^*,θ^*), τότε και η λογαριθμική πιθανοφάνεια $\log p(X,\theta)$ έχει ένα τοπικό μέγιστο στο θ^*. Παρομοίως, αν η $F(q,\theta)$ έχει ολικό μέγιστο (q^*,θ^*), τότε και η λογαριθμική πιθανοφάνεια $\log p(X|\theta)$ έχει ολικό μέγιστο στο θ^*. Μια συνέπεια αυτού είναι ότι αν η F είναι πιο ομαλή από την λογαριθμική πιθανοφάνεια, και παρουσιάζει λιγότερα τοπικά μέγιστα, τότε μεγιστοποιώντας την F είναι πιθανόν να πάρουμε καλύτερους εκτιμητές για τις παραμέτρους από ότι αν μεγιστοποιούσαμε την πιθανοφάνεια.

Είναι εύκολο να δείξουμε ότι η αντικειμενική συνάρτηση F που έχει οριστεί στην [61] αποτελεί ένα κάτω φράγμα της λογαριθμικής πιθανοφάνειας. Από την (2.68) έχουμε

$$\begin{aligned} F(q,\theta) &= \langle \log \frac{p(X,Z|\theta)}{q(Z)} \rangle_q \\ &= \langle \log \frac{p(Z|X,\theta)p(X|\theta)}{q(Z)} \rangle_q \\ &= \langle \log \frac{p(Z|X,\theta)}{q(Z)} \rangle_q + \log p(X|\theta) \end{aligned} \tag{2.69}$$

Αναδιατάσσοντας τους όρους στα δυο μέλη της ισότητας παίρνουμε

$$\log p(X|\theta) = F(q,\theta) - \langle \log \frac{p(Z|X,\theta)}{q(Z)} \rangle_q \tag{2.70}$$

και αφού ο όρος $-\langle \log \frac{p(Z|X,\theta)}{q(Z)} \rangle_q$ είναι η απόσταση Kullback-Liebler μεταξύ των κατανομών $q(Z)$ και $p(Z,X|\theta)$, που είναι μεγαλύτερη ή ίση του μηδέν, προκύπτει ότι $\log p(X|\theta) \geq F(q,\theta)$.

Επεκτείνοντας την έρευνα για την χρήση διαφορετικών αντικειμενικών συναρτήσεων στον EM, στις [5, 6] έχει προταθεί η μεγιστοποίηση ενός κάτω φράγματος της λογαριθμικής πιθανοφάνειας περιθωρίου $\log p(X)$.

$$\begin{aligned} \log p(X) &= \log \int p(X,Z,\theta)\, \mathrm{d}\theta\, \mathrm{d}Z \\ &= \log \int q(\theta,Z) \frac{p(X,Z,\theta)}{q(\theta,Z)} \mathrm{d}\theta\, \mathrm{d}Z \\ &\geq \int q(\theta,Z) \log \frac{p(X,Z,\theta)}{q(\theta,Z)} \mathrm{d}\theta\, \mathrm{d}Z \\ &= \mathcal{F}(q) \end{aligned} \tag{2.71}$$

όπου η ανισότητα στο προτελευταίο βήμα ισχύει λόγω της ανισότητας του Jensen. Για την μεγιστοποίηση της F ως προς την συναρτησιακή μορφή της κατανομής q έχει προταθεί μια προσεγγιστική μέθοδος, η οποία αποτελεί την ουσία της variational Bayesian μεθόδου. Ορίζουν την variational εκ των υστέρων κατανομή $q(Z,\theta)$ η οποία είναι περιορισμένη να αποτελείται από ένα γινόμενο παραγόντων

$$q(Z,\theta) = q_Z(Z) q_\theta(\theta) \tag{2.72}$$

έτσι ώστε δοθέντων των παρατηρήσεων X, οι κρυμμένες μεταβλητές Z να είναι ανεξάρτητες από τις παραμέτρους θ του μοντέλου. Ο περιορισμός αυτός επιτρέπει την εφαρμογή ενός

επαναληπτικού αλγορίθμου μεγιστοποίησης που είναι ανάλογος με τον EM. Στο αντίστοιχο E-βήμα μεγιστοποιούν την $\mathcal{F}$ ως προς q_Z για να πάρουν την βέλτιστη εκ των υστέρων κατανομή των χρυμμένων μεταβλητών

$$q_Z(Z) \propto \exp(\log p(X, Z|\theta))_{q_\theta} \tag{2.73}$$

Στο αντίστοιχο M-βήμα μεγιστοποιούν την $\mathcal{F}$ ως προς q_θ για να πάρουν την βέλτιστη εκ των υστέρων κατανομή επί των παραμέτρων

$$q_\theta(\theta) \propto \exp(\log p(X, Z|\theta))_{q_Z} p(\theta) \tag{2.74}$$

αντί για τις βέλτιστες παραμέτρους που εκτιμούνται στο τυπικό αλγόριθμο EM.

Μια σημαντική συνεισφορά αυτής της μεθόδου είναι η χρήση της πιθανοφάνειας περιθωρίου, η οποία είναι ένας τρόπος για τον έλεγχο της πολυπλοκότητας του χρησιμοποιούμενου μοντέλου. Για τον υπολογισμό της απαιτείται η περιθωριοποίηση των παραμέτρων του μοντέλου, και έτσι εξάγεται ένας μέσος όρος επί όλων των πιθανών παραμετροποιήσεων. Στην περίπτωση πολλών σημαντικών μοντέλων η πολυπλοκότητά τους ελέγχεται μηδενίζοντας κάποιες παραμέτρους, όπως σε μια μίξη κατανομών που μικραίνει ο αριθμός των συνιστωσών της μηδενίζοντας κάποιους από τους συντελεστές της. Έτσι η περιθωριοποίηση των παραμέτρων έμμεσα περιθωριοποιεί και την πολυπλοκότητα του μοντέλου, και η αντικειμενική συνάρτηση που μεγιστοποιείται περιέχει πληροφορία για την πολυπλοκότητα του μοντέλου. Επιπλέον εκτός από αυτή την διαισθητική θεώρηση, στην [6] αναφέρεται ότι στο όριο μεγάλου δείγματος $N \to \infty$, η $\mathcal{F}$ αντιστοιχεί στο κριτήριο Bayesian Information Criterion καθώς και στο ισοδύναμο Mnimum Description Length Criterion.

Επεκτείνοντας αυτή την μέθοδο στις [34, 24] έχουν προταθεί ελαφρώς διαφορετικές υποθέσεις για τις κατανομές που υπεισέρχονται στο μοντέλο. Η πρώτη τροποποίηση είναι ο ορισμός μιας εκ των προτέρων κατανομής $p(Z)$ επί των χρυμμένων μεταβλητών. Αυτό έχει σαν αποτέλεσμα η λύση του E-βήματος να διαφοροποιείται από τις [5, 6], όπου ορίζονται εκ των προτέρων κατανομές μόνο επί των θ. Η δεύτερη τροποποίηση είναι η ρητή απαίτηση η variational εκ των υστέρων κατανομή να παραγοντοποιείται και ως προς όλα τα στοιχεία του θ. Θεωρώντας λοιπόν το σύνολο $\Theta = \{Z, \theta\}$ των χρυμμένων μεταβλητών και όλων των παραμέτρων απαιτείται κατά την μεγιστοποίηση της $\mathcal{F}$ να ισχύει

$$q(\Theta) = \prod_i q_i(\Theta_i) \tag{2.75}$$

αυτό ωστόσο δεν επιφέρει πρακτικές αλλαγές στο M-βήμα που έχει προταθεί στις [5, 6]. Εξαιτίας του ότι στις [5, 6] ορίζουν ανεξάρτητες και συζυγείς εκ των προτέρων κατανομές επί των θ, έτσι οι εκ των υστέρων κατανομές είναι και αυτές ανεξάρτητες. Γενικά η εκ των προτέρων κατανομή μιας παραμέτρου είναι συζυγής της κατανομής των παρατηρήσεων αν η εκ των υστέρων κατανομή της παραμέτρου, όπως υπολογίζεται από το θεώρημα του Bayes, έχει την ίδια συναρτησιακή μορφή με την εκ των προτέρων κατανομή. Αυτό μας βοηθάει κατά την μεγιστοποίηση της $\mathcal{F}$, γιατί μπορούμε έτσι να έχουμε λύσεις για τις q_i σε κλειστή μορφή.

Είναι εύκολο να δείξουμε ότι η αντικειμενική συνάρτηση που μεγιστοποιούμε διαφέρει από την λογαριθμική πιθανοφάνεια περιθωρίου κατά την απόσταση Kullback-Liebler μεταξύ της variational κατανομής και της πραγματικής κατανομής των Θ. Για την διαφορά τους ισχύει:

$$\begin{aligned}
\log p(X) - \mathcal{F}(q) &= \int q(\Theta) \log p(X) \mathrm{d}\Theta - \int q(\Theta) \log \frac{p(X,\Theta)}{q(\Theta)} \mathrm{d}\Theta \\
&= -\int q(\Theta) \log \frac{p(\Theta,X)}{q(\Theta)p(X)} \mathrm{d}\Theta \\
&= -\int q(\Theta) \log \frac{p(\Theta|X)}{q(\Theta)} \mathrm{d}\Theta
\end{aligned} \tag{2.76}$$

Έτσι σε αναλογία με την (2.70) ισχύει

$$\mathcal{F}(q) = \log p(X) + \langle \log \frac{p(\Theta|X)}{q(\Theta)} \rangle_q \tag{2.77}$$

Παρατηρούμε ότι η πιθανοφάνεια περιθωρίου δεν εξαρτάται από την q, οπότε η μεγιστοποίηση της F ισοδυναμεί με ελαχιστοποίηση της απόστασης Kullback-Liebler μεταξύ της variational εκ των υστέρων κατανομής $q(\Theta)$ και της πραγματικής εκ των υστέρων κατανομής $p(\Theta|X)$. Η απόσταση αυτή μηδενίζεται όταν οι δυο κατανομές ταυτιστούν, ωστόσο λόγω του περιορισμού (2.75) η απόσταση δεν μπορεί να μηδενιστεί, και αποφεύγεται αυτή η τετριμμένη λύση. Το κίνητρο για την επιβολή του περιορισμού είναι ότι έτσι η F θα αποτελεί μια καλή προσέγγιση της πιθανοφάνειας περιθωρίου, αλλά θα είναι ταυτόχρονα αρκετά απλή ώστε να υπολογίζεται εύκολα. Στην πράξη επαληθεύεται αρκετές φορές, και έχει οδηγήσει σε αλγορίθμους με πολύ καλή επίδοση.

Ολοκληρώνοντας την ενότητα θα υπολογίσουμε τις συναρτήσεις $q_i(\Theta_i)$ που μεγιστοποιούν την $F(q)$. Μπορεί να αναποδειχθεί ότι η $F(q)$ είναι μια κοίλη συνάρτηση (κυρτή

προς τα κάτω) της q, δες για παράδειγμα στην [46] και τις αναφορές που παρατίθενται εκεί, εδώ θα περιοριστούμε στην επίλυση της εξίσωσης των μερικών παραγώγων (εξίσωση Euler). Για την συναρτησιακή μορφή των q_i οι μόνες υποθέσεις που κάνουμε είναι ότι αποτελούν κατανομές και πρέπει να ικανοποιείται ο περιορισμός (2.75). Η μεγιστοποίηση γίνεται χρησιμοποιώντας τυπικές τεχνικές της συναρτησιακής ανάλυσης [4]. Θα μεγιστοποιήσουμε την F ξεχωριστά για κάθε κατανομή, οπότε για δοθείσα q_i γράφουμε

$$\begin{aligned}\mathcal{F}(q_i) &= \int \prod_i q_i(\Theta_i) \log \frac{p(X,\Theta)}{\prod_i q_i(\Theta_i)} \, \mathrm{d}\Theta \\ &= \int q_i(\Theta_i) \langle \log p(X,\Theta) \rangle_{j \neq i} \, \mathrm{d}\Theta_i - \int q_i(\Theta_i) \langle \log \prod_i q_i(\theta_i) \rangle_{j \neq i} \, \mathrm{d}\Theta_i \end{aligned} \quad (2.78)$$

όπου χρησιμοποιήσαμε τον συμβολισμό $\langle \cdot \rangle_{j \neq i}$ για να δηλώσουμε την μέση τιμή ως προς όλες της κατανομές εκτός της $q_i(\Theta_i)$, δηλαδή ως προς το $\prod_{j \neq i} q_j(\Theta_j)$. Επιπλέον ισχύει

$$\langle \log \prod_i q_i(\Theta_i) \rangle_{j \neq i} = \sum_i \langle \log q_i(\Theta_i) \rangle_{j \neq i} = \log q_i(\Theta_i) + \sum_{j \neq i} \langle \log q_j(\Theta_j) \rangle_{j \neq i} \quad (2.79)$$

οπότε γράφουμε την $\mathcal{F}(q_i)$ ισοδύναμα ως

$$\begin{aligned}\mathcal{F}(q_i) &= \int q_i(\Theta_i) \langle \log p(X,\Theta) \rangle_{j \neq i} \, \mathrm{d}\Theta_i - \int q_i(\Theta_i) \log q_i(\Theta_i) \, \mathrm{d}\Theta_i \\ &+ \sum_{j \neq i} \langle \log q_j(\Theta_j) \rangle_{j \neq i} \end{aligned} \quad (2.80)$$

Για να βρούμε το μέγιστο πρέπει να λύσουμε την εξίσωση του Euler, που στην περίπτωσή μας είναι

$$\frac{\partial}{\partial q_i} \left(q_i(\Theta_i) \langle \log p(X,\Theta) \rangle_{j \neq i} - q_i(\Theta_i) \log q_i(\Theta_i) \right) = 0 \quad (2.81)$$

Όμως προκειμένου η q_i να αποτελεί κατανομή πρέπει να ισχύει

$$\int q_i(\Theta_i) \, \mathrm{d}\Theta_i = 1 \quad (2.82)$$

για αυτό εισάγουμε έναν πολλαπλασιαστή Lagrange λ και λύνουμε την εξίσωση

$$\begin{aligned}&\frac{\partial}{\partial q_i} \left(q_i(\Theta_i) \langle \log p(X,\Theta) \rangle_{j \neq i} - q_i(\Theta_i) \log q_i(\Theta_i) + \lambda q_i(\Theta_i) \right) = 0 \\ &\langle \log p(X,\Theta) \rangle_{j \neq i} - \log q_i(\Theta_i) - 1 + \lambda = 0 \\ &\log q_i(\Theta_i) = \langle \log p(X,\Theta) \rangle_{j \neq i} - 1 + \lambda \\ &q_i(\Theta_i) = \exp \left\{ \langle \log p(X,\Theta) \rangle_{j \neq i} \right\} \exp \left\{ -1 + \lambda \right\} \end{aligned} \quad (2.83)$$

Από τον περιορισμό της κατανομής (2.82) που επιβάλαμε ισχύει

$$\exp\{-1+\lambda\}\int \exp\{\langle \log p(X,\Theta)\rangle_{j\neq i}\}\,\mathrm{d}\Theta_i = 1 \tag{2.84}$$

και ισοδύναμα η (2.83) γράφεται

$$q_i(\Theta_i) = \frac{\exp\{\langle \log p(X,\Theta)\rangle_{j\neq i}\}}{\int \exp\{\langle \log p(X,\Theta)\rangle_{j\neq i}\}\,\mathrm{d}\Theta_i} \tag{2.85}$$

Αυτή η λύση εξαρτάται από όλες τις q_j για $j \neq i$, οπότε απαιτείται η επαναληπτική της εφαρμογή για όλες τις variational εκ των υστέρων κατανομές q_i. Η διαδοχική εφαρμογή αυτής της εξίσωσης για όλες τις q_i αποτελεί μια επανάληψη της variational Bayesian μεθόδου. Η σύγκλιση της μεθόδου μπορεί να ελεγχθεί παρακολουθώντας την σύγκλιση των τιμών της $\mathcal{F}$.

Ολοκληρώνοντας την ενότητα, ένα ενδιαφέρον ερώτημα που παραμένει είναι το πόσο αυστηρό γίνεται το φράγμα της πιθανοφάνειας περιθωρίου, όταν υιοθετούμε τον περιορισμό (2.75), ο οποίος είναι γνωστός σαν mean field προσέγγιση στη φυσική. Σύμφωνα με την συζήτηση που γίνεται στην [46], αυτό εξαρτάται από το κατά πόσο οι Θ είναι ανεξάρτητες σύμφωνα με την πραγματική εκ των υστέρων κατανομή $p(\Theta|X)$. Αν είναι σχεδόν ανεξάρτητες, τότε η προσέγγιση θα είναι πολύ καλή. Σε διαφορετική περίπτωση μπορεί να είναι πολύ τραχιά, όπως για παράδειγμα όταν η εξάρτηση οφείλεται στην ύπαρξη περισσότερων από μιας κορυφών στην πραγματική εκ των υστέρων κατανομή, τότε η variational εκ των υστέρων κατανομή θα περιγράψει μόνο μια από αυτές.

Κεφαλαιο 3

Ομαδοποιηση με Μιξη Κανονικων Κατανομων

Ένα από τα ερωτήματα που πρέπει να απαντηθεί κατά την εκπαίδευση σε προβλήματα χάθε είδους μάθηση είναι ο καθορισμός της πολυπλοκότητας του μοντέλου που χρησιμοποιούμε, που αναφέρετε και σαν *επιλογή μοντέλου*. Στην περίπτωση μιας μίξης πρέπει να αποφασίσουμε τον αριθμό των συνιστωσών που θα έχει. Σε αυτό το κεφάλαιο προτείνουμε μια αυξητική μέθοδο για την επιλογή μοντέλου και την εκπαίδευση μιας μίξης, που βασίζεται στην variational Bayesian προσέγγιση. Η μέθοδος προσθέτει συνιστώσες στην μίξη χρησιμοποιώντας έναν έλεγχο Bayesian διάσπασης: μια συνιστώσα διασπάται σε δύο νέες συνιστώσες και στη συνέχεια εφαρμόζονται variational Bayesian εξισώσεις ενημέρωσης μόνο στις παραμέτρους των δυο νέων συνιστωσών. Ως αποτέλεσμα, είτε και οι δύο συνιστώσες διατηρούνται στη μίξη, είτε μια από τις δύο αποδεικνύεται περιττή και απαλείφεται από την μίξη. Στην προσέγγισή αυτή το πρόβλημα επιλογής μοντέλου αντιμετωπίζεται τοπικά σε μια περιοχή του χώρου δεδομένων, έτσι μπορούμε να θέσουμε πιο κατατοπιστικές εκ των προτέρων πιθανότητες βασιζόμενοι στην τοπική κατανομή των δεδομένων. Για την υλοποίηση αυτής της προσέγγισης παρουσιάζουμε μια τροποποιημένη Bayesian μίξη, καθώς και έναν αλγόριθμο εκπαίδευσης που εφαρμόζει επαναληπτικά έναν έλεγχο διάσπασης σε χάθε συνιστώσα της μίξης χρησιμοποιώντας variational μεθοδολογία. Τα πειραματικά αποτελέσματα και η πειραματική σύγκριση με δύο άλλες γνωστές τεχνικές επιβεβαιώνουν την αποδοτικότητα της

προτεινόμενης προσέγγισης.

3.1 Εισαγωγή

Η μίξη κανονικών κατανομών είναι ένα πολύτιμο στατιστικό εργαλείο για την μοντελοποίηση κατανομών. Είναι τόσο ευπροσάρμοστη ώστε να μπορεί να προσεγγίσει οποιαδήποτε κατανομή με μεγάλη ακρίβεια, και επιπλέον μπορεί να ερμηνευτεί σαν σταθμισμένη ομαδοποίηση. Για αυτό τον λόγο έχουν χρησιμοποιηθεί ευρέως σε προβλήματα μάθησης με επίβλεψη και μάθησης χωρίς επίβλεψη, και έχουν μελετηθεί εκτενώς π.χ. [58]. Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο, μια μίξη μπορεί να εκπαιδευτεί μέσω του αλγορίθμου EM [26, 57], που βρίσκει εκτιμητές μέγιστης πιθανοφάνειας για τις παραμέτρους της μίξης. Αυτή είναι και η τυπική μέθοδος εκπαίδευσης, αν και παρουσιάζει κάποια μειονεκτήματα. Πιο συγκεκριμένα, ο EM συγκλίνει σε ένα τοπικό ελάχιστο της πιθανοφάνειας, το οποίο εξαρτάται από τις αρχικές τιμές των παραμέτρων. Ένα άλλο πολύ σημαντικό πρόβλημα που εμφανίζεται κατά την εκπαίδευση τέτοιων μοντέλων είναι ο προσδιορισμός του αριθμού των συνιστωσών που χρειάζονται για ένα δοθέν σύνολο δεδομένων. Λόγω της σημασίας του προβλήματος επιλογής μοντέλου πολλές μεθόδοι έχουν προταθεί για την επίλυσή του.

Η πιο απλή προσέγγιση στην επιλογή μοντέλου είναι η εκπαίδευση ενός πλήθους από μίξεις με διαφορετικό αριθμό συνιστώσών μέσω του EM, και στη συνέχεια η αξιολόγηση των λύσεων χρησιμοποιώντας ένα κατάλληλο κριτήριο. Το κριτήριο αυτό δεν μπορεί να είναι η πιθανοφάνεια, γιατί τετριμμένα μπορούμε να την μεγιστοποιήσουμε χρησιμοποιώντας τόσες συνιστώσες όσα τα πρότυπα, και τοποθετώντας το κέντρο κάθε μιας πάνω σε ένα πρότυπο. Για αυτό συνήθως χρησιμοποιείται το άθροισμα ενός όρου πιθανοφάνειας και ενός όρου ποινής, ο οποίος τιμωρεί μίξεις με μεγάλο αριθμό συνιστωσών. Παραδείγματα τέτοιων κριτηρίων είναι τα Akaike's Information Criterion, Bayesian Inference Criterion, Laplace Empirical Criterion, Minimum Message Length, δες [58] για μια ανασκόπηση και σύγκριση των κριτηρίων. Στο ίδιο πνεύμα έχει προταθεί στην [72] η χρήση της πιθανοφάνειας διασταυρωμένης επικύρωσης (cross-validated likelihood), όπου μίξεις διαφορετικής πολυπλοκότητας εφαρμόζονται στα δεδομένα εκπαίδευσης, και αξιολογούνται ως προς την πι-

θανοφάνεια ενός ξεχωριστού συνόλου επικύρωσης (validation set). Έχουν επίσης προταθεί μέθοδοι που εκπαιδεύουν την μίξη και ταυτόχρονα προσαρμόζουν των αριθμό των συνιστωσών της. Στην [29] το Minimum Message Length (MML) κριτήριο έχει ενσωματωθεί στην αντικειμενική συνάρτηση, και βελτιστοποιείται χρησιμοποιώντας τον EM. Ξεκινώντας με ένα μεγάλο αριθμό συνιστωσών, η μέθοδος βελτιστοποίησης προοδευτικά αφαιρεί συνιστώσες, και το MML κριτήριο χρησιμοποιείται για να προτείνει το καλύτερο μοντέλο. Αυτή η προσέγγιση είναι επίσης λιγότερο ευαίσθητη στην αρχικοποίηση από ότι ο τυπικός EM. Μία Bayesian προσέγγιση έχει προταθεί στην [68], όπου ο αριθμός των συνιστωσών αντιμετωπίζεται σαν τυχαία μεταβλητή και χρησιμοποιείται η μέθοδος δειγματοληψίας Reversible Jump Markov Chain Monte Carlo [38], η οποία ωστόσο είναι πολύ απαιτητική υπολογιστικά. Για να αντιμετωπιστούν τα δύσκολα ολοκληρώματα που εμφανίζονται στην Bayesian μέθοδο έχει προταθεί η χρήση της variational προσέγγισης [61, 46, 35], που καταλήγουν στην variational Bayesian μέθοδο που είναι παρόμοια με τον EM. Αυτή η γενική μέθοδος βελτιστοποίησης έχει χρησιμοποιηθεί σε αρκετές πρόσφατες εργασίες, όπως στις [5, 6, 24] για την εκπαίδευση μιας μίξης. Επίσης στην [80] έχει χρησιμοποιηθεί σε συνδυασμό με το αλγόριθμο Split and Merge EM [81]. Σε αυτή την εργασία έχουν καταλήξει σε μια αντικειμενική συνάρτηση που επιτρέπει την ταυτόχρονη εκτίμηση των παραμέτρων και του αριθμού των συνιστωσών της μίξης, και την εφάρμοσαν για την εκπαίδευση μιας μίξης κανονικών κατανομών καθώς επίσης και σε μια μίξη εμπειρογνωμόνων (mixture of experts) για παλινδρόμηση. Η variational Bayesian μέθοδος έχει επίσης χρησιμοποιηθεί στην [34] για την εκτίμηση των παραμέτρων και του αριθμού των συνιστωσών μιας μίξης από factor analyzers. Σε αυτή εφαρμόζουν μια διαδικασία γεννήσεων/θανάτων στις συνιστώσες για να λύσουν το πρόβλημα τις επιλογής μοντέλου.

Μια από τις πιο ενδιαφέρουσες προσεγγίσεις στον καθορισμό του αριθμού των συνιστωσών έχει προταθεί στην [24]. Είναι μια variational Bayesian μέθοδος για την βελτιστοποίηση της πιθανοφάνειας περιθωρίου δοθέντων των συντελεστών της μίξης. Η μέθοδος ξεκινά με ένα μεγάλο αριθμό συνιστωσών, και προοδευτικά αφαιρεί αυτές που βρίσκονται στην ίδια περιοχή του χώρου δεδομένων. Έχουμε διαπιστώσει ότι η μέθοδος είναι αποδοτική, αλλά εξαρτάται από τις παραμέτρους των εκ των προτέρων πιθανοτήτων. Όπως θα δούμε στην συνέχεια, αν και η μέθοδος δεν επιτρέπει να καλύπτουν την ίδια ομάδα δεδομένων

πολλές συνιστώσες, αν η εκ των προτέρων πιθανότητα του πίνακα ακριβείας (αντίστροφος πίνακας συνδιακύμανσης) των συνιστωσών δεν επιλεγεί κατάλληλα, τότε φτάνει συχνά μια λύση όπου μια συνιστώσα καλύπτει περισσότερες από μια ομάδες. Επιπλέον αν και η μέθοδος είναι ντετερμινιστική, το σημείο σύγκλισής της εξαρτάται από την αρχική επιλογή των παραμέτρων της μίξης.

Εμείς προτείνουμε μια Bayesian μέθοδο για την εκπαίδευση μιας μίξης κανονικών κατανομών που είναι ντετερμινιστική, με καλώς ορισμένη αρχικοποίηση, και αντιμετωπίζει επαρκώς το πρόβλημα επιλογής μοντέλου. Πρόκειται για μια αυξητική μέθοδο: ξεκινά με μια μόνο συνιστώσα και προοδευτικά προσθέτει συνιστώσες στην μίξη. Η διαδικασία προσθήκης μιας συνιστώσας βασίζεται σε έναν έλεγχο διάσπασης, που εφαρμόζεται σε κάθε μια από τις υπάρχουσες συνιστώσες. Σύμφωνα με αυτόν τον έλεγχο μια συνιστώσα αντικαθίσταται από δυο υπο-συνιστώσες και στη συνέχεια εφαρμόζονται variational Bayesian εξισώσεις ενημέρωσης στο συγκεκριμένο ζεύγος, ενώ οι υπόλοιπες συνιστώσες διατηρούνται αμετάβλητες. Χάρη στην εισαγωγή των εκ των προτέρων πιθανοτήτων στις παραμέτρους των κανονικών συνιστωσών, οι συνιστώσες ανταγωνίζονται η μια την άλλη. Αν η κατανομή των δεδομένων στην περιοχή της συνιστώσας που ελέγχουμε υποδεικνύει την ύπαρξη περισσότερων από μια ομάδων, τότε και οι δύο υπο-συνιστώσες διατηρούνται και έτσι ο αριθμός των συνιστωσών αυξάνεται. Διαφορετικά ο ανταγωνισμός μεταξύ των δυο συνιστωσών θα προκαλέσει την απαλοιφή της μιας, και συνεπώς την ανάκτηση της αρχικής συνιστώσας. Αυτή η στρατηγική της αυξητικής προσθήκης συνιστωσών διευκολύνει επίσης τον καθορισμό των παραμέτρων των εκ των προτέρων πιθανοτήτων, αφού αυτός μπορεί να στηριχθεί στις παραμέτρους της συνιστώσας που ελέγχουμε. Προκειμένου να εφαρμόσουμε αυτή την ιδέα απαιτείται μια τροποποίηση της Bayesian μίξης την οποία θα περιγράψουμε στην συνέχεια.

3.2 Variational Bayesian Επιλογή Μοντέλου

Ένας κατάλληλος τρόπος για τον έλεγχο της πολυπλοκότητας μιας μίξης είναι η προσαρμογή των συντελεστών της μίξης. Μια συνιστώσα απαλείφεται από την μίξη αν ο αντίστοιχος συντελεστής μηδενιστεί. Συνεπώς μπορούμε να θεωρήσουμε μια μίξη με μεγάλο αριθμό συ-

νιστωσών, και να μεγιστοποιήσουμε μια κατάλληλη αντικειμενική συνάρτηση ως προς τους συντελεστές της μίξης. Με αυτό τον τρόπο οι περιττές συνιστώσες θα απαλειφθούν, καθώς οι αντίστοιχοι συντελεστές τους θα μηδενιστούν. Η τυπική εύρεση εκτιμητών μέγιστης πιθανοφάνειας μέσω του EM δεν αποτελεί μια βιώσιμη λύση για αυτού του είδους την επιλογή μοντέλου, καθώς ένας συντελεστής της μίξης που συνεχώς φθίνει έχει ως αποτέλεσμα μία συνιστώσα της οποίας ο πίνακας συνδιακύμανσης έχει φθίνουσες ιδιοτιμές, δηλαδή οδηγεί στον σχηματισμό μιας ιδιόμορφης συνιστώσας. Η Bayesian προσέγγιση παρέχει μια λύση σε αυτό το πρόβλημα καθώς περιορίζει τις παραμέτρους των συνιστωσών μέσω των εκ των προτέρων πιθανοτήτων, έτσι αποτρέπει τον σχηματισμό ιδιόμορφων συνιστωσών. Με αυτό τον τρόπο μόνο ασήμαντες συνιστώσες αφαιρούνται από την μίξη. Ωστόσο σε πολλές περιπτώσεις η variational Bayesian προσέγγιση έχει δώσει βιώσιμες λύσεις για την εκπαίδευση Bayesian μίξεων. Στο υπόλοιπο της ενότητας αυτής περιγράφουμε εν συντομία και σχολιάζουμε την variational Bayesian προσέγγιση για μίξεις κανονικών κατανομών.

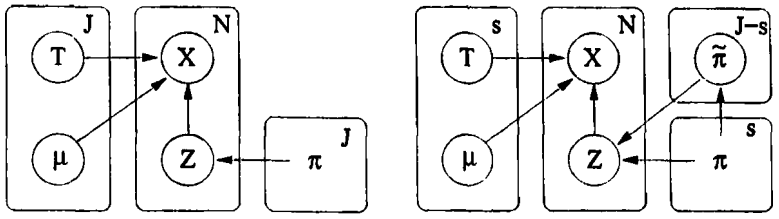
Έστω $X = \{x_n\}$ ένα σύνολο από N παρατηρήσεις, όπου κάθε $x_n \in \Re^d$ είναι ένα διάνυσμα χαρακτηριστικών. Επιπλέον θεωρούμε ότι η f είναι μια μίξη J κανονικών κατανομών

$$f(x) = \sum_{j=1}^{J} \pi_j \, \mathcal{N}(x|\mu_j, T_j) \tag{3.1}$$

όπου $\pi = \{\pi_j\}$ είναι οι συντελεστές της μίξης, $\mu = \{\mu_j\}$ τα κέντρα των συνιστωσών και $T = \{T_j\}$ οι πίνακες ακριβείας.

Η μοντελοποίηση των δεδομένων με την f υποδηλώνει την υπόθεση ότι για κάθε παρατήρηση x_n υπάρχει μια κρυμμένη μεταβλητή z_n που δείχνει την συνιστώσα που παρήγαγε την x_n. Έστω ότι $Z = \{z_n\}$ είναι το σύνολο αυτών των κρυμμένων μεταβλητών. Η z_n μπορεί να αναπαρασταθεί ως ένα J-διάστατο δυαδικό διάνυσμα, τέτοιο ώστε αν η j-στη συνιστώσα είναι υπεύθυνη για το x_n τότε $z_{jn} = 1$, αλλιώς $z_{jn} = 0$. Συνεπώς ισχύει ο περιορισμός $\sum_{j=1}^{J} z_{jn} = 1$. Επιπλέον η κατανομή του x_n δοθέντος της z_n είναι $\mathcal{N}(x_n|\mu_j, T_j)$, υποθέτοντας ότι $z_{jn} = 1$ για την συνιστώσα j.

Μια Bayesian μίξη λαμβάνεται θέτοντας εκ των προτέρων πιθανότητες στις παραμέτρους π, μ και T των συνιστωσών. Τυπικά χρησιμοποιούνται συζυγείς εκ των προτέρων



Σχήμα 3.1: (a) Το γραφικό μοντέλο που έχει προταθεί στην [24]. Τα πλαίσια υποδηλώνουν επανάληψη των τυχαίων μεταβλητών που περικλείουν, και ο ακριβής αριθμός των επαναλήψεων εμφανίζεται στην άνω δεξιά γωνία κάθε πλαισίου. Δεν κυκλώνουμε το π για να δηλώσουμε την ιδιαίτερη θεώρησή του ως παράμετρος χωρίς εκ των προτέρων κατανομή. (b) Το γραφικό μοντέλο που προτείνουμε, προσαρμοσμένο για τοπική επιλογή μοντέλου.

κατανομές, οι οποίες είναι για το μ_j κανονική $\mathcal{N}(\mu_j|0,\beta\,\mathcal{I})$, για το π Dirichlet

$$\mathcal{D}(\pi|\alpha_1,\ldots,\alpha_J) = \frac{\Gamma(\sum_{j=1}^J \alpha_j)}{\prod_{j=1}^J \Gamma(\alpha_j)} \prod_{j=1}^J \pi_j^{\alpha_j - 1}$$

με αναμενόμενη τιμή $\langle \pi_j \rangle = \frac{\alpha_j}{\sum_{i=1}^J \alpha_i}$, και για τον T_j Wishart

$$\mathcal{W}(T_j|\nu,V) = \frac{|T_j|^{(\nu-d-1)/2} \exp \operatorname{tr}\left\{-\frac{1}{2}VT_j\right\}}{2^{\nu d/2}\ \pi^{d(d-1)/4} |V|^{-n/2}\ \prod_{i=1}^d \Gamma((\nu+1-i)/2)}$$

με αναμενόμενη τιμή $\langle T_j \rangle = \nu V^{-1}$, όπου ν είναι οι βαθμοί ελευθερίας, και V είναι ο πίνακας διαβάθμισης. Αυτή η μίξη δεν είναι κατάλληλη για επιλογή μοντέλου με αλγόριθμους παρόμοιους με τον EM, καθώς η Dirichlet εκ των προτέρων κατανομή δεν επιτρέπει στους συντελεστές της μίξης να μηδενιστούν. Στην [24] έχει προταθεί ένα Bayesian μοντέλο που δεν υποθέτει εκ των προτέρων κατανομή για τους συντελεστές, οι οποίοι θεωρούνται απλοί παράμετροι και όχι τυχαίες μεταβλητές. Το γραφικό μοντέλο για αυτή την προσέγγιση απεικονίζεται στο Σχήμα 3.1(α).

Η Bayesian επιλογή μοντέλου επιτυγχάνεται μεγιστοποιώντας την πιθανοφάνεια περιθωρίου $p(X|\pi)$ που προκύπτει ολοκληρώνοντας τις μεταβλητές $\theta = \{Z, \mu, T\}$ από την από κοινού κατανομή $p(X,\theta|\pi)$

$$p(X|\pi) = \int p(X,\theta|\pi)\,\mathrm{d}\theta \tag{3.2}$$

και θεωρώντας τους συντελεστές της μίξης σαν παραμέτρους. Η variational προσέγγιση της Bayesian μεθόδου προτείνει την μεγιστοποίηση ενός κάτω φράγματος της λογαριθμικής

πιθανοφάνειας περιθωρίου

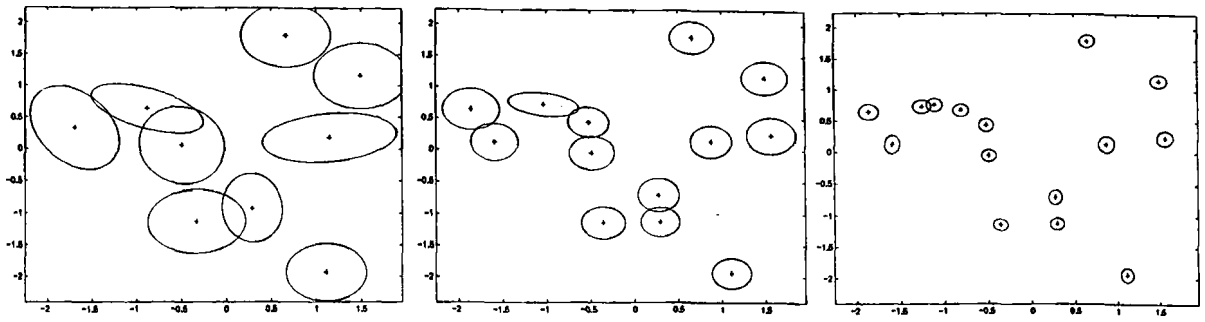
$$\mathcal{L}[q,\pi] = \int q(\theta) \log \frac{p(X,\theta|\pi)}{q(\theta)} \, \mathrm{d}\theta \leq \log p(X|\pi) \tag{3.3}$$

όπου $q(\theta)$ είναι μια αυθαίρετη κατανομή που προσεγγίζει την εκ των υστέρων κατανομή $p(\theta|X)$. Μια αξιοσημείωτη ιδιότητα είναι το ότι κατά την μεγιστοποίηση της $\mathcal{L}$, αν κάποιες από τις συνιστώσες βρίσκονται στην ίδια περιοχή του χώρου δεδομένων τότε υπάρχει μια ισχυρή τάση στην μίξη να απαλείψει τις περιττές συνιστώσες, αν τα δεδομένα σε αυτή την περισχή εξηγούνται ικανοποιητικά από λιγότερες συνιστώσες. Μια ερμηνεία αυτού του ανταγωνισμού λαμβάνεται από την ακόλουθη ανάλυση του variational φράγματος

$$\mathcal{L} = \int q(\theta) \log p(X|\theta) \, \mathrm{d}\theta - \int q(\theta) \log \frac{q(\theta)}{p(\theta|\pi)} \, \mathrm{d}\theta. \tag{3.4}$$

Ο πρώτος όρος αντιστοιχεί στην αναμενόμενη λογαριθμική πιθανοφάνεια, ως προς την $q(\theta)$. Ο δεύτερος όρος είναι η απόσταση Kullback-Leibler (KL) της $q(\theta)$ από την εκ των προτέρων $p(\theta|\pi)$. Όταν μια περιττή συνιστώσα αφαιρείται η απόσταση KL ελαττώνεται, και το κάτω φράγμα αυξάνεται ευνοώντας απλούστερα μοντέλα. Αν όμως υπάρχει ισχυρή ένδειξη από τα δεδομένα (δηλαδή η αναμενόμενη λογαριθμική πιθανοφάνεια έχει μεγάλη τιμή), τότε η συνιστώσα διατηρείται και το φράγμα αυξάνεται χάρη στην αύξηση του πρώτου όρου. Στην [6] εξετάζεται το πώς το Bayesian Information και το Minimum Message Length χριτήριο προκύπτουν σαν οριακές περιπτώσεις της variational μεγιστοποίησης της πιθανοφάνειας περιθωρίου.

Ο ανταγωνισμός μεταξύ των συνιστωσών οδήγησε σε μια νέα προσέγγιση για την αντιμετώπιση του προβλήματος της επιλογής μοντέλου: εκπαιδεύουμε μια μίξη με πολλές συνιστώσες και αφήνουμε τον ανταγωνισμό να απαλείψει τις περιττές. Αυτή είναι μια αποτελεσματική μέθοδος και γενικά δίνει την σωστή λύση, αν και παρουσιάζει κάποιες αδυναμίες. Η μέθοδος εξαρτάται από τις αρχικές τιμές των παραμέτρων της μίξης, και αυτό επηρεάζει την επιλογή μοντέλου ιδιαίτερα αν αρχικά δεν υπάρχουν αρκετές συνιστώσες. Επίσης αν αρχικοποιηθεί με πολλές συνιστώσες έχει μεγάλο υπολογιστικό κόστος για σύνολα με μεγάλο αριθμό δεδομένων και μεγάλο αριθμό χαρακτηριστικών. Εκτός αυτού η πιο σημαντική δυσκολία αφορά τον προσδιορισμό των παραμέτρων της εκ των προτέρων κατανομής του πίνακα ακριβείας T_j. Πιο συγκεκριμένα η εκ των προτέρων γνώση που ανατίθεται στον πίνακα διαβάθμισης V επηρεάζει το αποτέλεσμα της επιλογής μοντέλου, για παράδειγμα δες το Σχήμα 3.2. Η



Σχήμα 3.2: Χρήση του VBgmm για την μάθηση ενός τεχνητού συνόλου δεδομένων με τρεις διαφορετικούς πίνακες διαβάθμισης ($V = \beta\mathcal{I}$ όπου $\beta = 1, 0.25, 0.025$). Από αριστερά προς τα δεξιά, τα αποτελέσματα χρησιμοποιώντας πιο στενό πίνακα διαβάθμισης.

μέθοδος που περιγράφεται στην [24] (την ονομάζουμε VBgmm) έχει εφαρμοστεί σε ένα τεχνητό σύνολο δεδομένων με 208 διδιάστατα σημεία που σχηματίζουν 15 κανονικές ομάδες. Δοκιμάστηκαν τρεις διαφορετικοί πίνακες διαβάθμισης ($V = \beta\mathcal{I}$ όπου $\beta = 1, 0.25, 0.025$) για την εκπαίδευση μιας μίξης με 40 αρχικές συνιστώσες, που κατέληξαν σε λύσεις με 9, 13 και 15 συνιστώσες αντίστοιχα, δηλώνοντας την σημαντική εξάρτηση του αποτελέσματος από το V.

Έχουμε παρατηρήσει ότι αυτή είναι μια σταθερή τάση της μεθόδου: όσο πιο στενός είναι ο πίνακας διαβάθμισης που επιλέγεται τόσο περισσότερες συνιστώσες χρησιμοποιούνται στην τελική λύση. Εδώ χρησιμοποιούμε τον όρο στενός για να δηλώσουμε έναν πίνακα διαβάθμισης με συγκριτικά μικρότερές ιδιοτιμές. Εντούτοις δεν φαίνεται δυνατός ο εκ των προτέρων προσδιορισμός ενός βέλτιστου πίνακα. Αυτό το πρόβλημα γίνεται ακόμα πιο σημαντικό στην περίπτωση συνόλων δεδομένων που περιέχουν μαζί μικρές και μεγάλες ομάδες. Αν επιλεγεί ένας πλατύς πίνακας διαβάθμισης, τότε πολλές μικρές ομάδες θα καλυφθούν από μια συνιστώσα. Αν επιλεγεί ένας στενός πίνακας, τότε μια μεγάλη ομάδα μπορεί να καλυφθεί από πολλές συνιστώσες. Πρέπει να σημειωθεί ότι δεν έχουμε παρατηρήσει παρόμοια ευαισθησία της μεθόδου στην εκ των προτέρων κατανομή των κέντρων, η οποία επιλέγεται να είναι πλατιά και να μην περιορίζει τις πιθανές τιμές των κέντρων.

Συνοψίζοντας, το μειονέκτημα της μεθόδου είναι ότι χρησιμοποιώντας έναν αυθαίρετα πλατύ πίνακα διαβάθμισης δεν είναι δυνατόν να λάβουμε υπόψη μας τα χαρακτηριστικά των δεδομένων στην περιοχή όπου οι συνιστώσες ανταγωνίζονται μεταξύ τους. Με άλλα λόγια η μέθοδος λειτουργεί με μια καθολική εκ των προτέρων κατανομή για τον πίνακα ακριβείας,

ενώ μια τοπική εκ των προτέρων κατανομή φαίνεται πιο επιθυμητή. Αυτό μας οδήγησε στην ανάπτυξη μιας αυξητικής μεθόδου για την δημιουργία μιας μίξης. Σε κάθε βήμα της μεθόδου η μάθηση περιορίζεται στην περιοχή των δεδομένων που καταλαμβάνει μια συγκεκριμένη συνιστώσα j, έτσι μπορεί να προσδιοριστεί μια τοπική εκ των προτέρων κατανομή βασισμένη στον πίνακα ακριβείας T_j. Προκειμένου να πετύχουμε αυτή την συμπεριφορά χρειάζεται μια μετατροπή στο γραφικό μοντέλο που χρησιμοποιείται στον VBgmm για να περιοριστεί ο ανταγωνισμός σε ένα υποσύνολο τον συνιστωσών μόνο. Αυτή η ιδέα της τοπικής επιλογής μοντέλου παρουσιάζεται στην συνέχεια.

3.3 Ένα Bayesian Πλαίσιο για Τοπική Επιλογή Μοντέλου

Θεωρούμε ένα σύνολο παρατηρήσεων $X = \{x_n \in \Re^d | n = 1, \ldots, N\}$, και μια μίξη f με J κανονικές συνιστώσες

$$f_J(x) = \sum_{j=1}^{J} \pi_j \, \mathcal{N}(x|\mu_j, T_j). \tag{3.5}$$

Υποθέτουμε ότι ένα πλήθος από $J - s$ συνιστώσες περιγράφουν ικανοποιητικά τα δεδομένα στις αντίστοιχες περιοχές επιρροής· τότε τίθεται το ερώτημα: μπορούμε να βελτιστοποιήσουμε περισσότερο τις παραμέτρους των υπολοίπων συνιστωσών και παράλληλα να επιβάλουμε έναν μηχανισμό επιλογής μοντέλου; Με άλλα λόγια το πρόβλημα είναι το πώς να προσαρμόσουμε το μοντέλο της προηγούμενης ενότητας έτσι ώστε ο ανταγωνισμός μεταξύ των συνιστωσών να περιοριστεί σε ένα συγκεκριμένο υποσύνολό τους, ενώ οι υπόλοιπες να παραμείνουν ανεπηρέαστες.

Αυτό σημαίνει ότι χωρίζουμε τις συνιστώσες σε δύο ομάδες, στις "σταθερές" και τις "ελεύθερες", και εκτιμούμε μόνο τις παραμέτρους των τελευταίων. Εντούτοις πριν προχωρήσουμε σε αυτή την εκτίμηση είναι αναγκαίο να θέσουμε μια κατάλληλη εκ των προτέρων κατανομή στους συντελεστές μίξης των "σταθερών" συνιστωσών (τους ονομάζουμε "σταθερούς" συντελεστές), εμποδίζοντας έτσι την απαλοιφή τους από την μίξη. Στην συνέχεια οι "σταθεροί" συντελεστές αντιμετωπίζονται σαν τυχαίες μεταβλητές και περιθωριοποιούνται, οδηγώντας σε μια πιθανοφάνεια περιθωρίου που εξαρτάται μόνο από τους "ελεύθερους" συντελεστές. Μεγιστοποιώντας την πιθανοφάνεια περιθωρίου ως προς τους "ελεύθερους"

συντελεστές περιορίζουμε την αναζήτηση περιττών συνιστωσών μεταξύ των αντιστοίχων "ελεύθερων" συνιστωσών.

Το προτεινόμενο γραφικό μοντέλο απεικονίζεται στο Σχήμα 3.1(b). Είναι παρόμοια με το μοντέλο στο Σχήμα 3.1(a), με την διαφορά ότι έχει επιβληθεί μια εκ των προτέρων κατανομή επί των $J-s$ "ελεύθερων" συντελεστών $\tilde{\pi}$. Όπως και προηγουμένως, δοθέντος του συνόλου των κρυμμένων μεταβλητών $Z = \{z_{jn}\}$ ισχύει ότι

$$p(X|Z,\mu,T) = \prod_{n=1}^{N}\prod_{j=1}^{J}[\mathcal{N}(x_n|\mu_j,T_j)]^{z_{jn}} \tag{3.6}$$

υποθέτοντας ανεξάρτητες και ομοίως κατανεμημένες παρατηρήσεις. Η κατανομή του Z υποθέτοντας ανεξάρτητες και ομοίως κατανεμημένες κρυμμένες μεταβλητές είναι ένα γινόμενο πολυωνυμικών

$$p(Z|\pi,\tilde{\pi}) = \prod_{n=1}^{N}\prod_{j=1}^{s}\pi_j^{z_{jn}}\prod_{j=s+1}^{J}\tilde{\pi}_j^{z_{jn}} \tag{3.7}$$

δοθέντος του υποσυνόλου $\tilde{\pi} = \{\tilde{\pi}_j\}$ των "σταθερών" συντελεστών και του υποσυνόλου $\pi = \{\pi_j\}$ των "ελεύθερων" συντελεστών. Για ευκολία στον συμβολισμό και υποθέτοντας J συνιστώσες, μπορούμε πάντα να αναδιατάξουμε τους δείκτες των συνιστωσών έτσι ώστε οι πρώτες s να είναι οι "ελεύθερες". Τα υποσύνολα των συντελεστών είναι ξένα μεταξύ τους, και όλες οι τιμές τους είναι μη-αρνητικές και αθροίζουν στην μονάδα:

$$\sum_{j=1}^{s}\pi_j + \sum_{j=s+1}^{J}\tilde{\pi}_j = 1.$$

Το τυπικό Bayesian πλαίσιο υποθέτει συζυγή Dirichlet εκ των προτέρων κατανομή επί του συνόλου των συντελεστών μίξης. Όμως προκειμένου να εφαρμόσουμε την ιδέα μας, είναι αναγκαίο να ορίσουμε την υπό συνθήκη από κοινού κατανομή $p(\tilde{\pi}|\pi)$ των "σταθερών" συντελεστών δοθέντων των "ελεύθερων". Είναι γνωστό ότι αν η από κοινού κατανομή ενός συνόλου τυχαίων μεταβλητών είναι Dirichlet, τότε η από κοινού κατανομή περιθωρίου ενός υποσυνόλου τους είναι επίσης Dirichlet, δες [50]. Χρησιμοποιώντας το θεώρημα του Bayes μπορεί να προκύψει η υπό συνθήκη από κοινού κατανομή $p(\tilde{\pi}|\pi)$, η οποία είναι μια μη-τυπική Dirichlet με παραμέτρους α_j $(j = s+1,\ldots,J)$:

$$p(\tilde{\pi}|\pi) = \left(1-\sum_{j=1}^{s}\pi_j\right)^{-J+s}\frac{\Gamma(\sum_{j=s+1}^{J}\alpha_j)}{\prod_{j=s+1}^{J}\Gamma(\alpha_j)}\prod_{j=s+1}^{J}\left(\frac{\tilde{\pi}_j}{1-\sum_{k=1}^{s}\pi_k}\right)^{\alpha_j-1} \tag{3.8}$$

και αποτελεί την συζυγή εκ των προτέρων κατανομή των "σταθερών" συντελεστών μίξης $\tilde{\pi}$. Οι αναμενόμενες τιμές των $\tilde{\pi}_j$ και $\log \tilde{\pi}_j$ είναι

$$\langle \tilde{\pi}_j \rangle = \left(1 - \sum_{k=1}^{s} \pi_k\right) \frac{\sum_{n=1}^{N} \langle z_{jn} \rangle + \alpha_j}{\sum_{k=s+1}^{J} \left(\sum_{n=1}^{N} \langle z_{kn} \rangle + \alpha_k\right)} \tag{3.9}$$

$$\begin{aligned} \langle \log \tilde{\pi}_j \rangle &= \log\left(1 - \sum_{k=1}^{s} \pi_k\right) + \psi\left(\sum_{n=1}^{N} \langle z_{jn} \rangle + \alpha_j\right) \\ &\quad - \psi\left(\sum_{k=s+1}^{J} \sum_{n=1}^{N} \langle z_{kn} \rangle + \alpha_k\right) \end{aligned} \tag{3.10}$$

όπου $\psi(x)$ είναι η συνάρτηση δίγαμμα: $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x)$.

Ολοκληρώνοντας την περιγραφή του Bayesian μας μοντέλου υποθέτουμε μια κανονική και μια Wishart εκ των προτέρων κατανομή για τα μ και T αντίστοιχα

$$p(\mu) = \prod_{j=1}^{s} \mathcal{N}(\mu_j | 0, \beta\, \mathcal{I}) \tag{3.11}$$

$$p(T) = \prod_{j=1}^{s} \mathcal{W}(T_j | \nu, V). \tag{3.12}$$

Στην επόμενη ενότητα περιγράφουμε μια μέθοδο εκπαίδευσης για αυτό το μοντέλο, που βασίζεται στην μεγιστοποίηση της πιθανοφάνειας περιθωρίου.

3.4 Variational Bayesian Μάθηση με Τοπική Επιλογή Μοντέλου

Η μάθηση μέσα στο Bayesian πλαίσιο που περιγράψαμε μπορεί να επιτευχθεί μεγιστοποιώντας την πιθανοφάνεια περιθωρίου των δεδομένων, η οποία υπολογίζεται ολοκληρώνοντας ως προς τις χρυμμένες μεταβλητές την από κοινού κατανομή του γραφικού μοντέλου. Στην περίπτωσή μας, η πιθανοφάνεια περιθωρίου του X δοθέντος π υπολογίζεται ολοκληρώνοντας ως προς τα $\theta = \{Z, \mu, T, \tilde{\pi}\}$

$$p(X|\pi) = \sum_{Z} \int p(X, Z, \mu, T, \tilde{\pi}|\pi)\, \mathrm{d}\mu\, \mathrm{d}T\, \mathrm{d}\tilde{\pi}. \tag{3.13}$$

Ακολουθώντας την variational Bayesian μεθοδολογία [5, 6, 24, 34, 35] που στοχεύει στην μεγιστοποίηση ενός κάτω φράγματος $\mathcal{L}$ της λογαριθμικής πιθανοφάνειας περιθωρίου, μεγιστοποιούμε την

$$\begin{aligned} \mathcal{L}[q,\pi] &= \sum_{Z} \int q(Z,\mu,T,\tilde{\pi}) \log \frac{p(X,Z,\mu,T,\tilde{\pi}|\pi)}{q(Z,\mu,T,\tilde{\pi})} \,\mathrm{d}\mu \,\mathrm{d}T \,\mathrm{d}\tilde{\pi} && (3.14) \\ &\leq \log p(X|\pi) && (3.15) \end{aligned}$$

όπου q είναι μια αυθαίρετη κατανομή που προσεγγίζει την εκ των υστέρων κατανομή $p(Z,\mu,T,\tilde{\pi}|X,\pi)$. Η μεγιστοποίηση της $\mathcal{L}$ γίνεται επαναληπτικά, όπου σε κάθε επανάληψη εκτελούνται δύο βήματα (σε αναλογία με τον EM): πρώτα μεγιστοποιείται το φράγμα ως προς q, και στην συνέχεια μεγιστοποιείται ως προς π.

Για την μεγιστοποίηση ως προς q, επιλέγουμε την προσέγγιση mean-field [5, 6, 24, 34, 35], και θεωρούμε ότι το q είναι περιορισμένο να αποτελεί ένα γινόμενο της μορφής

$$q(\theta) = q_Z(Z)\, q_\mu(\mu)\, q_T(T)\, q_{\tilde{\pi}}(\tilde{\pi}).$$

Η μέθοδος δεν υποθέτει κάποια συγκεκριμένη μορφή για τους παράγοντες της q, αντιθέτως μεγιστοποιεί την $\mathcal{L}$ ως προς την συναρτησιακή μορφή των q_Z, q_μ, q_T και $q_{\tilde{\pi}}$. Η τυπική διαδικασία βελτιστοποίησης της συναρτησιακής ανάλυσης συνεπάγεται την χρήση της εξίσωσης Euler και περιορισμούς της μορφής πολλαπλασιαστών Lagrange για να εξασφαλίσουν ότι οι λύσεις είναι κατανομές, δες [4]. Η λύση για κάθε $\vartheta \in \theta$ είναι

$$q_\vartheta(\vartheta) = \frac{\exp\langle \log p(X,\theta|\pi)\rangle_{\theta-\vartheta}}{\int \exp\langle \log p(X,\theta|\pi)\rangle_{\theta-\vartheta} \,\mathrm{d}\vartheta} \qquad (3.16)$$

όπου οι αναμενόμενες τιμές $\langle\cdot\rangle_{\theta-\vartheta}$ υπολογίζονται ως προς όλες τις τυχαίες μεταβλητές εκτός της ϑ. Αναπτύσσοντας την προηγούμενη εξίσωση το αποτέλεσμα είναι το ακόλουθο σύνολο κατανομών:

$$\begin{aligned} q_Z(Z) &= \prod_{n=1}^{N} \prod_{j=1}^{s} r_{jn}^{z_{jn}} \prod_{j=s+1}^{J} \rho_{jn}^{z_{jn}} && (3.17) \\ q_\mu(\mu) &= \prod_{j=1}^{s} \mathcal{N}(\mu_j|m_j,S_j) && (3.18) \\ q_T(T) &= \prod_{j=1}^{s} \mathcal{W}(T_j|\eta_j,U_j) && (3.19) \\ q_{\tilde{\pi}}(\tilde{\pi}) &= \left(1-\sum_{k=1}^{s}\pi_k\right)^{-J+s} \frac{\Gamma(\sum_{j=s+1}^{J}\tilde{\alpha}_j)}{\prod_{j=s+1}^{J}\Gamma(\tilde{\alpha}_j)} \prod_{j=s+1}^{J} \left(\frac{\tilde{\pi}_j}{1-\sum_{k=1}^{s}\pi_k}\right)^{\tilde{\alpha}_j-1} && (3.20) \end{aligned}$$

Οι παράμετροι των κατανομών είναι οι ακόλουθες:

$$r_{jn} = \frac{\tilde{r}_{jn}}{\sum_{k=1}^{s}\tilde{r}_{kn}+\sum_{k=s+1}^{J}\tilde{\rho}_{kn}}, \text{ για } j=1,\ldots,s \tag{3.21}$$

$$\rho_{jn} = \frac{\tilde{\rho}_{jn}}{\sum_{k=1}^{s}\tilde{r}_{kn}+\sum_{k=s+1}^{J}\tilde{\rho}_{kn}}, \text{ για } j=s+1,\ldots,J \tag{3.22}$$

$$\begin{aligned}\tilde{r}_{jn} &= \pi_j \exp\left\{\frac{1}{2}\langle\log|T_j|\rangle\right. \\ &\quad \left. -\frac{1}{2}\mathrm{tr}\left\{\langle T_j\rangle\left(x_n x_n^T - x_n\langle\mu_j\rangle^T + \langle\mu_j\rangle x_n^T + \langle\mu_j\mu_j^T\rangle\right)\right\}\right\}\end{aligned} \tag{3.23}$$

$$\begin{aligned}\tilde{\rho}_{jn} &= \exp\left\{\frac{1}{2}\langle\log|T_j|\rangle + \langle\log\tilde{\pi}_j\rangle\right. \\ &\quad \left. -\frac{1}{2}\mathrm{tr}\left\{\langle T_j\rangle\left(x_n x_n^T - x_n\langle\mu_j\rangle^T + \langle\mu_j\rangle x_n^T + \langle\mu_j\mu_j^T\rangle\right)\right\}\right\}\end{aligned} \tag{3.24}$$

$$m_j = S_j^{-1}\langle T_j\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle x_n \tag{3.25}$$

$$S_j = \beta\,\mathcal{I} + \langle T_j\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle \tag{3.26}$$

$$\eta_j = \nu + \sum_{n=1}^{N}\langle z_{jn}\rangle \tag{3.27}$$

$$U_j = V + \sum_{n=1}^{N}\langle z_{jn}\rangle\left(x_n x_n^T - x_n\langle\mu_j\rangle^T - \langle\mu_j\rangle x_n^T + \langle\mu_j\mu_j^T\rangle\right) \tag{3.28}$$

$$\tilde{\alpha}_j = \alpha_j + \sum_{n=1}^{N}\langle z_{jn}\rangle. \tag{3.29}$$

Οι αναμενόμενες τιμές ως προς q που εμφανίζονται στις παραπάνω εξισώσεις ισούνται με: $\langle T_j\rangle = \eta_j U_j^{-1}$, $\langle\log|T_j|\rangle = \sum_{i=1}^{d}\psi(0.5(\eta_j+1-i)) + d\ln 2 - \ln|U_j|$, $\langle\mu_j\rangle = m_j$ και $\langle\mu_j\mu_j^T\rangle = S_j^{-1} + m_j m_j^T$. Όσο αφορά τις υπευθυνότητες (εκ των υστέρων πιθανότητες) μιας συνιστώσας j δοθέντος x_n, ισχύει για τις "ελεύθερες" συνιστώσες $\langle z_{jn}\rangle = r_{jn}$, $j=1,\ldots,s$, και για τις "σταθερές" συνιστώσες $\langle z_{jn}\rangle = \rho_{jn}$, $j=s+1,\ldots,J$. Οπότε χρησιμοποιώντας τις (3.9) και (3.10) έχουμε:

$$\langle\tilde{\pi}_j\rangle = \left(1-\sum_{k=1}^{s}\pi_k\right)\frac{\sum_{n=1}^{N}\rho_{jn}+\alpha_j}{\sum_{k=s+1}^{J}\left(\sum_{n=1}^{N}\rho_{kn}+\alpha_k\right)}$$

$$\langle\log\tilde{\pi}_j\rangle = \log\left(1-\sum_{k=1}^{s}\pi_k\right) + \psi\left(\sum_{n=1}^{N}\rho_{jn}+\alpha_j\right) - \psi\left(\sum_{k=s+1}^{J}\sum_{n=1}^{N}\rho_{kn}+\alpha_k\right).$$

Διαπιστώνεται ότι οι κατανομές εξαρτώνται η μια από την άλλη μέσω των αναμενόμενων τιμών, έτσι χρειάζεται επαναληπτική εκτίμηση των παραμέτρων τους. Ωστόσο στην πράξη ένα μόνο πέρασμα είναι αρκετό.

Μετά την μεγιστοποίηση της $\mathcal{L}$ ως προς q, το δεύτερο βήμα κάθε επανάληψης της μεθόδου εκπαίδευσης απαιτεί την μεγιστοποίηση της $\mathcal{L}$ ως προς π, σύμφωνα με τις ακόλουθες εξισώσεις ανανέωσης:

$$\pi_j = \left(1 - \sum_{k=s+1}^{J} \langle \tilde{\pi}_k \rangle\right) \frac{\sum_{n=1}^{N} \langle z_{jn} \rangle}{\sum_{k=1}^{s} \sum_{n=1}^{N} \langle z_{kn} \rangle}. \tag{3.30}$$

Η επαναληπτική μέθοδος που περιγράψαμε επαναλαμβάνεται μέχρι να συγκλίνει. Η σύγκλιση μπορεί να ελεγχθεί εξετάζοντας το variational φράγμα. Στην επόμενη ενότητα παρουσιάζουμε έναν αλγόριθμο που ενσωματώνει την τοπική επιλογή μοντέλου για να λύσει το πρόβλημα καθολικά.

3.5 Αυξητική Μάθηση Βασισμένη στην Διάσπαση Συνιστωσών

Εκμεταλλευτήκαμε την μέθοδο τοπικής επιλογής χαρακτηριστικών για να αναπτύξουμε μια αυξητική μέθοδο για την Bayesian εκπαίδευση μιας μίξης κανονικών κατανομών. Στην προσέγγισή μας οι συνιστώσες προσθέτονται σειριακά στην μίξη χρησιμοποιώντας την ακόλουθη διαδικασίας διάσπασης συνιστωσών: μία από τις συνιστώσες επιλέγεται και διασπάται κατάλληλα σε δυο νέες συνιστώσες. Αυτές τις συνιστώσες που προέκυψαν τις χειριζόμαστε σαν "ελεύθερες" και τις υπόλοιπες σαν "σταθερές", σύμφωνα με την ορολογία που εισαγάγαμε στην προηγούμενη ενότητα. Στην συνέχεια ορίζουμε την εκ των προτέρων κατανομή $p(T)$ βασισμένοι στα χαρακτηριστικά της συνιστώσας που διασπάσαμε, και εφαρμόζουμε την variational μάθηση με τοπική επιλογή μοντέλου που περιγράψαμε. Το αποτέλεσμα θα αντιστοιχεί σε μία από τρεις πιθανές περιπτώσεις. Στην πρώτη περίπτωση που οι δύο συνιστώσες περιγράφουν πολύ καλά τα δεδομένα της περιοχής τους, τότε και οι δύο διατηρούνται στην μίξη. Διαφορετικά η μέθοδος απαλείφει την μία, οπότε έχουμε την δεύτερη δυνατή περίπτωση. Υπάρχει επίσης μια σπάνια τρίτη περίπτωση, όπου και οι δύο νέες συνιστώσες

αφαιρούνται καθώς η συνιστώσα που διασπάσαμε είναι ασήμαντη (με πολύ μικρό συντελεστή μίξης). Μια τέτοια συνιστώσα τυγχάνει να περιγράφει λίγα απομονωμένα σημεία στην περιοχή μιας σημαντικής συνιστώσας. Μετά την διάσπαση η κυρίαρχη συνιστώσα περιγράφει και αυτά τα σημεία, και οι δύο νέες συνιστώσες αφαιρούνται από την μίξη. Στον αλγόριθμο που προτείνουμε δεν αποδεχόμαστε αυτή την σπάνια περίπτωση, γιατί μπορεί να οδηγήσει σε μια ατέρμονη επανάληψη. Ωστόσο μπορούμε να αφαιρέσουμε τέτοιου είδους συνιστώσες μετά τον τερματισμό του αλγορίθμου.

Ο έλεγχος διάσπασης εφαρμόζεται σειριακά σε όλες τις συνιστώσες, και η μέθοδος τερματίζει όταν όλες οι συνιστώσες έχουν ελεγχθεί και οι διασπάσεις απέτυχαν. Στην περίπτωση που μια διάσπαση πετύχει, τότε ο αριθμός των συνιστωσών της μίξης αυξάνεται και ένας νέος γύρος ελέγχων ακολουθεί. Ο προτεινόμενος αλγόριθμος συνοψίζεται στα ακόλουθα βήματα:

1. Θέτουμε $\beta := 1e-10$, και $\nu := d$.

2. Αρχικοποιούμε $J := 2$, $V := \mathrm{Cov}\{X\}$ και εκπαιδεύουμε την μίξη με τον αλγόριθμο VBgmm.

3. Αν μετά την σύγκλιση υπάρχει μια μόνο συνιστώσα, τότε σταματάμε.

4. Έστω C το σύνολο των J συνιστωσών που σχηματίζουν την μίξη f_J.

5. Ταξινομούμε τα στοιχεία του C σε φθίνουσα σειρά, σύμφωνα με την $|U_j|$.

6. Για κάθε συνιστώσα $c \in C$

   (α΄) Διασπάμε την c σε c_1 και c_2, σύμφωνα με τις (3.31)–(3.34), και σχηματίζουμε την f_{J+1}.

   (β΄) Έστω $F = \{c_1, c_2\}$ το σύνολο των ''ελεύθερων'' συνιστωσών, και $\bar{F}$ το σύνολο των ''σταθερών'' συνιστωσών με στοιχεία τις συνιστώσες της f_{J+1} εκτός από τις c_1 και c_2.

   (γ΄) Θέτουμε $\alpha_j := \sum_{n=1}^{N} \langle z_{jn} \rangle$ για $j \in \bar{F}$, και $V := \nu\lambda\mathcal{I}$ όπου λ είναι η μέγιστη ιδιοτιμή του U_c/η_c.

(δ΄) Εφαρμόζουμε επαναληπτικά τις (3.17)-(3.30) στις παραμέτρους της f_{J+1}, και μετά την σύγκλιση σχηματίζουμε την $f_{J'}$ με J' συνιστώσες.

(ε΄) Αν και οι δύο συνιστώσες στο F έχουν αφαιρεθεί, τότε

i. Σημειώνουμε την αποτυχία της διάσπασης.

ii. Συνεχίζουμε με την επόμενη συνιστώσα στο C (πηγαίνουμε στο βήμα 6α΄).

(ς΄) Αν μία από τις συνιστώσες στο F έχει αφαιρεθεί, τότε σημειώνουμε την αποτυχία της διάσπασης.

(ζ΄) Θέτουμε $J := J'$ και $f_J := f_{J'}$.

7. Αν όλες οι διασπάσεις έχουν σημειωθεί ως αποτυχημένες σταματάμε, αλλιώς πηγαίνουμε στο βήμα 4.

Για να περιγράψουμε τις λεπτομέρειες της διαδικασίας διάσπασης υποθέτουμε ότι κάποια συνιστώσα $\hat{\jmath}$ με κατανομή $\mathcal{N}(x|\mu_{\hat{\jmath}}, T_{\hat{\jmath}})$ πρέπει να διασπαστεί. Η ιδέα είναι ότι για να σχηματίσουμε την νέα μίξη απομακρύνουμε την συνιστώσα $\hat{\jmath}$, και εισάγουμε δυο νέες συνιστώσες με κατανομές $\mathcal{N}(x|\mu_{\hat{\jmath}1}, T_{\hat{\jmath}1})$ και $\mathcal{N}(x|\mu_{\hat{\jmath}2}, T_{\hat{\jmath}2})$ αντίστοιχα. Επιλέξαμε να τοποθετήσουμε τα δυο κέντρα τους κατά μήκος του πρωτεύοντος άξονα του πίνακα συνδιακύμανσης $T_{\hat{\jmath}}^{-1}$, και συμμετρικά σε σχέση με το κέντρο $\mu_{\hat{\jmath}}$. Τους συντελεστές μίξης των δύο συνιστωσών τους θέτουμε ίσους $\pi_{\hat{\jmath}1} = \pi_{\hat{\jmath}2} = \pi_{\hat{\jmath}}/2$, και τις παραμέτρους των συνιστωσών τις θέτουμε σύμφωνα με:

$$\mu_{\hat{\jmath}1} = \mu_{\hat{\jmath}} + \sqrt{\lambda}\, u \tag{3.31}$$

$$\mu_{\hat{\jmath}2} = \mu_{\hat{\jmath}} - \sqrt{\lambda}\, u \tag{3.32}$$

$$T_{\hat{\jmath}1} = T_{\hat{\jmath}} \tag{3.33}$$

$$T_{\hat{\jmath}2} = T_{\hat{\jmath}} \tag{3.34}$$

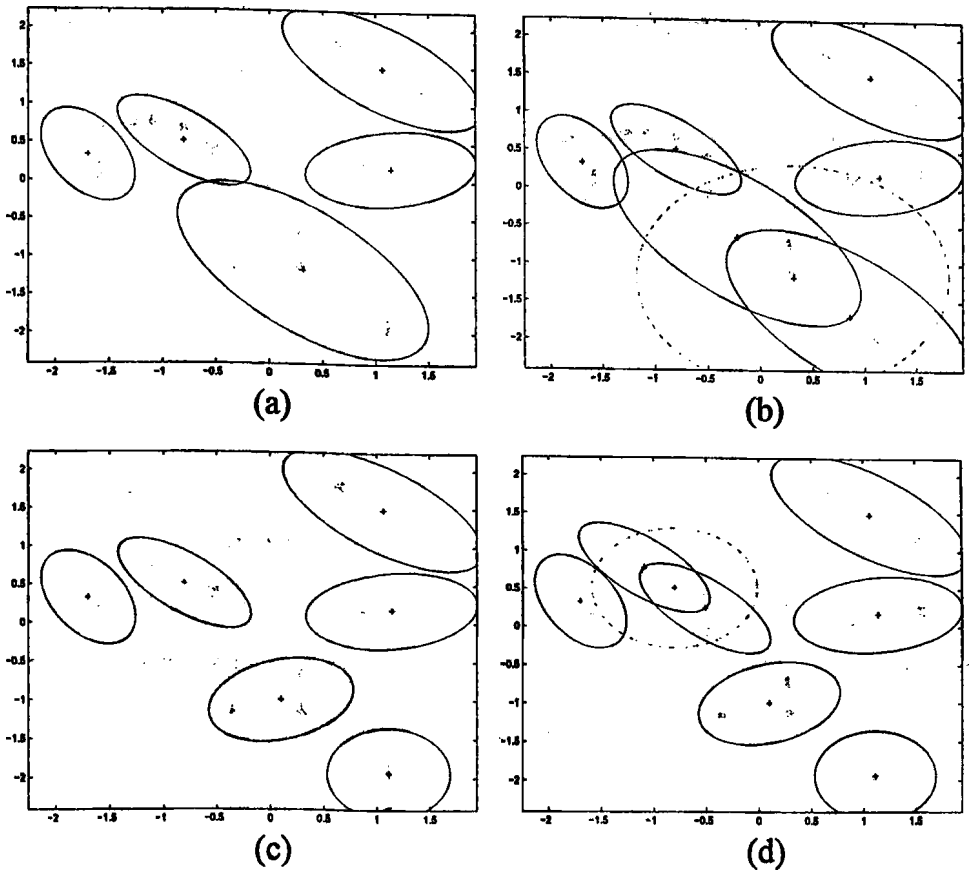
όπου λ είναι η μέγιστη ιδιοτιμή του $T_{\hat{\jmath}}^{-1}$ και u το αντίστοιχο ιδιοδιάνυσμα. Πρέπει να σημειώσουμε ότι κάναμε μια απλή και λογική επιλογή για την τοποθέτηση των κέντρων των νέων συνιστωσών, η οποία έχει επίσης χρησιμοποιηθεί και σε άλλες μεθόδους που υπεισέρχεται διάσπαση, π.χ. δες [32, 85] για πιο εξελιγμένες μεθόδους για τον προσδιορισμό του

άξονα διάσπασης. Άλλες επιλογές θα μπορούσαν επίσης να ελεγχθούν, π.χ. τυχαία επιλογή άξονα όπως στην [34], καθώς επίσης και πολλαπλοί έλεγχοι διάσπασης με διαφορετικές αρχικοποιήσεις των συνιστωσών.

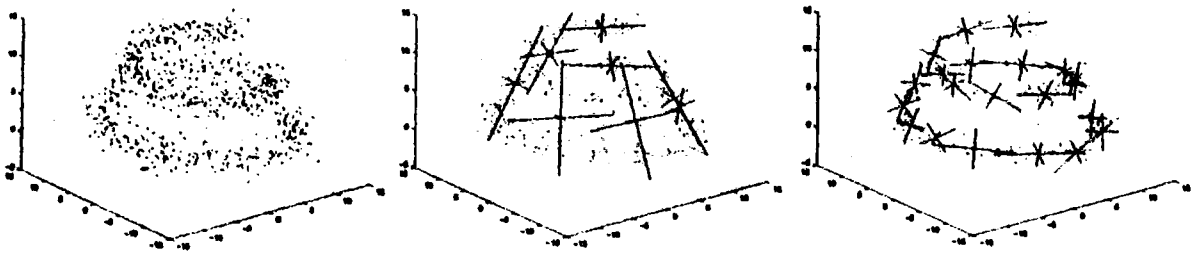
Ένα σημαντικό θέμα στην προτεινόμενη μέθοδο είναι ο προσδιορισμός του πίνακα διαβάθμισης V της εκ των προτέρων κατανομής $\mathcal{W}(\nu, V)$, με βάση την συνιστώσα που διασπάται. Θέτουμε $\nu = d$ (που είναι η μικρότερη επιτρεπόμενη τιμή), και θέλουμε η αναμενόμενη τιμή νV^{-1} της $\mathcal{W}(\nu, V)$ να είναι συγκρίσιμη με τον πίνακα ακριβείας T_j. Ωστόσο εμπειρικά παρατηρήσαμε ότι όταν θέτουμε $V = \nu T_j^{-1}$ η μέθοδος παρουσιάζει μια τάση να αποδέχεται περισσότερες διασπάσεις από ότι είναι απαραίτητο, και ότι πετυχαίνουμε καλύτερα αποτελέσματα αν ορίσουμε τον πίνακα διαβάθμισης να είναι λίγο πιο πλατύς. Σε αυτό το πνεύμα, επιλέξαμε να θέσουμε $V = \nu\lambda\mathcal{I}$, όπου λ είναι η μεγαλύτερη ιδιοτιμή του T_j^{-1}. Ένα παράδειγμα διάσπασης και προσδιορισμού της εκ των προτέρων κατανομής απεικονίζεται στο Σχήμα 3.3.

Μια άλλη πλευρά της μεθόδου αφορά την σειρά με την οποία οι συνιστώσες επιλέγονται για τον έλεγχο διάσπασης. Αν και η επίδοση δεν επηρεάζεται σημαντικά από την σειρά, παρατηρήσαμε ότι η μέθοδος επιταχύνεται (χάνει λιγότερους αποτυχημένους ελέγχους) αν δώσουμε προτεραιότητα στις πιο πλατιές συνιστώσες, όπου το πλάτος τους καθορίζεται από την ορίζουσα $|T_j^{-1}|$ του αντίστοιχου πίνακα συνδιακύμανσης.

Τέλος πρέπει να σημειώσουμε ότι προκειμένου να εφαρμόσουμε την μέθοδο τοπικής επιλογής μοντέλου, η μίξη πρέπει να αποτελείται από δύο τουλάχιστον συνιστώσες. Για να λάβουμε υπόψη μας το ενδεχόμενο τα δεδομένα μας να έχουν παραχθεί από μια μοναδική συνιστώσα, αρχικά εφαρμόζουμε τον αλγόριθμο VBgmm σε μια μίξη με δυο συνιστώσες, χρησιμοποιώντας ως εκ των προτέρων κατανομή του πίνακα διαβάθμισης τον αντίστροφο πίνακα συνδιακύμανσης των δεδομένων. Αν η εκπαίδευση καταλήξει σε μια μοναδική συνιστώσα σταματάμε, διαφορετικά εφαρμόζουμε τους έλεγχους διάσπασης στις δύο συνιστώσες που προέκυψαν.



Σχήμα 3.3: Τέσσερα στιγμιότυπα της διαδικασίας εκπαίδευσης. Ο αναμενόμενος πίνακας συνδιακύμανσης ως προς την εκ των προτέρων Wishart απεικονίζεται με διακεκομμένη γραμμή. (a) Μια ενδιάμεση λύση με 5 συνιστώσες. (b) Μια συνιστώσα διασπάται στα δύο. (c) Η μίξη μετά από variational Bayesian μάθηση. (d) Μια άλλη συνιστώσα επιλέγεται και διασπάται.



Σχήμα 3.4: Στα αριστερά, τα σπειροειδή δεδομένα. Στο μέσο, ένα ενδιάμεσο στάδιο του προτεινόμενου αλγορίθμου. Στα δεξιά, η τελική λύση. Για κάθε συνιστώσα της μίξης δείχνουμε τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης.

3.6 Πειράματα

Για την αξιολόγηση του προτεινόμενου αλγορίθμου (τον καλούμε VBgmmSplit) χρησιμοποιήσαμε τεχνητά και πραγματικά σύνολα δεδομένων, και συγκρίναμε την επίδοσή του με δυο ευρέως γνωστές μεθόδους: την μέθοδο που προτάθηκε στην [29] (την καλούμε MML-gmm[1]), και την μέθοδο που προτάθηκε στην [34] (την καλούμε VBmfa[2]).

Ο πρώτος έλεγχος του VBgmmSplit έγινε στα θορυβώδη συρρικνούμενα σπειροειδή δεδομένα [29, 34, 81] με σημεία που παράγονται σύμφωνα με την εξίσωση:

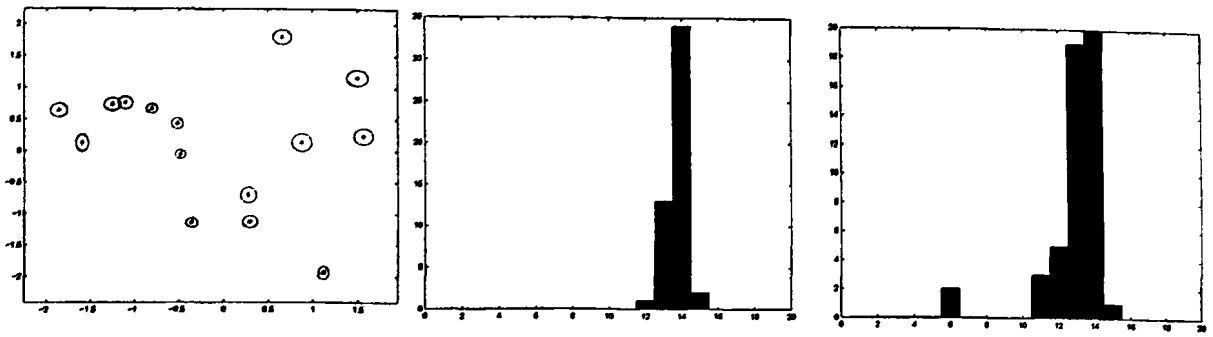
$$[x_1,\ x_2,\ x_3]^T = [(13 - 0.5t)\cos t,\ (0.5t - 13)\sin t,\ t]^T + [n_1,\ n_2,\ n_3]^T$$

όπου το t δειγματοληπτείται ομοιόμορφα στο διάστημα $[0, 4\pi]$, και τα n_1, n_2 και n_3 είναι ανεξάρτητα και όμοια κατανεμημένα σύμφωνα με την τυπική κανονική κατανομή $\mathcal{N}(0, 1)$. Εφαρμόσαμε τον αλγόριθμο σε ένα σύνολο 900 σημείων και το αποτέλεσμα ήταν μια μίξη με 21 συνιστώσες, όπως φαίνεται στο Σχήμα 3.4. Η λύση αυτή είναι παρόμοια με τα αποτελέσματα που αναφέρονται στις [29] και [34], όπου τα ίδια δεδομένα μαθεύτηκαν με μια μίξη από factor analyzers δίνοντας λύσεις με 13 συνιστώσες κατά μέσο όρο.

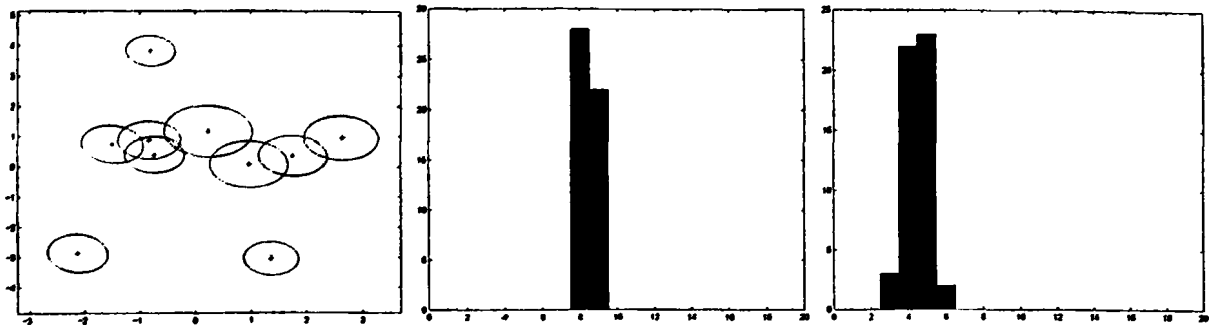
Συγκρίναμε επίσης τον αλγόριθμο με τεχνητά δεδομένα που σχηματίζουν κανονικές ομάδες, έτσι ώστε η μίξη που προκύπτει να ερμηνευθεί σαν ομαδοποίηση. Το πρώτο σύνολο (το ίδιο που χρησιμοποιήθηκε και στην Ενότητα 3.2) αποτελείται από 208 διδιάστατα σημεία

[1]Το λογισμικό είναι διαθέσιμο στο `http://www.lx.it.pt/~mtf/mixturecode.zip`.

[2]Το λογισμικό είναι διαθέσιμο στο `http://www.cse.buffalo.edu/faculty/mbeal/software/vbmfa/vbmfa.tar.gz`.

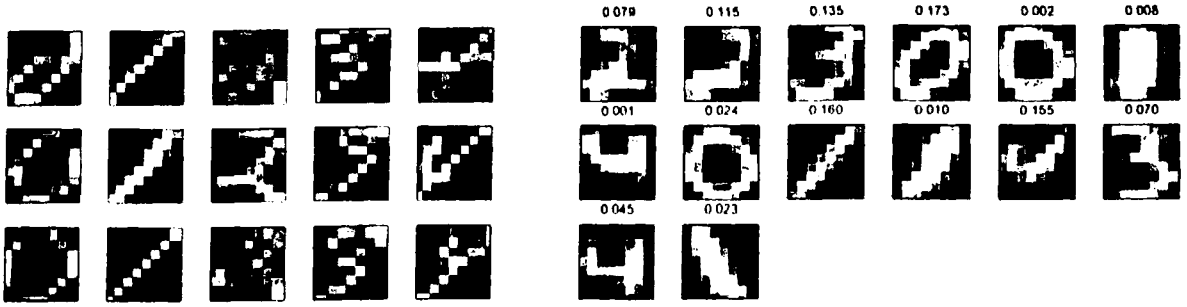


(a) Τεχνητά δεδομένα που σχηματίζουν 15 ομάδες.



(b) Τεχνητά δεδομένα που σχηματίζουν 10 ομάδες.

Σχήμα 3.5: Από αριστερά προς τα δεξιά: τα σημεία ομαδοποιημένα χρησιμοποιώντας τον VBgmmSplit, το ιστόγραμμα του αριθμού των συνιστωσών που βρήκε ο MMLgmm, και το ιστόγραμμα του αριθμού των συνιστωσών που βρήκε ο VBmfa.



Σχήμα 3.6: Στα αριστερά, κάποια παραδείγματα από το σύνολο ψηφίων. Στα δεξιά, το κέντρο κάθε συνιστώσας της μίξης που έμαθε ο VBgmmSplit. Πάνω από κάθε κέντρο εμφανίζεται ο αντίστοιχος συντελεστής μίξης.

που δημιουργήθηκαν από μια μίξη με 15 κανονικές συνιστώσες. Το Σχήμα 3.5(a) (αριστερά) απεικονίζει το αποτέλεσμα του VBgmmSplit. Εφαρμόσαμε επίσης τους MLgmm και VBmfa 50 φορές τον καθένα, και τις περισσότερες φορές έμαθαν τα δεδομένα χρησιμοποιώντας 14 συνιστώσες. Το Σχήμα 3.5(a) απεικονίζει τα ιστογράμματα του αριθμού των συνιστωσών που βρήκαν οι δύο αλγόριθμοι. Ο VBmfa εμφανίζει μεγαλύτερη διακύμανση στα αποτελέσματα από ότι ο MLgmm, και δυο φορές βρήκε λύσεις με μόνο 6 συνιστώσες. Στο επόμενο πείραμα χρησιμοποιήσαμε 505 10-διάστατα σημεία που προήλθαν από μια μίξη με 10 κανονικές συνιστώσες. Το Σχήμα 3.5(b) απεικονίζει το αποτέλεσμα του VBgmmSplit προβεβλημένο στα δύο ιδιοδιανύσματα με τις μεγαλύτερες ιδιοτιμές, και ιστογράμματα του αριθμού των συνιστωσών που βρήκαν ο MLgmm και ο VBmfa σε 50 εκτελέσεις. Ο MLgmm έδωσε καλύτερες λύσεις χρησιμοποιώντας 8 ή 9 συνιστώσες, ενώ ο VBmfa βρήκε 4 ή 5 συνιστώσες τις περισσότερες φορές.

Η επίδοση του αλγορίθμου ελέγχθηκε και σε πραγματικά δεδομένα. Πιο συγκεκριμένα, εφαρμόσαμε τον VBgmmSplit σε ένα σύνολο χειρόγραφων ψηφίων [45], το Σχήμα 3.6 (αριστερά) έχει κάποια παραδείγματα. Κάθε ψηφίο αποτελεί μια εικόνα από 8×8 εικονοστοιχεία, με 256 διαβαθμίσεις του γκρίζου. Θεωρήσαμε ένα σύνολο από 700 περιπτώσεις για κάθε ένα από τα ψηφία 0–4, και ομαδοποιήσαμε τα δεδομένα του με μια μίξη. Πριν την εκπαίδευση τυποποιήσαμε τα δεδομένα, ώστε κάθε χαρακτηριστικό τους να έχει μέση τιμή μηδέν και τυπική απόκλιση ένα. Τα κέντρα των συνιστωσών που βρήκε ο αλγόριθμος φαίνονται στο Σχήμα 3.6 (δεξιά). Αφού χρησιμοποιήσαμε κάθε αλγόριθμο για την εκπαίδευση μιας

Πίνακας 3.1: Ο αναμενόμενος αριθμός συνιστωσών και το σφάλμα ταξινόμησης (σε παρένθεση η τυπική απόκλιση) του αλγορίθμου VBgmm για το σύνολο χειρόγραφων ψηφίων. Τα αποτελέσματα αυτά πήραμε χρησιμοποιώντας μια Wishart εκ των προτέρων κατανομή με πίνακα διαβάθμισης $\beta\mathcal{I}$.

β	συνιστώσες	% σφάλμα	β	συνιστώσες	% σφάλμα
1	50.0 (0.0)	5.0 (1.0)	60	9.4 (1.5)	4.8 (3.9)
10	47.8 (1.6)	2.9 (1.3)	70	7.8 (0.8)	10.0 (4.3)
20	40.8 (1.6)	2.5 (1.0)	80	6.8 (0.8)	13.2 (8.4)
30	27.8 (2.7)	3.6 (3.6)	90	6.8 (0.4)	21.8 (10.5)
40	17.0 (1.7)	2.0 (0.6)	100	6.0 (0.0)	28.7 (4.1)
50	10.8 (0.8)	6.2 (4.2)	200	3.2 (0.4)	50.6 (6.1)

μίξης, συγκρίναμε τις λύσεις τους ως προς το σφάλμα "ταξινόμησης" για ένα ξεχωριστό σύνολο ελέγχου (200 περιπτώσεις για κάθε ψηφίο). Για να υπολογίσουμε αυτό το σφάλμα αναθέσαμε κάθε δεδομένο εκπαίδευσης στην συνιστώσα με την μεγαλύτερη υπευθυνότητα (εκ των υστέρων πιθανότητα). Μετά από αυτή την ομαδοποίηση αναθέσαμε σε κάθε συνιστώσα το ψηφίο που είχε η πλειονότητα των σημείων της. Προκειμένου να ταξινομήσουμε ένα άγνωστο δεδομένο ελέγχου υπολογίσαμε τις υπευθυνότητες κάθε συνιστώσας, και του αναθέσαμε το ψηφίο της συνιστώσας με την μεγαλύτερη υπευθυνότητα. Ο VBgmmSplit έδωσε μια μίξη με 14 συνιστώσες και σφάλμα ταξινόμησης 1.9%. Ο VBgmm δοκιμάστηκε επίσης με μια Wishart εκ των προτέρων κατανομή με διάφορους πίνακες διαβάθμισης, και τα αποτελέσματα συνοψίζονται στον πίνακα 3.1. Οι αναμενόμενες τιμές υπολογίστηκαν μετά από πέντε δοκιμές για κάθε πίνακα διαβάθμισης. Σε κάθε δοκιμή η μίξη είχε αρχικά 50 συνιστώσες, και η καλύτερη λύση είχε σφάλμα ταξινόμησης 2.0% με 17 συνιστώσες κατά μέσο όρο.

Εξαιτίας του ότι το σύνολο δεδομένων είχε υψηλή διάσταση και ήταν αραιό, ο αλγόριθμος MMLgmm έδωσε αποδεκτά αποτελέσματα κάτω υπό την υπόθεση ενός κοινού πίνακα συνδιακύμανσης για όλες τις συνιστώσες. Σε πέντε δοκιμές το μέσο σφάλμα ταξινόμησης ήταν 11.9% και ο μέσος αριθμός συνιστωσών ήταν 19.8. Υπό την υπόθεση ενός ξεχωριστού διαγώνιου πίνακα συνδιακύμανσης για κάθε συνιστώσα, το μέσο σφάλμα σε πέντε δοκιμές

ήταν 35.9%, και ο μέσος αριθμός συνιστωσών 7. Η μίξη που εκπαιδεύτηκε και στα δύο πειράματα είχε αρχικά 50 συνιστώσες.

Χρησιμοποιήθηκε επίσης ο αλγόριθμος VBmfa για την εκπαίδευση μιας μίξης από factor analyzers. Πρέπει να σημειώσουμε ότι για αυτό το σύνολο δεδομένων ο αλγόριθμος παρουσίασε ευαισθησία στην τιμή της μέγιστης επιτρεπόμενης διάστασης των factor analyzers, η οποία έπρεπε να καθοριστεί εξ αρχής. Για να πάρουμε αποτελέσματα συγκρίσιμα με τον VBgmmSplit σε χρόνο εκτέλεσης και αριθμό συνιστωσών, η μέγιστη διάσταση κάθε factor analyzer ορίστηκε να είναι 10 μετά από δοκιμές. Για πέντε εκτελέσεις το μέσο σφάλμα ήταν 11.4% και ο μέσος αριθμός συνιστωσών 14.

3.7 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάστηκε μια αυξητική μέθοδο για την εκπαίδευση και επιλογή μοντέλου σε μια μίξης κανονικών κατανομών. Η μέθοδος αυτή βελτιώνει την Bayesian προσέγγιση που προτάθηκε στην [24] και η οποία προσφέρει ένα μηχανισμό για τον ανταγωνισμό μεταξύ των συνιστωσών που βρίσκονται στην ίδια περιοχή του χώρου δεδομένων και απαλοιφή των περιττών. Όμως εκτός από το ζήτημα της αρχικοποίησης, αυτή η προσέγγιση παρουσιάζει ευαισθησία στις παραμέτρους της εκ των προτέρων κατανομής του πίνακα ακριβείας. Όπως δείξαμε, είναι δύσκολο να προσδιορίσουμε κατάλληλες τιμές, ειδικά στην περίπτωση δεδομένων που σχηματίζουν ομάδες διαφορετικού μεγέθους.

Η προτεινόμενη μέθοδος ξεπερνά αυτή την δυσκολία προσθέτοντας σειριακά συνιστώσες στη μίξη χρησιμοποιώντας ένα Bayesian έλεγχο διάσπασης, όπου μια συνιστώσα διασπάται σε δύο νέες και στη συνέχεια εφαρμόζονται variational εξισώσεις ενημέρωσης στις παραμέτρους αυτών των συνιστωσών. Ως αποτέλεσμα είτε και δύο συνιστώσες διατηρούνται στην μίξη ή μία αποδεικνύεται περιττή και αφαιρείται. Η προσέγγισή μας επιτρέπει τον προσδιορισμό μιας διαφορετικού τοπικής εκ των προτέρων κατανομής $p(T)$ για κάθε έλεγχο διάσπασης, και οι παράμετροί της προσδιορίζονται παίρνοντας υπόψη τα χαρακτηριστικά του πίνακα ακριβείας της συνιστώσας που ελέγχουμε. Επιπλέον η προτεινόμενη μέθοδος είναι ντετερμινιστική και δεν εξαρτάται από την αρχικοποίηση της μίξης, όπως συμβαίνει μα άλ-

λες μεθόδους. Όπως φαίνεται από τα πειραματικά αποτελέσματα και τις συγκρίσεις με δυο άλλες γνωστές μεθόδους η προτεινόμενη μέθοδος αντιμετωπίζει επαρκώς το πρόβλημα της επιλογής μοντέλου για μια μίξη κανονικών κατανομών.

Η μελλοντική μας έρευνα θα επικεντρωθεί στον έλεγχο και την τελειοποίηση δύο θεμάτων. Το πρώτο είναι η αναζήτηση εναλλακτικών τρόπων για τον καθορισμό της τοπικής εκ των προτέρων κατανομής του πίνακα ακριβείας. Εκτός αυτού, είναι δυνατό να εκτελέσουμε πολλαπλούς ελέγχους διάσπασης για την ίδια συνιστώσα με τον πίνακα διαβάθμισης να πλαταίνει σταδιακά προκειμένου να αποκτήσουμε ένα μέτρο για ευρωστία του ελέγχου διάσπασης. Το δεύτερο θέμα είναι να εξετάσουμε εναλλακτικούς τρόπους για την αρχικοποίηση των κέντρων των νέων συνιστωσών που προκύπτουν από την διάσπαση (π.χ. στην [34] επιλέγονται τυχαία). Επίσης είναι δυνατό να εκτελέσουμε πολλαπλές διασπάσεις για μια συνιστώσα, με διαφορετική αρχικοποίηση των συνιστωσών που προκύπτουν. Τέλος άλλα θέματα που πρέπει να εξεταστούν είναι το πώς κλιμακώνεται η μέθοδος ως προς την διάσταση και το πλήθος των δεδομένων, η δυνατότητα να εκτελούμε ταυτόχρονα ελέγχους σε πολλές συνιστώσες, και η χρήση της μεθόδου σε εφαρμογές διαφόρων ερευνητικών πεδίων.

Κεφαλαιο 4

Επιλογη Χαρακτηριστικων για Ομαδοποιηση

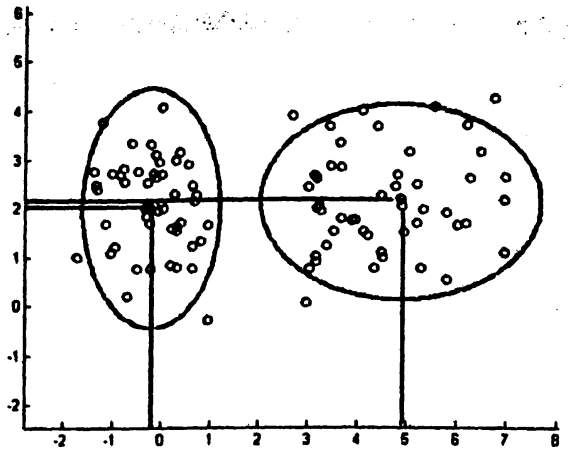
Σε αυτό το κεφάλαιο παρουσιάζουμε μια Bayesian μέθοδο για την εκπαίδευση μιας μίξης που αντιμετωπίζει ταυτόχρονα το πρόβλημα της επιλογής χαρακτηριστικών και της επιλογής μοντέλου [23]. Η μέθοδος στηρίζεται στον συνδυασμό μιας μίξης που λαμβάνει υπόψη της τα εξέχοντα χαρακτηριστικά των προτύπων, και μια Bayesian μέθοδο μάθησης που εκτιμά τον αριθμό των συνιστωσών της μίξης. Η προτεινόμενη μέθοδος ακολουθεί την variational Bayesian προσέγγιση και μπορεί να βελτιστοποιεί ταυτόχρονα τις παραμέτρους της μίξης, τον αριθμό των συνιστωσών της και την σημαντικότητα των χαρακτηριστικών των δεδομένων. Τα πειραματικά αποτελέσματα με τεχνητά και πραγματικά δεδομένα υψηλής διάστασης αποτελούν ενδείξεις για την αποτελεσματικότητα της μεθόδου.

4.1 Εισαγωγή

Η μίξη κανονικών κατανομών αποτελεί μια ευρέως χρησιμοποιούμενη μέθοδο σε προβλήματα μάθησης χωρίς επίβλεψη. Η προσέγγιση της κατανομής των δεδομένων με μια μίξη μπορεί να ερμηνευθεί σαν ταυτοποίηση των ομάδων που σχηματίζουν τα δεδομένα με τις συνιστώσες της μίξης. Η εκτίμηση των παραμέτρων μιας μίξης με καθορισμένο αριθμό συνιστωσών γίνε-

ται μεγιστοποιώντας την πιθανοφάνεια με τον αλγόριθμο ΕΜ ή κάποια παραλλαγή του [58]. Εκτός από τον καθορισμό του αριθμού των συνιστωσών, ένα ερώτημα που προκύπτει φυσιολογικά είναι η εύρεση των εξεχόντων χαρακτηριστικών, ειδικά σε προβλήματα με δεδομένα υψηλής διάστασης. Διαισθητικά, τα εξέχοντα χαρακτηριστικά είναι εκείνα που διευκολύνουν την εκπαίδευση και δίνουν ομάδες με τις επιθυμητές ιδιότητες. Όσο αφορά τις μίξεις κανονικών κατανομών, τα εξέχοντα χαρακτηριστικά των δεδομένων πρέπει να περιγράφουν μια κατανομή που εμφανίζει πολλές ελλειψοειδής ομάδες, και σε ικανοποιητική απόσταση μεταξύ τους ώστε να "διακρίνονται" εύκολα. Σε αντιπαράθεση με τα χαρακτηριστικά που περιγράφουν δεδομένα με ομοιόμορφη κατανομή ή κατανομή με μια μόνο κορυφή, και τα οποία δεν παρέχουν πληροφορία σχετικά με την ύπαρξη ομάδων στα δεδομένα. Τέτοια "ασήμαντα" χαρακτηριστικά μπορούν να προκαλέσουν σύγχυση κατά την εκπαίδευση αυξάνοντας την πολυπλοκότητα του μοντέλου, για παραδείγματα δες [13, 51]. Ένα παράδειγμα χαρακτηριστικών με διαφορετική σημαντικότητα παρουσιάζεται στο Σχήμα 4.1. Σημειώνουμε ότι η επιλογή των χαρακτηριστικών και η επιλογή του αριθμού των συνιστωσών είναι αλληλοεξαρτώμενα προβλήματα. Χρησιμοποιώντας διαφορετικά υποσύνολα χαρακτηριστικών είναι δυνατό να πάρουμε διαφορετικές εκτιμήσεις για τον αριθμό των ομάδων που σχηματίζονται, για επιπλέον σχόλια δες [27]. Είναι επίσης αναμενόμενο ότι χρησιμοποιώντας περισσότερα χαρακτηριστικά μπορούμε να οδηγηθούμε σε πιο πολύπλοκες κατανομές των δεδομένων, και συνεπώς περισσότερες ομάδες. Για το λόγο αυτό είναι σημαντικό να επιτευχθεί ταυτόχρονη επιλογή χαρακτηριστικών και επιλογή μοντέλου.

Για να αντιμετωπίσουμε την επιλογή χαρακτηριστικών και μοντέλου παρουσιάζουμε μια variational Bayesian μέθοδο για την εκπαίδευση μιας μίξης δύο επιπέδων, η εκπαίδευση της οποίας βασίζεται στην μεγιστοποίηση ενός κάτω φράγματος της πιθανοφάνειας περιθωρίου. Χρησιμοποιούμε το μοντέλο που προτάθηκε στην [51], δηλαδή μια μίξη κανονικών κατανομών που ενσωματώνει μια διαδικασία για τον προσδιορισμό των εξεχόντων χαρακτηριστικών, και η σημαντικότητα κάθε χαρακτηριστικού μετριέται σαν μια πιθανότητα. Έτσι όταν αυτή η πιθανότητα παίρνει τιμές κοντά στο μηδέν το χαρακτηριστικό πρακτικά δεν λαμβάνεται υπόψη. Αυτή η προσέγγιση είναι ελκυστική, καθώς δεν απαιτεί τον διεξοδικό έλεγχο όλων των πιθανών υποσυνόλων των χαρακτηριστικών, που γενικά είναι ανέφικτος. Σύμφωνα με το Bayesian πλαίσιο ορίζουμε εκ των προτέρων κατανομές για τις παραμέτρους της μίξης,



Σχήμα 4.1: Διδιάστατα δεδομένα που έχουν προέλθει από την μίξη δυο κανονικών συνιστωσών. Η προβολή των δεδομένων στον οριζόντιο άξονα θα παρουσιάσει δυο ομάδες, ενώ η προβολή τους στον κατακόρυφο άξονα μόνο μια.

και μεγιστοποιούμε την κατανομή περιθωρίου δοθέντων των συντελεστών μίξης και της σημαντικότητας των χαρακτηριστικών. Για την βελτιστοποίηση χρησιμοποιούμε variational μεθόδους για να εξάγουμε έναν αλγόριθμο παρόμοιο του EM [61], ακολουθώντας την προσέγγιση στις [6, 24].

4.2 Επιλογή Χαρακτηριστικών για Μάθηση Χωρίς Επίβλεψη

Το πρόβλημα της επιλογής χαρακτηριστικών, αν και έχει μελετηθεί εκτενώς σε σχέση με την ταξινόμηση, μόνο πρόσφατα άρχισε να εξετάζεται παράλληλα με την ομαδοποίηση. Δύο βασικές προσεγγίσεις έχουν προταθεί· στην προσέγγιση του "περιτυλίγματος" (wrapper) ο αλγόριθμος ομαδοποίησης περιβάλλεται από έναν αλγόριθμο επιλογής ενός υποσυνόλου των χαρακτηριστικών. Ο αλγόριθμος επιλογής εκτελεί μια αναζήτηση για την εύρεση ενός καλού υποσυνόλου, και χρησιμοποιεί τον αλγόριθμο ομαδοποίησης σαν τμήμα της συνάρτησης που αποτιμά το υποψήφιο υποσύνολο χαρακτηριστικών. Η δεύτερη προσέγγιση αντιμετωπίζει την ομαδοποίηση και την επιλογή χαρακτηριστικών ταυτόχρονα, ορίζοντας μια κατάλληλη αντικειμενική συνάρτηση. Η βελτιστοποίηση αυτής της συνάρτησης αποφέρει ένα υποσύνο-

λο χαρακτηριστικών και την ομαδοποίηση των δεδομένων στον χώρο που ορίζουν. Στην συνέχεια περιγράφουμε συνοπτικά κάποιες αντιπροσωπευτικές μεθόδους.

Στην [27] χρησιμοποιούν την προσέγγιση περιτυλίγματος. Ψάχνουν στον χώρο των υποσυνόλων των χαρακτηριστικών, και αποτιμούν κάθε υποψήφιο υποσύνολο ομαδοποιώντας τα δεδομένα και χρησιμοποιώντας εν συνεχεία κάποιο κατάλληλο μέτρο. Για να ψάξουν τον χώρο των υποσυνόλων χρησιμοποιούν σειριακή προς τα εμπρός (forward) αναζήτηση ξεκινώντας με το κενό σύνολο, και προσθέτονας ένα χαρακτηριστικό τη φορά. Για να εντοπίσουν το καλύτερο υποσύνολο χαρακτηριστικών χρησιμοποιούν το κριτήριο της μέγιστης πιθανοφάνειας. Για την ομαδοποίηση των δεδομένων χρησιμοποιούν μια μίξη κανονικών κατανομών που εκπαιδεύουν με τον EM. Για να εκτιμήσουν τον αριθμό των συνιστωσών της μίξης δοκιμάζουν με τη σειρά να συγχωνεύσουν δυο συνιστώσες, και χρησιμοποιούν το Bayesian Information Criterion για να επιλέξουν το καλύτερο από την αλληλουχία μοντέλων που δημιουργούν.

Στην [51] ακολουθούν την δεύτερη προσέγγιση και ορίζουν την σημαντικότητα των δεδομένων ως πιθανότητα. Χρησιμοποιούν μίξη κανονικών κατανομών για την ομαδοποίηση, και υποθέτουν ανεξάρτητα χαρακτηριστικά δοθείσης μιας συνιστώσας. Το μέτρο της σημαντικότητας του χαρακτηριστικού είναι η πιθανότητα με την οποία η κατανομή των παρατηρήσεων για αυτό το χαρακτηριστικό εξαρτάται από τις συνιστώσες. Αν με μεγάλη πιθανότητα η κατανομή των παρατηρήσεων του χαρακτηριστικού είναι παρόμοια για κάθε συνιστώσα, τότε το χαρακτηριστικό έχει μικρή σημαντικότητα. Για να εκπαιδεύσουν την μίξη χρησιμοποιούν το κριτήριο MML και μια παραλλαγή του αλγορίθμου EM που μπορεί να απαλείφει συνιστώσες από την μίξη. Όπως υποστηρίζουν στην [51], η μέθοδος μπορεί να θεωρηθεί σαν MAP προσέγγιση με improper prior κατανομές για τους συντελεστές μίξης και την σημαντικότητα των χαρακτηριστικών.

Στην [13] προτείνουν μια Bayesian μέθοδο. Χρησιμοποιούν μια μίξη κανονικών κατανομών για την ομαδοποίηση, και ορίζουν συζυγείς εκ των προτέρων κατανομές για όλες τις παραμέτρους της μίξης. Επιπλέον θέτουν hyperprior κατανομές στις παραμέτρους των εκ των προτέρων κατανομών των κέντρων και των συντελεστών της μίξης. Χρησιμοποιώντας μια εκ των προτέρων κατανομή για τα κέντρα που παίρνει μεγάλες τιμές όταν όλα είναι πολύ κοντά, προσπαθούν να εντοπίσουν τα ασήμαντα χαρακτηριστικά και να συγκεντρώσουν τις

αντίστοιχες τιμές των κέντρων γύρω από θέσεις κοινές για όλες τις συνιστώσες. Για την εκτίμηση των παραμέτρων της μίξης προσφεύγουν στην MAP προσέγγιση.

Στην [53] προτείνεται η εφαρμογή της ανάλυσης κυρίων συνιστωσών, και στη συνέχεια η εκπαίδευση μιας μίξης στα δεδομένα που προβάλλονται σε κάποιες από τις κύριες συνιστώσες. Για να επιλέξουν ένα υποσύνολο από τις κύριες συνιστώσες υποθέτουν ότι για ένα δεδομένο τα πρώτα k χαρακτηριστικά του ακολουθούν μια μίξη κανονικών κατανομών, και τα υπόλοιπα χαρακτηριστικά ακολουθούν κανονική κατανομή. Χειρίζονται το k σαν τυχαία μεταβλητή, και προτείνουν ένα Bayesian πλαίσιο με χρήση της μεθόδου Markov Chain Monte Carlo.

Η μέθοδος που προτείνουμε χρησιμοποιεί το ίδιο μοντέλο που προτάθηκε στην [51] για να περιγράψει την σημαντικότητα των χαρακτηριστικών, αλλά ενοποιεί την επιλογή μοντέλου και χαρακτηριστικών κάτω από ένα κοινό Bayesian πλαίσιο. Η Minimum Message Length προσέγγιση που χρησιμοποιήθηκε στην [51] βασίζεται σε ένα στατιστικό κριτήριο και υπόκειται σε διάφορες απλοποιήσεις και υποθέσεις. Η μέθοδός μας χρησιμοποιώντας, το Bayesian πλαίσιο, αναμένουμε να έχει καλύτερη επίδοση ιδιαίτερα για αραιά σύνολα δεδομένων.

Πρέπει να σημειωθεί ότι η προσέγγισή μας στην επιλογή μοντέλου υποθέτει μια στάθμιση των χαρακτηριστικών, και ο συντελεστής στάθμισης κάθε χαρακτηριστικού είναι ο ίδιος για όλες τις ομάδες των δεδομένων. Μια διαφορετική προσέγγιση που μελετήθηκε στις [31, 42] είναι η ομαδοποίηση σε υποχώρους, που υποθέτει διαφορετικούς συντελεστές στάθμισης των χαρακτηριστικών για κάθε ομάδα. Έτσι κάθε ομάδα διαφοροποιείται από τις υπόλοιπες σε ένα ιδιαίτερο υποχώρο.

4.3 Ένα Bayesian Πλαίσιο για Επιλογή Χαρακτηριστικών

Υποθέτουμε ένα σύνολο δεδομένων $X = \{x^n | n = 1, \ldots, N\}$, όπου κάθε x^n είναι ένα διάνυσμα σε ένα d-διάστατο χώρο, και θέλουμε να μοντελοποιήσουμε αυτά τα δεδομένα εκπαιδεύοντας μια μίξη κανονικών κατανομών. Επιπλέον υποθέτουμε ότι τα χαρακτηριστικά των δεδομένων είναι ανεξάρτητα δοθέντος μιας συνιστώσας της μίξης. Κάποια από τα χα-

ρακτηριστικά μπορεί να είναι ασήμαντα για την μοντελοποίηση ενώ κάποια άλλα μπορεί να είναι πιο σημαντικά. Αντί να θεωρήσουμε ότι υπάρχει μια αυστηρή διάκριση μεταξύ σημαντικών και ασήμαντων χαρακτηριστικών, θεωρούμε ότι η σημαντικότητα ενός χαρακτηριστικού μετριέται σαν μια πιθανότητα. Έτσι δοθείσης μιας συνιστώσας, θεωρούμε ότι ένα χαρακτηριστικό του x παράγεται από μια μίξη δύο μονοδιάστατων υπο-συνιστωσών, όπως προτάθηκε στη [51]. Η πρώτη υπο-συνιστώσα είναι διαφορετική για κάθε συνιστώσα της μίξης και παράγει "χρήσιμα" δεδομένα, ενώ η δεύτερη υπο-συνιστώσα είναι κοινή για όλες τις συνιστώσες και παράγει "θορυβώδη" δεδομένα.

Προτείνουμε την ενσωμάτωση αυτού του μοντέλου στο Bayesian πλαίσιο που προτάθηκε στην [24] για την εκτίμηση του αριθμού των συνιστωσών μιας μίξης. Υποθέτουμε ότι το σύνολο X έχει παραχθεί από το γραφικό μοντέλο που παρουσιάζεται στο Σχήμα 4.2. Αν η μίξη έχει J συνιστώσες τότε πρόκειται για την παρακάτω κατανομή:

$$f(x) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \varphi(x_i), \tag{4.1}$$

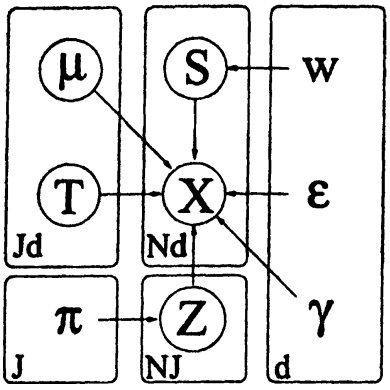
$$\varphi(x_i) = w_i \, \mathcal{N}(x_i; \mu_{ji}, \tau_{ji}) + (1 - w_i)\mathcal{N}(x_i; \varepsilon_i, \gamma_i). \tag{4.2}$$

Αυτό το γραφικό μοντέλο υποδηλώνει την εξάρτηση μιας παρατήρησης x^n από την j-οστή συνιστώσα μέσω μιας κρυμμένης μεταβλητής z_j^n, όπου $z_j^n \in \{0,1\}$ και $\sum_j z_j^n = 1$. Αν η x^n προήλθε από την j-οστή συνιστώσα, τότε η τιμή της z_j^n είναι μονάδα, διαφορετικά είναι μηδέν. Η σημαντικότητα ενός χαρακτηριστικού εκφράζεται από τις κρυμμένες μεταβλητές s_i^n, όπου $s_i^n \in \{0,1\}$. Αν η τιμή της s_i^n είναι μονάδα, τότε το i-οστό χαρακτηριστικό του x^n έχει προέλθει από την "χρήσιμη" υπο-συνιστώσα, διαφορετικά έχει προέλθει από την "θορυβώδη".

Δοθέντων των συνόλων των κρυμμένων μεταβλητών $Z = \{z_j^n\}$ και $S = \{s_i^n\}$, τα δεδομένα θεωρούνται ότι είναι ανεξάρτητα και προήλθαν από μια κανονική κατανομή

$$p(X|Z, \mu, T, S, \varepsilon, \gamma) = \prod_{n=1}^{N} \prod_{j=1}^{J} \left[\prod_{i=1}^{d} \mathcal{N}(x_i^n; \mu_{ji}, \tau_{ji})^{s_i^n} \mathcal{N}(x_i^n; \varepsilon_i, \gamma_i)^{1-s_i^n} \right]^{z_j^n}. \tag{4.3}$$

Τα σύνολα $\mu = \{\mu_{ji}\}$ και $T = \{\tau_{ji}\}$ συγκεντρώνουν τα κέντρα και τους πίνακες ακριβείας των "χρήσιμων" υπο-συνιστωσών. Αντίστοιχα, $\varepsilon = \{\varepsilon_i\}$ και $\gamma = \{\gamma_i\}$ είναι τα σύνολα παραμέτρων των "θορυβωδών" υπο-συνιστωσών. Η κατανομή των κρυμμένων μεταβλητών



Σχήμα 4.2: Γραφικό μοντέλο για την δημιουργία δεδομένων από μια Bayesian μίξη που προβλέπει θορυβώδη χαρακτηριστικά. Τα κυκλωμένα σύμβολα δηλώνουν τυχαίες μεταβλητές, διαφορετικά δηλώνουν παραμέτρους του μοντέλου. Τα πλαίσια υποδηλώνουν επανάληψη των μεταβλητών, και ο αριθμός των επαναλήψεων εμφανίζεται στην κάτω αριστερή γωνία κάθε πλαισίου.

Z δοθέντων των συντελεστών μίξης $\pi = \{\pi_j\}$, και η κατανομή των κρυμμένων μεταβλητών S δοθείσης της σημαντικότητας των χαρακτηριστικών $w = \{w_i\}$ είναι:

$$p(Z|\pi) = \prod_{n=1}^{N} \prod_{j=1}^{J} \pi_j^{z_j^n}, \tag{4.4}$$

$$p(S|w) = \prod_{n=1}^{N} \prod_{i=1}^{d} w_i^{s_i^n} (1 - w_i)^{1 - s_i^n}. \tag{4.5}$$

Η πιθανοφάνεια των παρατηρούμενων δεδομένων δοθέντων των παραμέτρων υπολογίζεται περιθωριοποιώντας της κρυμμένες μεταβλητές Z και S από την από κοινού κατανομή $p(X, Z, S|\pi, \mu, T, w, \varepsilon, \gamma)$

$$p(X|\pi, \mu, T, w, \varepsilon, \gamma) = \prod_{n=1}^{N} \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \varphi(x_i^n). \tag{4.6}$$

Αυτή είναι η συνήθης ποσότητα που βελτιστοποιείται στο πλαίσιο την μέγιστης πιθανοφάνειας. Ωστόσο αυτή η αντικειμενική συνάρτηση δεν μπορεί να χρησιμοποιηθεί για την επιλογή μοντέλου. Έτσι δεν μας βοηθάει, αφού στόχος μας είναι και ο προσδιορισμός του αριθμού συνιστωσών της μίξης. Στην [51] αυτό το πρόβλημα αντιμετωπίζεται χρησιμοποιώντας το κριτήριο Minimum Message Length, παράλληλα με μια παραλλαγή του EM που μπορεί να απαλείφει σταδιακά συνιστώσες από την μίξη. Στην μέθοδό μας υιοθετούμε

την προσέγγιση στην επιλογή μοντέλου που προτάθηκε στην [24], και περιγράφηκε στο προηγούμενο κεφάλαιο. Συγκεκριμένα, εισάγουμε εκ των προτέρων κατανομές για τα μ και T που είναι αντίστοιχα μια κανονική και μια γάμμα

$$p(\mu) = \prod_{j=1}^{J} \prod_{i=1}^{d} \mathcal{N}(\mu_{ji}; m_i, c), \tag{4.7}$$

$$p(T) = \prod_{j=1}^{J} \prod_{i=1}^{d} \mathcal{G}(\tau_{ji}; \alpha, \beta), \tag{4.8}$$

και ολοκληρώνουμε ως προς αυτές για να πάρουμε την κατανομή περιθωρίου. Οι υπερ-παράμετροι m, c, α και β ελέγχουν τις εκ των προτέρων κατανομές, και παίρνουν τιμές που ορίζουν κατανομές πλατιές και χωρίς πληροφορία. Πιο συγκεκριμένα, θέτουμε το m ίσο με την μέση τιμή όλων των δεδομένων, ενώ $c = \alpha = \beta = 10^{-16}$ που είναι μια πολύ μικρή τιμή κοντά στην ακρίβεια της μηχανής. Η μέθοδος δεν επηρεάζεται από τόσο μικρές τιμές των υπερ-παραμέτρων, γιατί έτσι δεν παρέχουν καμία πληροφορία. Όπως θα δούμε στην συνέχεια, ουσιαστικά δεν επηρεάζουν την εκτίμηση των variational εκ των υστέρων κατανομών. Παρατηρήστε ότι δεν θέσαμε εκ των προτέρων κατανομές επί των συντελεστών της μίξης π και την σημαντικότητα των χαρακτηριστικών w, και τις χειριζόμαστε σαν παραμέτρους του μοντέλου. Θέτοντας κάποιους συντελεστές της μίξης ίσους με το μηδέν μπορούμε να διαγράφουμε συνιστώσες, και αντίστοιχα μηδενίζοντας την σημαντικότητα να υποβαθμίζουμε χαρακτηριστικά.

4.4 Variational Bayesian Μάθηση για Επιλογή Χαρακτηριστικών

Για να απλοποιήσουμε τους συμβολισμούς ορίζουμε το $\theta = \{Z, \mu, T, S\}$ να είναι το σύνολο των τυχαίων μεταβλητών και $\vartheta = \{\pi, w, \varepsilon, \gamma\}$ το σύνολο των παραμέτρων. Η μέθοδος που προτείνουμε εκτιμά τις παραμέτρους ϑ του μοντέλου μεγιστοποιώντας την κατανομή περιθωρίου $p(X|\vartheta)$:

$$p(X|\vartheta) = \sum_{Z,S} \int p(X, \theta|\vartheta) \mathrm{d}\mu \, \mathrm{d}T, \tag{4.9}$$

ως προς τους συντελεστές μίξης π, την σημαντικότητα των χαρακτηριστικών w και τις παραμέτρους της "θορυβώδους" συνιστώσας. Θεωρώντας κατάλληλες εκ των προτέρων κατανομές για τις μεταβλητές των συνιστωσών και περιθωριοποιώντας αυτές αναμένουμε να εξομαλύνουμε την πιθανοφάνεια (4.6), και να πάρουμε μια πιθανοφάνεια περιθωρίου που είναι πιο εύρωστη ως προς την υπερ-εκπαίδευση. Αυτή η μεθοδολογία έχει προταθεί στην [24] για βελτιστοποίηση ως προς τους συντελεστές π, και την επιλογή του αριθμού των συνιστωσών μιας τυπικής μίξης κανονικών κατανομών με αξιόλογα αποτελέσματα.

Επειδή η ολοκλήρωση στην (4.9) δεν γίνεται αναλυτικά, χρησιμοποιούμε την variational προσέγγιση που προτείνει την μεγιστοποίηση ενός κάτω φράγματος $\mathcal{L}$ του λογαρίθμου της πιθανοφάνειας περιθωρίου:

$$\mathcal{L}[Q,\vartheta] = \sum_{Z,S} \int Q(\theta) \log \frac{p(X,\theta|\vartheta)}{Q(\theta)} \mathrm{d}\mu \, \mathrm{d}T \tag{4.10}$$

$$\leq \log p(X|\vartheta). \tag{4.11}$$

Το φράγμα $\mathcal{L}$ είναι συνάρτηση μιας αυθαίρετης κατανομής $Q(\theta)$ που προσεγγίζει την εκ των υστέρων κατανομή $p(\theta|X,\vartheta)$. Προκειμένου να μεγιστοποιήσουμε το $\mathcal{L}$ υιοθετούμε μια διαδικασία που εκτελεί επαναληπτικά δύο βήματα: αρχικά μεγιστοποιείται το φράγμα ως προς την Q, και στη συνέχεια ως προς το ϑ.

Σύμφωνα με την προσέγγιση mean-field, δεν υποθέτουμε κάποια συγκεκριμένη συναρτησιακή μορφή για την Q,εκτός του ότι είναι ένα γινόμενο της μορφής

$$Q(\theta) = Q_Z(Z)Q_\mu(\mu)Q_T(T)Q_S(S).$$

Μεγιστοποιώντας το $\mathcal{L}$ ως προς την συναρτησιακή μορφή των Q_Z, Q_μ, Q_T και Q_S, η τυπική συναρτησιακή ανάλυση παρέχει λύσεις με την ακόλουθη γενική μορφή:

$$Q(\theta_i) = \frac{\exp\langle P(X,\theta|\vartheta)\rangle_{k\neq i}}{\int \exp\langle P(X,\theta|\vartheta)\rangle_{k\neq i} \mathrm{d}\theta_i}, \tag{4.12}$$

όπου $\langle\cdot\rangle_{k\neq i}$ δηλώνει την αναμενόμενη τιμή ως προς τις κατανομές $Q_k(\theta_k)$ για όλα τα $k \neq i$.

Για το μοντέλο μας η (4.12) δίνει:

$$Q_Z(Z) = \prod_{n=1}^{N}\prod_{j=1}^{J} r_{jn}^{z_j^n}, \tag{4.13}$$

$$Q_\mu(\mu) = \prod_{j=1}^{J}\prod_{i=1}^{d} \mathcal{N}(\mu_{ji}; m_{ji}^v, c_{ji}^v), \tag{4.14}$$

$$Q_T(T) = \prod_{j=1}^{J}\prod_{i=1}^{d} \mathcal{G}(\tau_{ji}; \alpha_{ji}^v, \beta_{ji}^v), \tag{4.15}$$

$$Q_S(S) = \prod_{n=1}^{N}\prod_{i=1}^{d} \rho_{in}^{s_i^n}(1-\rho_{in})^{1-s_i^n}. \tag{4.16}$$

Οι variational παράμετροι $r_{jn}, m_{ji}^v, c_{ji}^v, \alpha_{ji}^v, \beta_{ji}^v$ και ρ_{in} προκύπτουν από την μεγιστοποίηση και ρυθμίζουν τις κατανομές που παραγοντοποιούν την Q. Οι variational παράμετροι με την σειρά τους ορίζονται χρησιμοποιώντας τις μέσες τιμές των z_j^n, μ_{ji}, τ_{ji}, s_i^n και συναρτήσεις αυτών. Χρησιμοποιώντας την συναρτησιακή μορφή των Q_Z, Q_μ, Q_T και Q_S μπορούμε να βρούμε τις αντίστοιχες μέσες τιμές, και να τις χρησιμοποιήσουμε στους ορισμούς των variational παραμέτρων. Μέτα από πράξεις παίρνουμε τις επόμενες εξισώσεις:

$$r_{jn} = \frac{\pi_j \tilde{r}_{jn}}{\sum_{j=1}^{J} \pi_j \tilde{r}_{jn}}, \tag{4.17}$$

$$\tilde{r}_{jn} = \exp\left\{\frac{1}{2}\sum_{i=1}^{d}\rho_{in}\left[\psi(\alpha_{ji}^v) - \log\beta_{ji}^v\right] - \frac{1}{2}\sum_{i=1}^{d}\rho_{in}\frac{\alpha_{ji}^v}{\beta_{ji}^v}\left[\left(x_i^n - m_{ji}^v\right)^2 + \frac{1}{c_{ji}^v}\right]\right\}, \tag{4.18}$$

$$m_{ji}^v = \frac{c\, m_i + (\alpha_{ji}^v/\beta_{ji}^v)\sum_{n=1}^{N} r_{jn}\rho_{in}x_i^n}{c + (\alpha_{ji}^v/\beta_{ji}^v)\sum_{n=1}^{N} r_{jn}\rho_{in}}, \tag{4.19}$$

$$c_{ji}^v = c + \frac{\alpha_{ji}^v}{\beta_{ji}^v}\sum_{n=1}^{N} r_{jn}\rho_{in}, \tag{4.20}$$

$$\alpha_{ji}^v = \alpha + \frac{1}{2}\sum_{n=1}^{N} r_{jn}\rho_{in}, \tag{4.21}$$

$$\beta_{ji}^v = \beta + \frac{1}{2}\sum_{n=1}^{N} r_{jn}\rho_{in}\left[\left(x_i^n - m_{ji}^v\right)^2 + \frac{1}{c_{ji}^v}\right], \tag{4.22}$$

$$\rho_{in} = \frac{w_i \tilde{\rho}_{in}}{w_i \tilde{\rho}_{in} + (1 - w_i)\xi_{in}}, \tag{4.23}$$

$$\tilde{\rho}_{in} = \exp\left\{ \frac{1}{2} \sum_{j=1}^{J} r_{jn} \left[\psi(\alpha_{ji}^v) - \log \beta_{ji}^v \right] - \frac{1}{2} \sum_{j=1}^{J} r_{jn} \frac{\alpha_{ji}^v}{\beta_{ji}^v} \left[(x_i^n - m_{ji}^v)^2 + \frac{1}{c_{ji}^v} \right] \right\}, \tag{4.24}$$

$$\xi_{in} = \exp\left\{ -\frac{1}{2} \gamma_i (x_i^n - \varepsilon_i)^2 + \frac{1}{2} \log \gamma_i \right\}, \tag{4.25}$$

όπου $\psi(x) = \mathrm{d} \log \Gamma(x)/\mathrm{d}x$. Η μεγιστοποίηση του $\mathcal{L}$ ως προς την Q στοχεύει στο να βρεθεί ένα ικανοποιητικό κάτω φράγμα για την λογαριθμική πιθανοφάνεια περιθωρίου. Αν και η ακριβής μεγιστοποίηση του $\mathcal{L}$ ως προς τις variational παραμέτρους είναι αδύνατη, αφού αλληλοεξαρτώνται με μη γραμμικό τρόπο, μπορούμε να εκτιμήσουμε το φράγμα ανανεώνοντας επαναληπτικά τις παραμέτρους χρησιμοποιώντας τις (4.17) ως (4.24). Μια ανάλογη προσέγγιση ακολουθείται στην [24].

Μετά την μεγιστοποίηση του $\mathcal{L}$ ως προς την Q, το δεύτερο βήμα της μεθόδου απαιτεί την μεγιστοποίηση της $\mathcal{L}$ ως προς τις π_j, w_i, ε_i και γ_i. Μηδενίζοντας τις μερικές παραγώγους του $\mathcal{L}$ ως προς τις παραμέτρους, παίρνουμε τις ακόλουθες εξισώσεις

$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} r_{jn}, \tag{4.26}$$

$$w_i = \frac{1}{N} \sum_{n=1}^{N} \rho_{in}, \tag{4.27}$$

$$\varepsilon_i = \frac{\sum_{n=1}^{N} \rho_{in} x_i^n}{\sum_{n=1}^{N} \rho_{in}}, \tag{4.28}$$

$$\frac{1}{\gamma_i} = \frac{\sum_{n=1}^{N} \rho_{in} (x_i^n - \varepsilon_i)^2}{\sum_{n=1}^{N} \rho_{in}}. \tag{4.29}$$

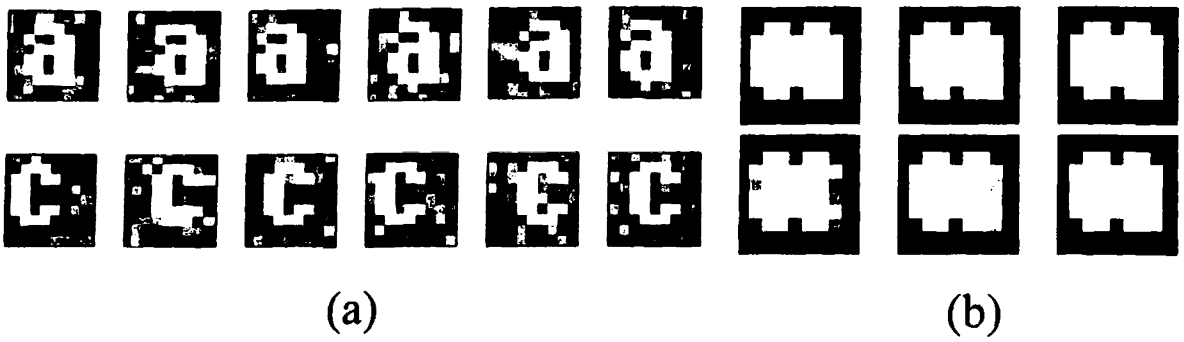
Η παραπάνω διαδικασία των δύο βημάτων επαναλαμβάνεται μέχρι να συγκλίνει. Η σύγκλιση μπορεί να ελεγχθεί εξετάζοντας το variational φράγμα. Μια σημαντική ιδιότητα αυτής της διαδικασίας είναι ότι δεν επιτρέπει σε δυο κανονικές συνιστώσες με παρόμοιες παραμέτρους να περιγράφουν την ίδια ομάδα δεδομένων. Έτσι μια από τις δύο επικρατεί και η άλλη απαλείφεται. Ξεκινώντας με ένα μεγάλο αριθμό συνιστωσών, ο ανταγωνισμός μεταξύ τους έχει σαν αποτέλεσμα ένα μοντέλο όπου οι περιττές συνιστώσες έχουν απορριφθεί. Ταυτόχρονα η ανανέωση των w_i επιτρέπει τον υπολογισμό της σημαντικότητας των

χαρακτηριστικών.

4.5 Πειράματα

Συγκρίναμε τη μέθοδό μας (την καλούμε varFnMS) με την μέθοδο που έχει προταθεί στην [51] (την καλούμε FnMs) για την ομαδοποίηση τεχνητών και πραγματικών συνόλων δεδομένων υψηλής διάστασης. Εκτελέσαμε επίσης τα ίδια πειράματα με την μέθοδο που έχει προταθεί στην [24] (την καλούμε varMS). Η πρώτη σειρά πειραμάτων αφορούσε την ομαδοποίηση τεχνητών δεδομένων. Πιο συγκεκριμένα, δημιουργήσαμε 9×9 εικόνες με διαβαθμίσεις του γκρίζου, όπου κάθε μια περιείχε το σχήμα του χαρακτήρα "a" ή "c". Το σχήμα σε κάθε εικόνα τοποθετήθηκε σε μία από τις τρεις δυνατές θέσεις, έτσι ώστε 41 εικονοστοιχεία κατά μήκος του ορίου της εικόνας ανήκαν πάντα στο υπόβαθρο. Οι εντάσεις των εικονοστοιχείων του υποβάθρου προήλθαν με δειγματοληψία από μια κανονική κατανομή $\mathcal{N}(0.4, 12 \cdot 10^{-3})$, και οι εντάσεις των εικονοστοιχείων του προσκηνίου από μια κανονική κατανομή $\mathcal{N}(0.85, 0.4 \cdot 10^{-3})$. Στην συνέχεια όλες οι εντάσεις κανονικοποιήθηκαν στο $[0, 1]$. Στο Σχήμα 4.3(a) εμφανίζονται μερικές από τις εικόνες που χρησιμοποιήθηκαν. Είναι προφανές ότι υπάρχουν έξι ομάδες στα δεδομένα (τρεις θέσεις $\times$ δυο γράμματα), και ότι τουλάχιστον 41 εικονοστοιχεία δεν περιέχουν πληροφορία χρήσιμη για την ομαδοποίηση, δηλαδή αντιστοιχούν σε ασήμαντα χαρακτηριστικά. Εφαρμόσαμε τις τρεις μεθόδους σε σύνολα με διαφορετικό συνολικό αριθμό εικόνων κάθε φορά, χρησιμοποιώντας πάντα ίσο αριθμό εικόνων για όλες τις ομάδες. Για κάθε σύνολο εκπαίδευσης εκτελέσαμε το πείραμα 10 φορές, εκπαιδεύοντας μια μίξη με 30 αρχικές συνιστώσες. Για σύνολα με 180, 240 και 300 εικόνες η μέθοδός μας εντόπισε σωστά τις έξι ομάδες 4, 10 και 10 φορές αντίστοιχα. Η μέθοδος FnMs εντόπισε τις έξι ομάδες 0, 5 και 10 φορές αντίστοιχα, επηρεασμένη από την μείωση του πλήθους των εικόνων. Το Σχήμα 4.3(b) προσφέρει μια οπτική αναπαράσταση τις σημαντικότητας των χαρακτηριστικών όπως προσδιορίστηκε από τις δυο μεθόδους. Η μέθοδος varMS δεν κατάφερε καμία φορά να εκτιμήσει των σωστό αριθμό των ομάδων, βρίσκοντας κατά μέσο όρο 12 ομάδες και για τα τρία σύνολα εικόνων.

Για τα πειράματα σε πραγματικά σύνολα δεδομένων χρησιμοποιήσαμε την βάση "mul-



(a) (b)

Σχήμα 4.3: (a) Ένα δείγμα από τις τεχνητές εικόνες. (b) Η σημαντικότητα των χαρακτηριστικών όπως εκτιμήθηκε από την varFnMS στην πάνω σειρά, και από την FnMS στην κάτω. Από αριστερά προς τα δεξιά, τα αποτελέσματα για τα σύνολα με 180, 240 και 300 εικόνες αντίστοιχα.

tiple feature database" που χρησιμοποιήθηκε στην [47], και είναι διαθέσιμη από το UCI repository [11]. Αποτελείται από τα χαρακτηριστικά χειρόγραφων ψηφίων ("0"–"9"), όπου για κάθε κατηγορία υπάρχουν 200 παραδείγματα δίνοντας συνολικά 2000 εικόνες. Τα ψηφία αναπαριστώνται με διάφορα σύνολα χαρακτηριστικών. Εμείς χρησιμοποιήσαμε τρία σύνολα δεδομένων, το πρώτο περιγράφει τα ψηφία με Zernike moments (47 χαρακτηριστικά), το δεύτερο με συντελεστές Fourier (76 χαρακτηριστικά) και το τρίτο με profile correlations (216 χαρακτηριστικά). Η επίδοση των αλγορίθμων στην ομαδοποίηση μετρήθηκε με το "σφάλμα ταξινόμησης" σε ένα ανεξάρτητο σύνολο ελέγχου. Για να το υπολογίσουμε για δοθείσα ομαδοποίηση του συνόλου εκπαίδευσης, αναθέτουμε σε κάθε ομάδα την κατηγορία που αντιστοιχεί στην πλειοψηφία των δεδομένων της. Στην συνέχεια αναθέτουμε σε κάθε δεδομένο ελέγχου την κατηγορία της ομάδας στην οποία ανήκει και τη συγκρίνουμε με την πραγματική κατηγορία του. Για να εκτιμήσουμε την αναμενόμενη τιμή του σφάλματος και τον αναμενόμενο αριθμό συνιστωσών εκτελέσαμε κάθε πείραμα 20 φορές, χωρίζοντας το σύνολο των δεδομένων στην μέση και διατηρώντας την αναλογία των κατηγοριών για να δημιουργήσουμε τα σύνολα εκπαίδευσης και ελέγχου. Τα αποτελέσματα συγκεντρώνονται στους Πίνακες 4.1–4.3 για την εκπαίδευση μιας μίξης με 30, 50 και 60 αρχικές συνιστώσες αντίστοιχα.

Η μέθοδός μας εμφανίζει μικρότερο σφάλμα, αλλά χρησιμοποιεί περισσότερες συνιστώσες από την FnMS. Όμως και οι δύο μέθοδοι καταλήγουν πάντα σε παραπλήσιο αριθμό

Πίνακας 4.1: Αναμενόμενο σφάλμα και αριθμός συνιστωσών χρησιμοποιώντας varFnMS, varMS καιFnMs, με 30 αρχικές συνιστώσες. Σε παρένθεση η αντίστοιχη τυπική απόκλιση.

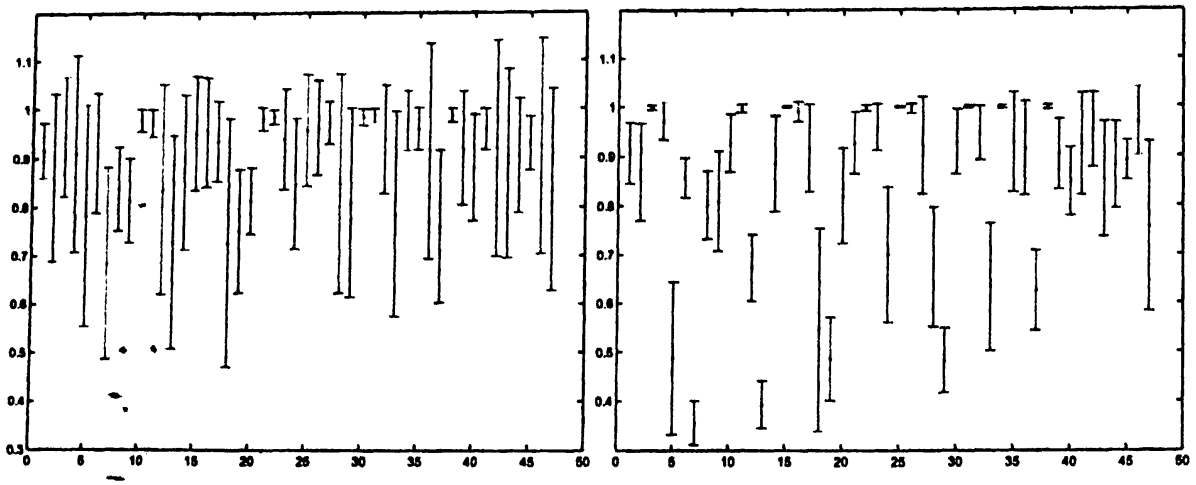
		Zernike	*Fourier*	*Profile*
varFnMS	*σφάλμα*	0.39 (0.07)	0.35 (0.06)	0.13 (0.01)
	συνιστ.	26.8 (6.4)	24.6 (7.2)	26.3 (3.7)
varMS	*σφάλμα*	0.37 (0.02)	0.34 (0.02)	0.14 (0.01)
	συνιστ.	29.5 (0.5)	27.5 (1.2)	27.6 (1.5)
FnMS	*σφάλμα*	0.53 (0.02)	0.50 (0.07)	0.77 (0.04)
	συνιστ.	10.7 (1.2)	6.1 (0.9)	2.3 (0.7)

Πίνακας 4.2: Αναμενόμενο σφάλμα και αριθμός συνιστωσών χρησιμοποιώντας varFnMS, varMS καιFnMs, με 50 αρχικές συνιστώσες. Σε παρένθεση η αντίστοιχη τυπική απόκλιση.

		Zernike	*Fourier*	*Profile*
varFnMS	*σφάλμα*	0.37 (0.03)	0.32 (0.02)	0.12 (0.01)
	συνιστ.	28.1 (2.3)	25.0 (1.6)	29.6 (2.4)
varMS	*σφάλμα*	0.35 (0.02)	0.31 (0.02)	0.11 (0.01)
	συνιστ.	44.6 (1.9)	37.6 (2.9)	41.3 (2.1)
FnMS	*σφάλμα*	0.53 (0.01)	0.52 (0.03)	0.76 (0.04)
	συνιστ.	10.8 (1.1)	5.7 (0.7)	2.3 (0.4)

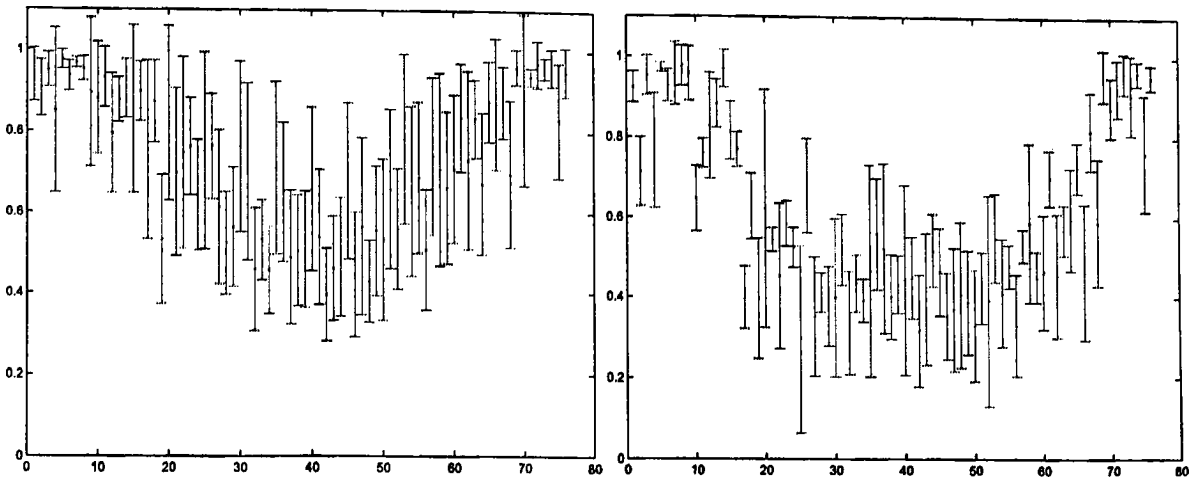
Πίνακας 4.3: Αναμενόμενο σφάλμα και αριθμός συνιστωσών χρησιμοποιώντας varFnMS, varMS καιFnMs, με 60 αρχικές συνιστώσες. Σε παρένθεση η αντίστοιχη τυπική απόκλιση.

		Zernike	*Fourier*	*Profile*
varFnMS	*σφάλμα*	0.37 (0.03)	0.30 (0.02)	0.11 (0.01)
	συνιστ.	32.8 (2.0)	29.5 (2.7)	35.3 (2.5)
varMS	*σφάλμα*	0.33 (0.01)	0.30 (0.02)	0.10 (0.01)
	συνιστ.	50.6 (2.3)	41.7 (3.1)	46.7 (2.7)
FnMS	*σφάλμα*	0.53 (0.2)	0.52 (0.04)	0.77 (0.04)
	συνιστ.	10.9 (0.9)	5.7 (0.8)	2.2 (0.4)



Σχήμα 4.4: Η σημαντικότητα των κατά Zernike moments χρησιμοποιώντας την varFnMS (αριστερά) και την FnMS (δεξιά).

συνιστωσών ανεξάρτητα από τον αρχικό τους αριθμό, οπότε οι λύσεις που δίνουν είναι διαφορετικές αλλά συνεπείς. Από την άλλη μεριά, η varMS επηρεάζεται από τον αρχικό αριθμό συνιστωσών, και έχει την τάση να διατηρεί τις περισσότερες από αυτές. Επίσης ενδιαφέρον για την varFnMS είναι το ότι καθώς ο αριθμός των χαρακτηριστικών αυξάνει ο αριθμός των συνιστωσών μεταβάλλεται λίγο αλλά το σφάλμα μικραίνει αισθητά. Προφανώς η μέθοδος εκμεταλλεύεται το μεγάλο πλήθος χαρακτηριστικών για να βελτιώσει την λύση της, και δεν επηρεάζεται από αραίωση των δεδομένων. Όσο αφορά την εκτιμώμενη σημαντικότητα των χαρακτηριστικών παρουσιάζουμε ραβδογράμματα (error-bars) στα Σχήματα. 4.4–4.6 για μίξεις που αρχικοποιήθηκαν με 30 συνιστώσες. Παρατηρήστε ότι στο Σχήμα 4.5 οι συντελεστές Fourier τείνουν να γίνουν λιγότερο σημαντικοί για την ομαδοποίηση καθώς πλησιάζουμε τις μεσαίες συχνότητες, και στο Σχήμα 4.7 η αναμενόμενη σημαντικότητα όπως εκτιμάται από την varFnMS παρουσιάζει ένα τοπικό ελάχιστο κάθε 12 χαρακτηριστικά. Σαν ένα γενικό σχόλιο, η FnMS παρέχει μικρότερες τιμές για την σημαντικότητα, ενώ η varFnMS είναι πιο συντηρητική.

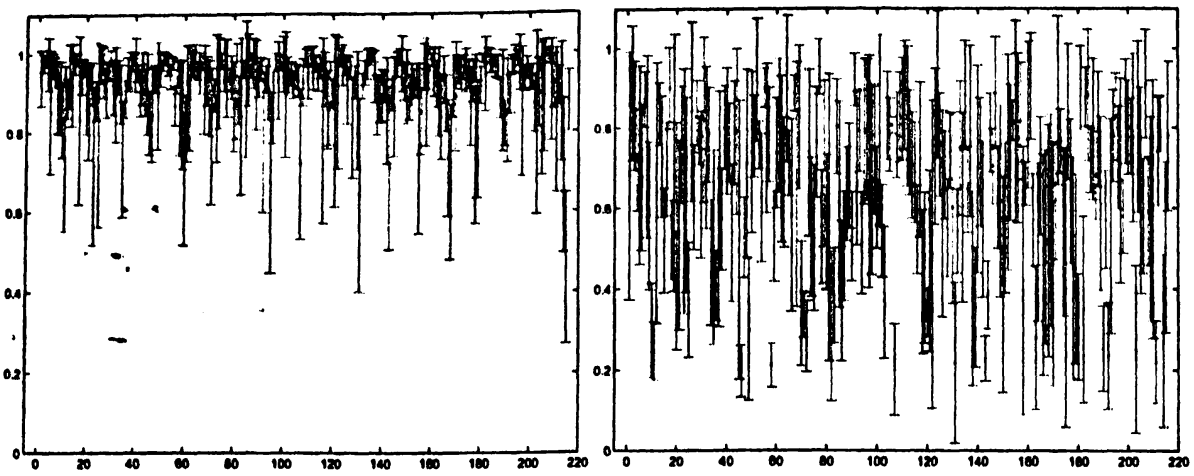


Σχήμα 4.5: Η σημαντικότητα των συντελεστών Fourier χρησιμοποιώντας την varFnMS (αριστερά) και την FnMS (δεξιά).

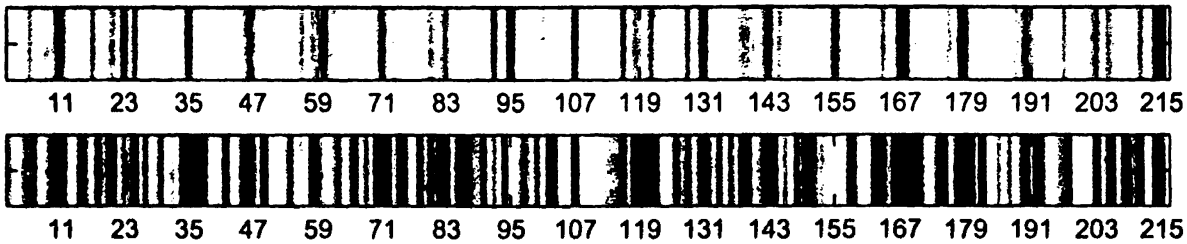
4.6 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάσαμε μια variational Bayesian μέθοδο για την εκπαίδευση μιας μίξης, η οποία μπορεί αυτόματα να καθορίζει τον αριθμό των συνιστωσών και την σημαντικότητα των χαρακτηριστικών [23]. Τα πειράματά μας έδειξαν ότι η μέθοδος υπερέχει της προσέγγισης που προτάθηκε στην [51] και που βασίζεται στο Minimum Message Length κριτήριο όταν εφαρμόζεται σε αραιά σύνολα δεδομένων, και αυτό αποτελεί ένδειξη της σπουδαιότητας της Bayesian προσέγγισης που υιοθετήσαμε. Όπως αναμέναμε, η μέθοδος που βασίζεται στο Minimum Message Length κριτήριο χρειάζεται περισσότερα δεδομένα για να εκμεταλλευτεί πλήρως το μοντέλο για την σημαντικότητα των χαρακτηριστικών. Επίσης η μέθοδός μας παρουσιάζει πιο συνεπή συμπεριφορά από τη μέθοδο που προτάθηκε στην [24], όσον αφορά τον αριθμό των συνιστωσών που χρησιμοποιούνται. Αυτό ήταν αναμενόμενο, καθώς η τελευταία δεν εφαρμόζει επιλογή χαρακτηριστικών, και για δεδομένα υψηλής διάστασης συναντά δυσκολίες κατά την εκπαίδευση.

Ο βασικός περιορισμός της προτεινόμενης μεθόδου είναι το ότι τα χαρακτηριστικά θεωρούνται ανεξάρτητα δοθείσης μιας συνιστώσας της μίξης. Σημαντική μελλοντική κατεύθυνση έρευνας είναι η γενίκευση του μοντέλου, έτσι ώστε να χρησιμοποιεί έναν πλήρη πίνακα συνδιακύμανσης για κάθε συνιστώσα της μίξης και να είναι δυνατή ταυτόχρονα η εκτίμηση της σημαντικότητας των χαρακτηριστικών.



Σχήμα 4.6: Η σημαντικότητα των profile correlations χρησιμοποιώντας την varFnMS (αριστερά) και την FnMS (δεξιά).



Σχήμα 4.7: Η μέση τιμή της σημαντικότητας των profile correlations χρησιμοποιώντας την varFnMS (επάνω) και την FnMS (χάτω). Κάθε στήλη αντιστοιχεί σε ένα χαρακτηριστικό, και οι εντάσεις έχουν διαβαθμιστεί έτσι ώστε το μαύρο να αντιστοιχεί στην ελάχιστη αναμενόμενη σημαντικότητα και το λευκό στην μέγιστη.

Κεφαλαιο 5

Κατατμηση Εικονων με Μιξη Κανονικων Κατανομων

Η μίξη κανονικών κατανομών έχει αποδειχθεί ένα αποδοτικό εργαλείο για την μοντελοποίηση και την κατάτμηση εικόνων. Ωστόσο πολλά από τα ζητήματα που αφορούν την εκτίμηση των παραμέτρων μιας μίξης για την μοντελοποίηση εικόνων δεν έχουν βρει μια κατάλληλη λύση, όπως είναι ο αριθμός των συνιστωσών της μίξης και ο μεγάλος όγκος δεδομένων εκπαίδευσης για εικόνες με τυπική ανάλυση (π.χ. 256×256). Προτείνουμε την σταδιακή ομαδοποίηση των εικονοστοιχειων, και παρουσιάζουμε μια μέθοδο που αυξάνει σταδιακά το πλήθος των εικονοστοιχείων που χρησιμοποιούνται για την επαναλαμβανόμενη κατάτμηση της εικόνας. Σε κάθε επανάληψη της μεθόδου έχουμε ενσωματώσει έναν αλγόριθμο εκπαίδευσης της μίξης που βασίζεται στην variational Bayesian προσέγγιση, και αντιμετωπίζει αποδοτικά το πρόβλημα επιλογής μοντέλου. Τα πειραματικά αποτελέσματα έδειξαν ότι η μέθοδος έχει ικανοποιητικό χρόνο εκτέλεσης και προσαρμόζει με επιτυχία των αριθμό των περιοχών που χρησιμοποιεί στην πολυπλοκότητα της εικόνας.

5.1 Εισαγωγή

Η κατάτμηση εικόνων είναι ένα από τα κεντρικά προβλήματα της επεξεργασίας εικόνας, και αφορά την ομαδοποίηση των εικονοστοιχείων χρησιμοποιώντας κατάλληλα χαρακτηριστικά τους [63]. Το πρώτο βήμα για την κατάτμηση μιας εικόνας είναι η εξαγωγή ενός διανύσματος χαρακτηριστικών για κάθε εικονοστοιχείο, όπως το χρώμα, η υφή κ.λ.π. Στην συνέχεια επιχειρείται η ομαδοποίηση των διανυσμάτων αυτών, και οι ομάδες που καθορίζονται προβάλλονται πίσω στον χώρο των εικονοστοιχείων. Μια προσέγγιση στην ομαδοποίηση αποτελεί η χρήση μεθόδων φασματικής ομαδοποίησης (spectral clustering), όπως για παράδειγμα στην [71]. Μια διαφορετική προσέγγιση είναι η εκτίμηση της κατανομής των χαρακτηριστικών, και εξαγωγή ομάδων από τα χαρακτηριστικά αυτής της στατιστικής αναπαράστασης. Η στατιστική αναπαράσταση εικόνων μπορεί να γίνει με διάφορους τρόπους, οι πρώτες προσπάθειες στηρίζονταν σε ιστογράμματα, ενώ αργότερα παρουσιάστηκαν μέθοδοι βασισμένοι στη μίξη κατανομών, όπως για παράδειγμα στις [70, 37]. Η μίξη κατανομών έχει αναγνωριστεί σαν ένα εύχρηστο εργαλείο για την μοντελοποίηση εικόνων. Πιο συγκεκριμένα, μια εικόνα θεωρείται σαν ένα σύνολο περιοχών, όπου κάθε περιοχή αναπαριστάται από μια κανονική κατανομή, και το σύνολο όλων των περιοχών της εικόνας αναπαριστάται από την μίξη των αντίστοιχων κανονικών κατανομών.

Όταν ορίζουμε της περιοχές της εικόνας θέλουμε να είναι όσο δυνατόν πιο συμπαγείς, δηλαδή να μην έχουν πολλές οπές και να μην αποτελούνται από πολλές μικρότερες ασύνδετες περιοχές. Ένα σημαντικό ζήτημα είναι το πως θα επιβάλουμε αυτή την χωρική συνοχή, και μπορούμε να το πετύχουμε με δύο τρόπους. Ο πρώτος είναι θέτοντας μια κατάλληλη εκ των προτέρων πιθανότητα στις παραμέτρους της μίξης για να οδηγηθούμε στο Spatially Variant Mixture Model SVMM [69]. Η δεύτερη είναι μια *απευθείας αντιμετώπιση*, όπου η *θέση* κάθε εικονοστοιχείου μέσα στην εικόνα χρησιμοποιείται σαν ένα επιπλέον χαρακτηριστικό του κατά την μοντελοποίηση της εικόνας. Με αυτό τον τρόπο, για εικονοστοιχεία που βρίσκονται μακρυά στην εικόνα αυξάνεται η απόστασή τους και στον χώρο που γίνεται η ομαδοποίηση. Ενώ για εικονοστοιχεία που βρίσκονται κοντά στην εικόνα η απόστασή τους καθορίζεται μόνο από τα υπόλοιπα χαρακτηριστικά.

Σε αυτή την εργασία χρησιμοποιούμε χαρακτηριστικά χρώματος και θέσης για την περιγραφή των εικονοστοιχείων. Εξάγουμε τα χαρακτηριστικά του χρώματος αναπαριστώντας

κάθε εικονοστοιχείο με ένα τριδιάστατο διάνυσμα στον χώρο χρώματος (L, a, b). Τα χαρακτηριστικά θέσης κάθε εικονοστοιχείου αποτελούν ένα διδιάστατο διάνυσμα με στοιχεία την γραμμή και την στήλη στην οποία βρίσκεται μέσα στην εικόνα. Η αντιμετώπιση της κατάτμησης της εικόνας σαν ομαδοποίηση σημείων αυτού του πενταδιάστατου χώρου με μια μίξη κανονικών κατανομών έχει κάποιες ελκυστικές ιδιότητες. Η πρώτη είναι ότι τα χαρακτηριστικά θέσης περιθωριοποιούνται εύκολα από την κατανομή, και μπορούμε να πάρουμε έτσι μια εκτίμηση για την κατανομή μόνο για το χρώμα. Η δεύτερη είναι ότι κάθε συνιστώσα της μίξης έχει συγκεκριμένη θέση και καλύπτει συγκεκριμένο χώρο μέσα στην εικόνα. Έτσι μπορούμε να εύκολα να εξάγουμε μια μίξη για την κατανομή του χρώματος σε μια αυθαίρετη περιοχή, χρησιμοποιώντας τις συνιστώσες που είναι ενεργές σε αυτή την περιοχή. Τέλος μπορούμε να αναθέσουμε τιμές στα χαρακτηριστικά χρώματος σε περιοχές της εικόνας που δεν έχουν χρησιμοποιηθεί κατά την ομαδοποίηση. Περιθωριοποιώντας τα χαρακτηριστικά χρώματος από την μίξη, μπορούμε υπολογίσουμε για τα εικονοστοιχεία της άγνωστης περιοχής την συνιστώσα με την μεγαλύτερη εκ των υστέρων πιθανότητα. Στη συνέχεια τους αναθέτουμε τα χαρακτηριστικά χρώματος που έχει το κέντρο της υπεύθυνης συνιστώσας. Αυτό μας επιτρέπει να εκμεταλλευτούμε την πλεονάζουσα πληροφορία των εικονοστοιχείων, και να εκπαιδεύσουμε την μίξη χρησιμοποιώντας μόνο ένα αντιπροσωπευτικό υποσύνολό τους. Τα υπόλοιπα εικονοστοιχεία μπορούμε να τα χρησιμοποιήσουμε σαν ένα *σύνολο ελέγχου* για να εκτιμήσουμε την "επίδοση κατάτμησης" υπολογίζοντας την απόσταση μεταξύ της αρχικής εικόνας και της "κατατμημένης", που προκύπτει θέτοντας σε όλα τα εικονοστοιχεία που ανήκουν στην ίδια συνιστώσα το χρώμα του κέντρου της. Αναπτύσσοντας περισσότερο αυτήν την ιδέα προτείνουμε την χρήση μιας μεθοδολογίας "ενεργητικής κατάτμησης", όπου η εκπαίδευση ξεκινά με ένα μικρό υποσύνολο των δεδομένων και σταδιακά προστίθενται επιπλέον κατάλληλα επιλεγμένα δεδομένα καθώς η μάθηση προχωρά. Με αυτό τον τρόπο είναι δυνατό να κατασκευάσουμε ένα αντιπροσωπευτικό μοντέλο για την εικόνα χρησιμοποιώντας μόνο ένα μικρό μέρος των εικονοστοιχείων.

Η προτεινόμενη μέθοδος "ενεργητικής κατάτμησης" μπορεί να υλοποιηθεί σαν ένας αλγόριθμος *περιτυλίγματος* (wrapper) γύρω από έναν αλγόριθμο ομαδοποίησης τον οποίο αντιμετωπίζει σαν "μαύρο κουτί". Ένα σημαντικό πρόβλημα του αλγόριθμου ομαδοποίησης που πρέπει να αντιμετωπιστεί είναι η επιλογή του αριθμού των ομάδων. Η χρήση της μίξης για

την ομαδοποίηση μας επιτρέπει να χρησιμοποιήσουμε κάποιες από τις μεθόδους που έχουν προταθεί για την ταυτόχρονη εκπαίδευση της μίξης και την επιλογή του αριθμού συνιστωσών της, όπως στις [24, 29]. Ωστόσο επιλέγουμε τον αλγόριθμο *αυξητικής μάθησης βασισμένης στην διάσπαση συνιστωσών* που αναπτύξαμε στην Ενότητα 3.5, αφού η σταδιακή μέθοδος προσθήκης συνιστωσών συνάδει περισσότερο με την σταδιακή προσθήκη δεδομένων.

5.2 Ενεργητική Κατάτμηση

Η μέθοδος κατάτμησης που προτείνουμε είναι επαναληπτική, και κάθε επανάληψή της αποτελείται από δυο στάδια. Στο πρώτο εφαρμόζεται ο αλγόριθμος εκπαίδευσης μια μίξης κανονικών κατανομών με ένα περιορισμένο σύνολο σημείων, τα οποία αποτελούν τα χαρακτηριστικά επιλεγμένων εικονοστοιχείων. Έτσι προκύπτει μια μίξη κανονικών κατανομών. Στο δεύτερο στάδιο ο αλγόριθμος ενεργητικής κατάτμησης χρησιμοποιεί την μίξη για κατατμήσει την εικόνα και να επιλέξει επιπλέον εικονοστοιχεία, τα χαρακτηριστικά των οποίων προστίθενται στο σύνολο εκπαίδευσης.

Ο αλγόριθμος εκπαίδευσης της μίξης που χρησιμοποιούμε ακολουθεί την variational Bayesian μέθοδο θέτοντας κατάλληλες εκ των προτέρων κατανομές στις παραμέτρους της, που επιβάλλουν τον ανταγωνισμό μεταξύ των συνιστωσών για την επικράτησή τους σε μια περιοχή. Οι συντελεστές μίξης των πλεοναζόντων συνιστωσών μικραίνουν συνεχώς, μέχρι που μηδενίζονται και οι αντίστοιχες συνιστώσες πρακτικά απαλείφονται από την μίξη. Ωστόσο ο αλγόριθμος δεν επιτρέπει σε όλες τις συνιστώσες να ανταγωνίζονται ταυτόχρονα. Σε κάθε επανάληψή του επιλέγει μια συνιστώσα την οποία διασπά κατάλληλα σε δύο νέες συνιστώσες, στις οποίες εφαρμόζεται variational Bayesian μάθηση ενώ οι υπόλοιπες διατηρούνται σταθερές. Αυτό μας δίνει την δυνατότητα να θέσουμε εκ των προτέρων κατανομές στις παραμέτρους των νέων συνιστωσών χρησιμοποιώντας πληροφορία από την συνιστώσα που διασπάστηκε. Μια λεπτομερής περιγραφή αυτής της variational Bayesian προσέγγισης με τοπική επιλογή μοντέλου παρατίθεται στο Κεφάλαιο 3.

Έστω $S = \{x_n\}$ το σύνολο των N σημείων της αρχικής εικόνας, όπου x_n ανήκει στον πενταδιάστατο χώρο που ορίζουν το χρώμα (L, a, b) και η θέση (γραμμή, στήλη) των

εικονοστοιχείων. Η εφαρμογή της μεθόδου εκπαίδευσης με αυτό το σύνολο δεδομένων μας δίνει μια μίξη J κανονικών κατανομών

$$p(x_n) = \sum_{j=1}^{J} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j) \tag{5.1}$$

όπου μ_j είναι το κέντρο και Σ_j ο πίνακας συνδιακύμανσης της j-οστής συνιστώσας, και οι συντελεστές της μίξης π_j είναι θετικοί και αθροίζουν στην μονάδα. Για κάθε συνιστώσα j της μίξης μπορούμε να υπολογίσουμε την υπευθηνότητά της δοθέντος ενός σημείου x_n, δηλαδή την εκ των υστέρων πιθανότητα

$$p(j|x_n) = \frac{\pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}{p(x_n)}. \tag{5.2}$$

Στη συνέχεια μπορούμε να ομαδοποιήσουμε τα δεδομένα εκπαίδευσης σε J ομάδες $S_1, \ldots, S_J$, όπου ένα σημείο x_n ανήκει στην $S_{\hat{j}}$ αν

$$\hat{j} = \arg\max_j p(j|x_n). \tag{5.3}$$

Η κατάτμηση της εικόνας χρησιμοποιώντας την μίξη p μπορεί να θεωρηθεί ως μια απεικόνιση $f : x_n \in S \rightarrow f(x_n; p) \in S^*$, σύμφωνα με την οποία για να κατασκευάσουμε την κατατμημένη εικόνα αναθέτουμε σε κάθε σημείο $x_n \in S$ που ανήκει στην ομάδα S_j το χρώμα που περιέχεται στο μ_j, έτσι δημιουργούμε το σύνολο S^* των σημείων της καταμημένης εικόνας.

Ο αλγόριθμος εκπαίδευσης της μίξης μπορεί να εφαρμοστεί στο σύνολο της εικόνας, ωστόσο για να επιταχύνουμε την εκτέλεσή του εφαρμόζουμε μια διαφορετική προσέγγιση. Πιο συγκεκριμένα, επιλέγουμε ένα μικρό αντιπροσωπευτικό σύνολο σημείων, όπου η δειγματοληψία γίνεται ομοιόμορφα ως προς την θέση τους. Με τα σημεία αυτά εκπαιδεύουμε μια μίξη κανονικών κατανομών. Στην συνέχεια επιλέγουμε επιπλέον σημεία με ένα κατάλληλο κριτήριο, και τα προσθέτουμε στο σύνολο εκπαίδευσης. Μετά επαναλαμβάνουμε την εκπαίδευση της μίξης στο επαυξημένο σύνολο σημείων, αρχικοποιώντας τις παραμέτρους της μίξης στις τιμές που προέκυψαν από την προηγούμενη εκτέλεση του αλγόριθμου εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται συνεχώς, μέχρι να ικανοποιηθεί ένα κριτήριο τερματισμού.

Η επιλογή των σημείων που προστίθονται στο σύνολο εκπαίδευσης βασίζεται στην τετραγωνική απόσταση $e(x_n)$ των σημείων που περιγράφουν το n-οστό εικονοστοιχείο πριν και

μετά την κατάτμηση, δηλαδή $e(x_n) = ||x_n - f(x_n; p)||^2$. Σε κάθε επανάληψη προσθέτουμε τα K σημεία με την μεγαλύτερη απόσταση. Για να τερματιστεί αυτή η διαδικασία αθροίζουμε τις τετραγωνικές αποστάσεις όλων των σημείων της εικόνας, και αν το τετραγωνικό σφάλμα αυξήθηκε απορρίπτουμε την τρέχουσα μίξη και υιοθετούμε αυτή που εκπαιδεύτηκε στην προηγούμενη επανάληψη. Αν αυτό συμβεί τέσσερις φορές συνεχόμενα τότε τερματίζουμε τον αλγόριθμο.

Ο προτεινόμενος αλγόριθμος συνοψίζεται στα ακόλουθα βήματα:

1. Έστω $S = \{s_n | \, n = 1, \ldots, N\}$ το σύνολο όλων των σημείων της εικόνας.
2. Θέτουμε $t = 0$ και $E = \infty$.
3. Κατασκευάζουμε μια μίξη p_0 με μια συνιστώσα για το S.
4. Επιλέγουμε ομοιόμορφα ως προς την θέση τους K_0 σημεία του S, και δημιουργούμε το X.
5. Δημιουργούμε το σύνολο $Y = S - X$.
6. Θέτουμε $t = t+1$, και εκπαιδεύουμε μια μίξη p_t για το X αρχικοποιώντας με την p_{t-1}.
7. Υπολογίζουμε το $E_t = \sum_{s \in S} ||s - f(s; p_t)||^2$.
8. Αν $E_t \geq E_{t-1}$, τότε θέτουμε $E_t = E_{t-1}$, $p_t = p_{t-1}$, και καταγράφουμε την αποτυχία.
9. Αν αποτύχουμε τέσσερις φορές συνεχόμενα τερματίζουμε τον αλγόριθμο.
10. Για κάθε $y \in Y$ υπολογίζουμε το $e(y) = ||y - f(y; p_t)||^2$.
11. Επιλέγουμε τα K σημεία $y \in Y$ με την μεγαλύτερη απόσταση $e(y)$, και δημιουργούμε το σύνολο $Y_{\max}$.
12. Ανανεώνουμε τα $X = X \cup Y_{\max}$ και $Y = Y - Y_{\max}$.
13. Επιστρέφουμε στο βήμα 6.

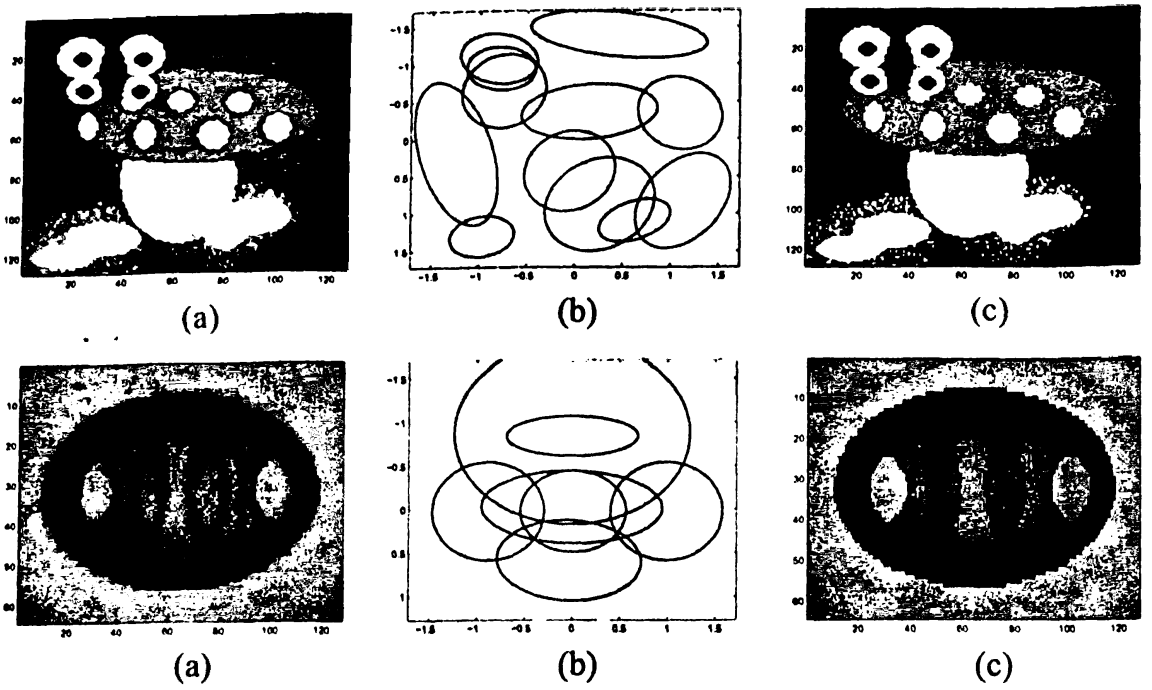
Αυτή η ενεργητική προσέγγιση στην κατάτμηση επιταχύνει σημαντικά την εκπαίδευση της μίξης, καθώς τις περισσότερες φορές χρησιμοποιείται μόνο ένα μικρό μέρος των δεδομένων για την εκτίμηση των παραμέτρων της. Επιπλέον σε κάθε επανάληψη παίρνουμε μια κατάτμηση που βασίζεται σε περισσότερα εικονοστοιχεία, έτσι προκύπτει μια διαδοχή κατατμήσεων με αυξανόμενη πληροφορία.

5.3 Πειράματα

Για να ελέγξουμε την επίδοση της προτεινόμενης μεθόδυο, διεξάγαμε πειράματα με τεχνητές και φυσικές εικόνες. Για κάθε εικόνα ακολουθήσαμε τα παρακάτω βήματα. Πρώτα δημιουργήσαμε το σύνολο με τα χαρακτηριστικά χρώματος και θέσης για όλα τα εικονοστοιχεία, και στη συνέχεια τα τυποποιήσαμε ώστε η κατανομή κάθε χαρακτηριστικού να έχει μέση τιμή μηδέν και τυπική απόκλιση ένα. Στην συνέχεια εφαρμόσαμε τον αλγόριθμο κατάτμησης που παρουσιάσαμε στην προηγούμενη ενότητα. Σε κάθε εκτέλεση, το αρχικό σύνολο εκπαίδευσης περιείχε $K_0 = 500$ σημεία, και σε κάθε επανάληψη του αλγορίθμου προσθέταμε $K = 100$ σημεία.

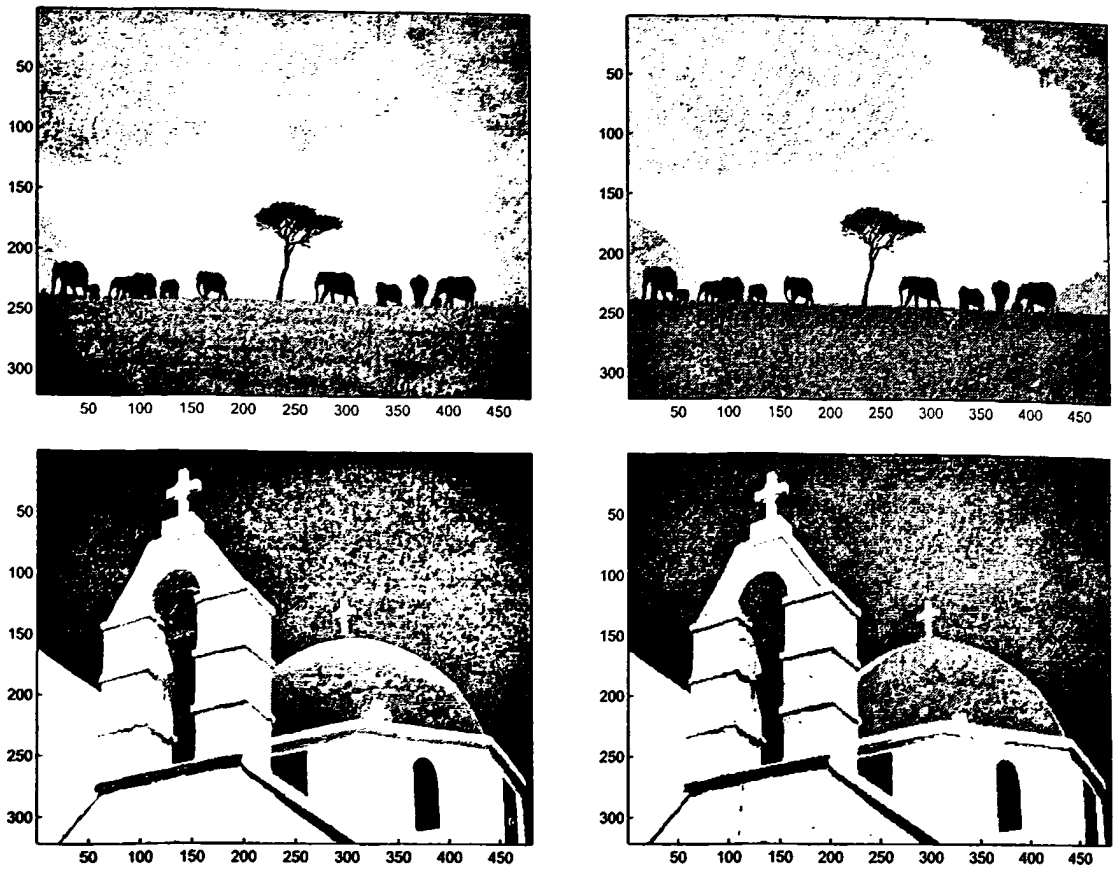
Στο Σχήμα 5.1 εμφανίζεται η κατάτμηση δυο τεχνητών εικόνων με διαστάσεις 128×128 (πάνω) και 128×64 (κάτω). Για να εμφανίσουμε τις συνιστώσες της μίξης και τα σημεία που επιλέχθηκαν χρησιμοποιήσαμε μόνον τα χαρακτηριστικά θέσης των σημείων. Για την επίδειξη της κατάτμησης σε φυσικές εικόνες, χρησιμοποιήσαμε δύο εικόνες από την βάση εικόνων Berkley Segmentation Data Set (BSDS) [55] με διαστάσεις 481×321. Το Σχήμα 5.2 παρουσιάζει την κατάτμηση της εικόνας #253036 (πάνω) και #118035 (κάτω). Η κατάτμηση της πρώτης εικόνας έγινε με μια μίξη 13 κανονικών συνιστωσών χρησιμοποιώντας 1000 σημεία. Η κατάτμηση της δεύτερης εικόνας έγινε με μια μίξη 8 κανονικών συνιστωσών χρησιμοποιώντας 800 σημεία.

Ο Πίνακας 5.1 συγκεντρώνει το τετραγωνικό σφάλμα και τον αριθμό των ομάδων για κάθε επανάληψη. Για λόγους σύγκρισης εμφανίζονται και οι μετρήσεις για μια παραλλαγή του αλγορίθμου κατάτμησης, όπου τα σημεία επιλέγονται ομοιόμορφα σε κάθε του επανάληψη. Όπως φαίνεται η προτεινόμενη μέθοδος χρησιμοποιεί περισσότερα σημεία και περισσότερες



Σχήμα 5.1: Δύο τεχνητές εικόνες. (a) Αρχική εικόνα. (b) Τα επιλεγμένα σημεία και οι συνιστώσες της μίξης. (c) Η κατάτμηση της εικόνας χρησιμοποιώντας την μίξη.

ομάδες για την κατάτμησή τους. Όπως αναμενόταν μείωσε πολύ το τετραγωνικό σφάλμα, σχεδόν στο μισό της αρχικής του τιμής. Αντίθετα η τυχαία μέθοδος επιλογής σταμάτησε πολύ γρήγορα, χρησιμοποιώντας λίγες ομάδες και σημεία. Η κατάτμηση της πρώτης εικόνας έγινε με μια μίξη 5 κανονικών συνιστωσών χρησιμοποιώντας 700 σημεία. Η κατάτμηση της δεύτερης εικόνας έγινε με μια μίξη 4 κανονικών συνιστωσών χρησιμοποιώντας 600 σημεία. Ταυτόχρονα το τετραγωνικό σφάλμα μειώθηκε πολύ λίγο. Αυτό αποτελεί ένδειξη ότι επιλέγοντας ομοιόμορφα λίγα σημεία η πληροφορία που έχουμε για την εικόνα δεν αυξάνεται αρκετά για να βελτιώσουμε την εκτίμησή μας. Προφανώς χρειάζεται κάθε φορά να επιλέγουμε τυχαία πολύ περισσότερα στοιχεία για να πάρουμε μια πιο λεπτομερή κατάτμηση. Σημειώνουμε ότι το συνολικό πλήθος των στοιχείων κάθε εικόνας ήταν της τάξης των 140000, ωστόσο είχαμε την λύση για την κατάτμηση της εικόνας σε μερικά λεπτά χρησιμοποιώντας έναν τυπικό προσωπικό υπολογιστή.



Σχήμα 5.2: Δύο φυσικές εικόνες από την βάση BSDS. Στην αριστερή στήλη οι αρχικές εικόνες, και στη δεξιά στήλη οι κατατμημένες εικόνες.

Πίνακας 5.1: Σύγκριση της προτεινόμενης μεθόδου επιλογής με την ομοιόμορφη τυχαία επιλογή. Για κάθε επανάληψη εμφανίζεται το τετραγωνικό σφάλμα και σε παρένθεση ο αριθμός των ομάδων.

BSDS #253036				*BSDS* #118035			
ενεργητική		*τυχαία*		*ενεργητική*		*τυχαία*	
1516.2	(4)	1516.2	(4)	1143.7	(4)	1143.7	(4)
1163.2	(6)	1300.6	(5)	1017.5	(5)	1047.7	(4)
1099.0	(7)	1276.8	(5)	840.3	(6)		
903.6	(9)			827.4	(8)		
800.5	(11)						
744.5	(13)						

5.4 Συμπεράσματα

Αντιμετωπίσαμε το πρόβλημα της κατάτμησης εικόνων με την χρήση μιας μίξης κανονικών κατανομών σαν πρόβλημα ομαδοποίησης, και προτείναμε μια επαναληπτική μέθοδο κατάτμησης που εκμεταλλεύεται ένα υποσύνολο των εικονοστοιχείων. Σε κάθε επανάληψη η μέθοδος βελτιώνει την τρέχουσα ομαδοποίηση επιλέγοντας με ένα κατάλληλο κριτήριο επιπλέον εικονοστοιχεία, τα οποία προσθέτει στο σύνολο εκπαίδευσης της μίξης. Παράλληλα, χρησιμοποιώντας την variational Bayesian μεθοδολογία του Κεφαλαίου 3 για την εκπαίδευση της μίξης αντιμετωπίσαμε επιτυχώς το πρόβλημα της επιλογής του αριθμού των συνιστωσών, και κατά συνέπεια του αριθμού των περιοχών της κατάτμησης. Η πειραματική εφαρμογή της μεθόδου σε ένα διάφορες εικόνες έδωσε ικανοποιητικά αποτελέσματα, καθώς ήταν γρήγορη χωρίς να θυσιάζει την ποιότητα της κατάτμησης.

Ένα ζήτημα που αξίζει περισσότερη μελέτη είναι το κριτήριο επιλογής των στοιχείων που προσθέτουμε στο σύνολο εκπαίδευσης. Προτείναμε την χρήση της απόστασης μεταξύ της αρχικής και της κατατμημένης εικόνας, και μια προφανής εναλλακτική επιλογή είναι η διαφορά της πιθανοφάνειας. Ωστόσο θα πρέπει να δοθεί βάρος σε κριτήρια που επικεντρώνονται την ποιότητα κατάτμησης. Επίσης η μέθοδος μπορεί να εφαρμοστεί χωρίς αλλαγή και για επιπλέον χαρακτηριστικά εκτός του χρώματος, όπως είναι η υφή. Η ενσωμάτωσή τέτοιων χαρακτηριστικών ίσως βελτιώσει το κριτήριο για την ενεργή επιλογή εικονοστοιχείων. Ένα άλλο χαρακτηριστικό της μεθόδου που αξίζει περισσότερη προσοχή είναι ότι παράγει μια διαδοχή από κατατμήσεις αυξανόμενης πολυπλοκότητας. Στις εφαρμογές ομαδοποίησης συνόλου εικόνων και αναζήτησης εικόνων σε βάσεις είναι συνηθισμένη τακτική η σύγκριση των εικόνων να βασίζεται σε σύγκριση των μίξεων που μοντελοποιούν την κατανομή τους, όπως για παράδειγμα στις [70, 37]. Μπορούμε ίσως να εκμεταλλευτούμε την αλληλουχία των αναπαραστάσεων που δίνει η προτεινόμενη μέθοδος για να δώσουμε έμφαση είτε στην ταχύτητα είτε στην λεπτομέρεια των απαιτούμενων συγκρίσεων. Μια άλλη εκδοχή είναι η δημιουργία μιας ιεραρχίας από ομαδοποιήσεις που προκύπτει χρησιμοποιώντας αναπαραστάσεις διαφορετικής πολυπλοκότητας για κάθε επίπεδο.

Κεφαλαιο 6

Αυξητικη Μαθηση με το Νευρωνικο Δικτυο PRBF

Το πιθανοτικό δίκτυο ακτινικών συναρτήσεων βάσης (PRBF) αποτελεί μια πιθανοτική παραλλαγή του νευρωνικού δικτύου RBF για προβλήματα ταξινόμησης, η οποία επεκτείνει την τυπική μίξη κανονικών κατανομών επιτρέποντας σε όλες τις κατηγορίες να μοιράζονται τις ίδιες συνιστώσες της μίξης. Η τυπική μέθοδος εκπαίδευσης του PRBF για ταξινόμηση χρησιμοποιεί τον αλγόριθμο EM και το τελικό δίκτυο εξαρτάται από την αρχικοποίησή του. Στο παρόν κεφάλαιο παρουσιάζουμε μια μέθοδο για την αυξητική εκπαίδευση του δικτύου PRBF [22]. Η προτεινόμενη μέθοδος ξεκινά με μια μόνο συνιστώσα, και σταδιακά προσθέτει περισσότερες συνιστώσες σε κατάλληλα επιλεγμένες θέσεις στο χώρο δεδομένων. Η προσθήκη μιας επιπλέον συνιστώσας βασίζεται σε ένα κριτήριο για τον εντοπισμό μιας περιοχής του χώρου δεδομένων που είναι σημαντική από την σκοπιά της ταξινόμησης. Μετά από την προσθήκη όλων των συνιστωσών, η μέθοδος διασπά κάθε συνιστώσα σε υπο-συνιστώσες που κάθε μια αντιστοιχεί σε διαφορετική κατηγορία.

6.1 Εισαγωγή

Ένα από τα θεμελιώδη προβλήματα στην μηχανική μάθηση είναι η ταξινόμηση. Πρόκειται για την δημιουργία ενός ταξινομητή για πρότυπα άγνωστης κατηγορίας, δοθέντος ενός συνόλου εκπαίδευσης με ταξινομημένα πρότυπα, δηλαδή πρότυπα γνωστής κατηγορίας. Κάθε πρότυπο ανήκει σε μια κατηγορία, και όλες οι δυνατές κατηγορίες είναι γνωστές. Μια στατιστική προσέγγιση στο πρόβλημα της ταξινόμησης είναι η δημιουργία ενός μοντέλου που εκτιμά την κατανομή $p(x|k)$ των προτύπων δοθείσης της κατηγορίας k και αντίστοιχη εκ των προτέρων κατανομή $P(k)$ για κάθε κατηγορία k. Στην συνέχεια, χρησιμοποιώντας το θεώρημα του Bayes υπολογίζονται οι εκ των υστέρων κατανομές $P(k|x)$:

$$P(k|x) = \frac{p(x|k)P(k)}{\sum_{\ell} p(x|\ell)P(\ell)}. \tag{6.1}$$

Προκειμένου να ταξινομηθεί ένα άγνωστο πρότυπο x, σύμφωνα με τον κανόνα ταξινόμησης του Bayes επιλέγεται η κατηγορία με την μέγιστη εκ των υστέρων πιθανότητα $P(k|x)$.

Μια συμβατική προσέγγιση στην εκτίμηση των κατανομών δοθείσης της κατηγορίας είναι κάθε κατανομή κατηγορίας $p(x|k)$ να εκτιμάται χρησιμοποιώντας μια ξεχωριστή μίξη κατανομών. Όπως έχει ήδη αναφερθεί στα προηγούμενα κεφάλαια, μια μίξη κατανομών [58] είναι ένας γραμμικός συνδυασμός q κατανομών, όπου $q(x) = \sum_{j=1}^{J} w_j f_j(x)$. Οι συντελεστές μίξη w_j είναι μη-αρνητικοί και αθροίζουν στην μονάδα, ενώ οι συνιστώσες f_j της μίξης είναι συνήθως κανονικές κατανομές. Αυτή η μέθοδος εκτιμά την κατανομή κάθε κατηγορίας ανεξάρτητα από τις άλλες, θωρώντας μόνο τα πρότυπα εκπαίδευσης της συγκεκριμένης κατηγορίας. Στην περίπτωση που υποθέσουμε μια κανονική συνιστώσα κεντραρισμένη σε κάθε πρότυπο, τότε έχουμε το πιθανοτικό μοντέλο που προτάθηκε στην [73]. Αυτό το μοντέλο υποθέτει πάρα πολλές συνιστώσες, δεν απαιτεί ιδιαίτερη εκπαίδευση και η επίδοσή του εξαρτάται από τον ευρετικό καθορισμό τις ακτίνας των κανονικών συνιστωσών.

Το δίκτυο PRBF [76, 77, 79] αποτελεί μια εναλλακτική προσέγγιση για την εκτίμηση των κατανομών κατηγορίας. Είναι ένα νευρωνικό δίκτυο RBF[10] προσαρμοσμένο να δίνει στις k εξόδους του τις αντίστοιχες κατανομές κατηγορίας $p(x|k)$. Αφού πρόκειται για δίκτυο τύπου RBF, οι συνιστώσες του (συναρτήσεις βάσης) μοιράζονται μεταξύ των κατηγοριών, και κάθε κατανομή κατηγορίας εκτιμάται χρησιμοποιώντας όλα τα διαθέσιμα πρότυπα και όχι μόνο τα πρότυπα της συγκεκριμένη κατηγορίας, όπως γίνεται σε μια τυπική μίξη κανονικών

κατανομών. Προκειμένου να εκπαιδεύσουμε το δίκτυο εφαρμόζουμε τον αλγόριθμο ΕΜ για την μεγιστοποίηση της πιθανοφάνειας των δεδομένων ως προς τις παραμέτρους του δικτύου [26, 57, 10, 77]. Επιπλέον σύμφωνα με την [78] η γενικευτική ικανότητα του δικτύου βελτιώνεται αν μετά από την εκπαίδευση οι συνιστώσες διασπώνται, έτσι ώστε οι νέες υποσυνιστώσες που προκύπτουν να μην είναι κοινές για όλες τις κατηγορίες. Θα αναφερόμαστε σε αυτή την προσέγγιση ως ιεραρχική εκπαίδευση του PRBF.

Ένα σημαντικό θέμα στην εκπαίδευση του PRBF είναι η αρχικοποίηση των συνιστωσών του, επειδή επηρεάζει το σημείο σύγκλισης του ΕΜ. Όπως ξέρουμε ο ΕΜ είναι ένας αλγόριθμος τοπικής αναζήτησης, οπότε είναι εγγυημένη η σύγκλισή του σε ένα τοπικό μέγιστο της πιθανοφάνειας το οποίο όμως μπορεί να βρίσκεται μακριά από το ολικό μέγιστο, δες για παράδειγμα την [9]. Η επίδραση της αρχικοποίησης στην επίδοση της μεθόδου φίνεται και στα πειραματικά αποτελέσματα που παραθέτουμε σε επόμενη ενότητα. Μια μερική λύση είναι η πολλαπλή επανεκκίνηση του ΕΜ με διαφορετικές αρχικοποιήσεις, και η διατήρηση της καλύτερης λύσης. Μια άλλη προσέγγιση είναι η εφαρμογή του αλγορίθμου k-means (που χρησιμοποιείται συχνά στην εκπαίδευση του δικτύου RBF) για την εύρεση αποδεκτών αρχικών τιμών. Ωστόσο το πρόβλημα μεταφέρεται έτσι στην αρχικοποίηση του k-means. Επιπλέον το κίνητρό μας είναι να αντιμετωπίσουμε το πρόβλημα της αρχικοποίησης με έναν τρόπο που θα προάγει την ταξινόμηση. Αυτό δεν είναι δυνατό με κανένα αλγόριθμο ομαδοποίησης, αφού δεν μπορεί να λάβει υπόψη του πληροφορία για τις κατηγορίες.

Για να αντιμετωπίσουμε το πρόβλημα της αρχικοποίησης προτείνουμε μια αυξητική μέθοδο εκπαίδευσης για το PRBF δίκτυο, όπου οι συνιστώσες προσθέτονται σειριακά σε επιλεγμένα σημεία του χώρου δεδομένων [22]. Η βασική ιδέα είναι η τοποθέτηση των συνιστωσών κοντά στο όριο απόφασης. Το όριο απόφασης ενός ταξινομητή διαιρεί των χώρο των δεδομένων σε μη επικαλυπτόμενες περιοχές, οι οποίες αντιστοιχούν σε διαφορετικές κατηγορίες. Αναμένουμε ότι μια καλή εκτίμηση των κατανομών κατηγοριών γύρω από το όριο απόφασης είναι αρκετή για να εξασφαλίσει καλή γενικευτική ικανότητα. Πρόσφατες μέθοδοι που επικεντρώνονται στο όριο απόφασης του RBF ταξινομητή περιγράφονται στις [54, 62]. Η μέθοδος που προτείνουμε είναι ντετερμινιστική, δεν εξαρτάται από την αρχικοποίηση του δικτύου, και μπορεί εύκολα να συνδυαστεί με κριτήρια επιλογής μοντέλου προκειμένου να βρεθεί και ο κατάλληλος αριθμός των συνιστωσών του δικτύου. Τα πειραματικά αποτελέ-

σματα δείχνουν ότι η μέθοδος είναι ανώτερη της ιεραρχικής εκπαίδευσης του PRBF [78], και έχει συγκρίσιμη επίδοση με μεθόδους βασισμένες σε Support Vector Machines (SVM) [25].

Πρέπει να σημειώσουμε ότι διάφορες μέθοδοι [65, 49, 84, 43, 44] έχουν προταθεί για την αυξητική εκπαίδευση του RBF δικτύου, και βασίζονται επίσης στην ιδέα της σταδιακής προσθήκης συνιστωσών κατά την εκπαίδευση. Ωστόσο αυτές οι μέθοδοι επικεντρώνονται σε σειριακά δεδομένα (on-line μάθηση) και σε προβλήματα παλινδρόμησης (προσέγγισης συναρτήσεων). Δεν είναι ξεκάθαρο το πώς μπορούν να προσαρμοστούν αυτές οι μέθοδοι στο πλαίσιο του PRBF, το οποίο είναι μια στατιστική μέθοδος στο πρόβλημα της ταξινόμησης, και έχει θεμελιώδεις διαφορές και σε σχέση με το μοντέλο που χρησιμοποιείται (το οποίο είναι μια μίξη κανονικών κατανομών) και σε σχέση με το πρόβλημα που επιλύει. Για αυτό η προσέγγιση που προτείνουμε είναι πολύ διαφορετική.

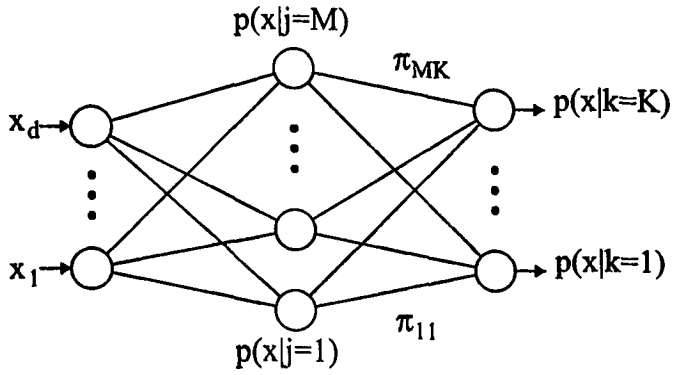
6.2 Το Πιθανοτικό Δίκτυο RBF για Ταξινόμηση

Θεωρούμε ένα πρόβλημα ταξινόμησης με K κατηγορίες, όπου K είναι γνωστό και κάθε πρότυπο ανήκει σε μια και μόνο κατηγορία. Μας δίνεται ένα σύνολο δεδομένων εκπαίδευσης $X = \{(x^{(n)}, y^{(n)}), n = 1, \ldots, N\}$ όπου $x^{(n)}$ είναι ένα d-διάστατο πρότυπο, και $y^{(n)}$ είναι μια ετικέτα $k \in \{1, \ldots, K\}$ που δείχνει την κατηγορία του προτύπου $x^{(n)}$. Το αρχικό σύνολο X μπορεί να χωριστεί χρησιμοποιώντας τις ετικέτες σε K ανεξάρτητα υποσύνολα X_k, έτσι ώστε κάθε υποσύνολο περιέχει μόνο πρότυπα της αντίστοιχης κατηγορίας. Έστω N_k ο αριθμός των προτύπων της κατηγορίας k, δηλαδή $N_k = |X_k|$.

Υποθέτουμε ότι έχουμε ένα δίκτυο με M συναρτήσεις βάσης οι οποίες είναι κατανομές, όπως στο Σχήμα 6.1. Στο δίκτυο PRBF όλες οι συνιστώσες του $f_j(x) = p(x|j)$ χρησιμοποιούνται για την εκτίμηση των υπό συνθήκη κατανομών όλων των κατηγοριών, θεωρώντας ότι όλες οι συνιστώσες βρίσκονται σε μια κοινή δεξαμενή (common-pool) [76, 77]. Οπότε κάθε κατανομή κατηγορίας $p(x|k)$ μοντελοποιείται με μια μίξη της μορφής:

$$p(x|k) = \sum_{j=1}^{M} \pi_{jk} f_j(x), \; k = 1, \ldots, K \tag{6.2}$$

όπου $f_j(x)$ είναι η συστατική κατανομή j, ενώ οι συντελεστές μίξης π_{jk} αποτελούν την



Σχήμα 6.1: Το πιθανοτικό δίκτυο RBF.

εκ των προτέρων πιθανότητα με την οποία ένα πρότυπο προέρχεται από την συνιστώσα j δοθέντος ότι ανήκει στην κατηγορία k. Οι συντελεστές είναι μη-αρνητικοί και αθροίζονται στην μονάδα $\sum_{j=1}^{M} \pi_{jk} = 1,\ k = 1, \ldots, K$. Μόλις υπολογιστούν οι έξοδοι $p(x|k)$ του δικτύου, η κατηγορία ενός άγνωστου προτύπου x καθορίζεται χρησιμοποιώντας τον κανόνα απόφασης του Bayes, δηλαδή το x ανατίθεται στην κατηγορία με την μέγιστη εκ των υστέρων πιθανότητα $P(k|x)$ που υπολογίζεται από την (6.1). Οι απαιτούμενες εκ των προτέρων πιθανότητες είναι οι $P_k = N_k/N$, σύμφωνα με την λύση μέγιστης πιθανοφάνειας.

Είναι επίσης χρήσιμο να ορίσουμε τις εκ των υστέρων πιθανότητες που εκφράζουν την πεποίθηση ότι η συνιστώσα j είναι υπεύθυνη για ένα πρότυπο x δοθέντος ότι ανήκει στην κατηγορία k. Αυτή η πιθανότητα υπολογίζεται με το θεώρημα του Bayes

$$P(j|x,k) = \frac{\pi_{jk} f_j(x)}{\sum_{i=1}^{M} \pi_{ik} f_i(x)}. \tag{6.3}$$

Στην συνέχεια της εργασίας θα αναφερόμαστε σε κανονικές συνιστώσες της μορφής:

$$f_j(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right\} \tag{6.4}$$

όπου $\mu_j \in \Re^d$ είναι το κέντρο της συνιστώσας j, ενώ Σ_j είναι ο αντίστοιχος $d \times d$ πίνακας συνδιακύμανσης. Το πλήρες διάνυσμα παραμέτρων Θ αποτελείται από τους συντελεστές μίξης, τα κέντρα και τους πίνακες συνδιακύμανσης $\Theta = \{\pi_{jk}, \mu_j, \Sigma_j | \ \forall j, k\}$. Όπως περιγράψαμε στην Ενότητα 2.3.2 το PRBF αποτελεί μια ειδική περίπτωση του δικτύου RBF, και γενικεύει την μίξη κανονικών κατανομών.

Για την εκτίμηση των παραμέτρων του PRBF, ο αλγόριθμος EM μπορεί να εφαρμοστεί

για την μεγιστοποίηση της πιθανοφάνειας:

$$\mathcal{L}(\Theta) = \sum_{k=1}^{K} \sum_{x \in X_k} \log p(x|k). \tag{6.5}$$

Όπως είναι γνωστό, ο ΕΜ είναι μια επαναληπτική διαδικασία με δύο βήματα σε κάθε επανάληψη. Κατά το Ε-βήμα υπολογίζονται οι εκ των υστέρων κατανομές (υπευθυνότητες) χρησιμοποιώντας τις τρέχουσες εκτιμήσεις για τους $\pi_{jk}^{(t)}$, $\mu_j^{(t)}$ και $\Sigma_j^{(t)}$, σύμφωνα με [77]:

$$P^{(t)}(j|x,k) = \frac{\pi_{jk}^{(t)} f_j(x; \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{i=1}^{M} \pi_{ik}^{(t)} f_i(x; \mu_i^{(t)}, \Sigma_i^{(t)})}. \tag{6.6}$$

Κατά το M-βήμα οι εκτιμήσεις των παραμέτρων ανανεώνονται σύμφωνα με τις εξισώσεις:

$$\mu_j^{(t+1)} = \frac{\sum_{k=1}^{K} \sum_{x \in X_k} P^{(t)}(j|x,k)x}{\sum_{\ell=1}^{K} \sum_{x \in X_\ell} P^{(t)}(j|x,\ell)} \tag{6.7}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{k=1}^{K} \sum_{x \in X_k} P^{(t)}(j|x,k)(x - \mu_j^{(t+1)})(x - \mu_j^{(t+1)})^T}{\sum_{\ell=1}^{K} \sum_{x \in X_\ell} P^{(t)}(j|x,\ell)} \tag{6.8}$$

$$\pi_{jk}^{(t+1)} = \frac{1}{N_k} \sum_{x \in X_k} P^{(t)}(j|x,k), \ k = 1, \ldots, K. \tag{6.9}$$

Ο ΕΜ τελικά συγκλίνει σε ένα τοπικό μέγιστο της πιθανοφάνειας.

6.3 Η Ιεραρχική Εκπαίδευση του Πιθανοτικού Δικτύου RBF

Στην [78] έχει προταθεί μια ιεραρχική μέθοδος εκπαίδευσης για το δίκτυο PRBF με δύο στάδια. Στο ΕΜ-στάδιο, ένα δίκτυο PRBF με M συνιστώσες εκπαιδεύεται με τον αλγόριθμο ΕΜ (6.6-6.9). Μετά την σύγκλιση του ΕΜ μπορεί να υπάρχουν συνιστώσες τοποθετημένες σε περιοχές που οι κατηγορίες των δεδομένων επικαλύπτονται. Αυτό μπορεί να συμβεί αν δεν έχουμε εκτιμήσει σωστά τον αριθμό M των συνιστωσών. Προκειμένου να αυξηθεί η γενικευτική επίδοση του δικτύου, στην [78] προτείνεται η διάσπαση τέτοιων συνιστωσών. Έτσι στο στάδιο διάσπασης της ιεραρχικής εκπαίδευσης, κάθε συνιστώσα του δικτύου διασπάται σε K το πολύ υπο-συνιστώσες, που αντιστοιχούν στις κατηγορίες των προτύπων για τα οποία είναι υπεύθυνη. Αυτό μπορούμε να το πετύχουμε υπολογίζοντας την εκ των υστέρων πιθανότητα $P(j|x,k)$ μέσω της (6.3) για κάθε συνιστώσα, και να αποφασίσουμε αν είναι

υπεύθυνη για πρότυπα πολλών κατηγοριών. Έτσι για κάθε πρότυπο $x \in X$ υπολογίζουμε την $P(j|x,k)$, και ελέγχουμε αν $\sum_{x \in X_k} P(j|x,k) > 0$ για περισσότερες από μια κατηγορίες k. Αν ισχύει αυτό, τότε αφαιρούμε την συνιστώσα j από το δίκτυο και προσθέτουμε μια ξεχωριστή υπό-συνιστώσα για κάθε κατηγορία. Αν διασπάσουμε μια συνιστώσα j, η υποσυνιστώσα που προκύπτει για την κατηγορία k είναι μια κανονική κατανομή $f_{jk} = p(x|j,k)$ με κέντρο μ_{jk}, πίνακα συνδιακύμανσηςΣ_{jk} και συντελεστή μίξης π_{jk}. Αυτές τις παραμέτρους θέτουμε σύμφωνα με τις παρακάτω εξισώσεις [78]:

$$\pi_{jk} = \frac{1}{N_k} \sum_{x \in X_k} P(j|x,k) \tag{6.10}$$

$$\pi_{j\ell} = 0, \ \forall \ell \neq k \tag{6.11}$$

$$\mu_{jk} = \frac{\sum_{x \in X_k} P(j|x,k)x}{\sum_{x \in X_k} P(j|x,k)} \tag{6.12}$$

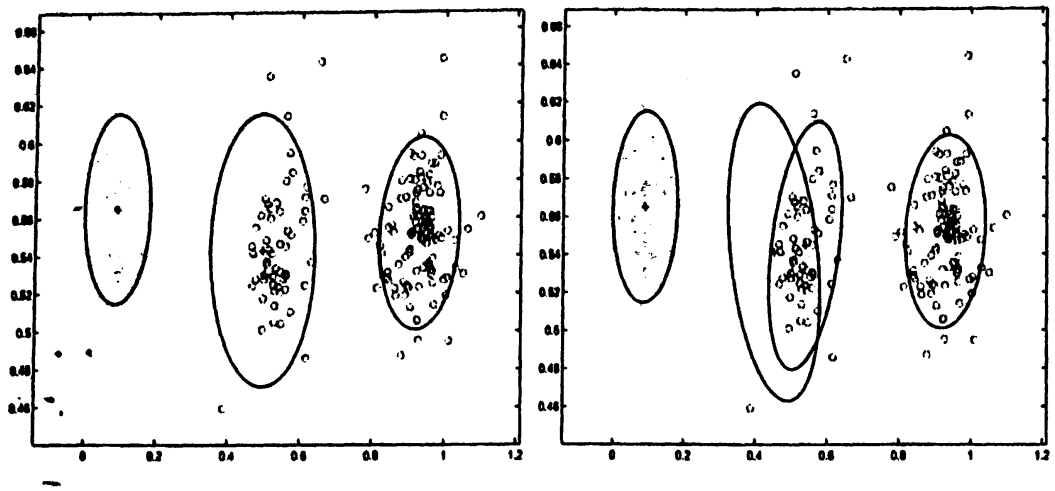
$$\Sigma_{jk} = \frac{\sum_{x \in X_k} P(j|x,k)(x-\mu_{jk})(x-\mu_{jk})^T}{\sum_{x \in X_k} P(j|x,k)}. \tag{6.13}$$

Μετά την διάσπαση για κάθε κατηγορία k υπάρχουν M_k συνιστώσες, και η έξοδος k του δικτύου είναι

$$p(x|k) = \sum_{j=1}^{M_k} \pi_{jk} f_{jk}(x), \ k = 1, \ldots, K. \tag{6.14}$$

Χρησιμοποιώντας τις προαναφερθείσες εξισώσεις, οι συνιστώσες που καλύπτουν περιοχές με πρότυπα πολλών κατηγοριών διασπώνται σε υπο-συνιστώσες που είναι υπεύθυνες για συγκεκριμένες κατηγορίες. Στο Σχήμα 6.2 παρουσιάζεται ένα χαρακτηριστικό παράδειγμα διάσπασης μιας συνιστώσας. Μια παρατήρηση που μπορεί να γίνει από την άποψη της ταξινόμησης είναι το ότι η πλήρης εκμετάλλευση της διάσπασης γίνεται όταν όλες οι συνιστώσες του PRBF είναι τοποθετημένες σε περιοχές που καλύπτουν το όριο απόφασης, όπως για παράδειγμα στο Σχήμα 6.3. Αυτή η παρατήρηση μας οδήγησε στην ανάπτυξη μιας αυξητικής μεθόδου για την τοποθέτηση των συνιστωσών του δικτύου που εκπαιδεύεται στο EM-στάδιο.

Έχει αποδειχθεί στην [78] ότι η προσθήκη του σταδίου διάσπασης από την μια εγγυάται την αύξηση της πιθανοφάνειας, και από την άλλη δίνει ένα δίκτυο με βελτιωμένη γενικευτική ικανότητα σε σχέση με το δίκτυο που εκπαιδεύεται στο EM-στάδιο. Ωστόσο αν η εφαρμογή του EM που προηγείται δεν έχει καταλήξει σε ικανοποιητική λύση, τότε η διάσπαση μπορεί να δώσε λύση που είναι πολύ κατώτερη από την βέλτιστη. Εφόσον ενδιαφερόμαστε για



Σχήμα 6.2: Η διαδικασία διάσπασης μιας συνιστώσας. Η κεντρική συνιστώσα είναι τοποθετημένη σε μια περιοχή με πρότυπα δύο κατηγοριών, και διασπάται σε δύο υπο-συνιστώσες σε δυο υπο-συνιστώσες που είναι υπεύθυνες για πρότυπα μιας κατηγορίας μόνο.

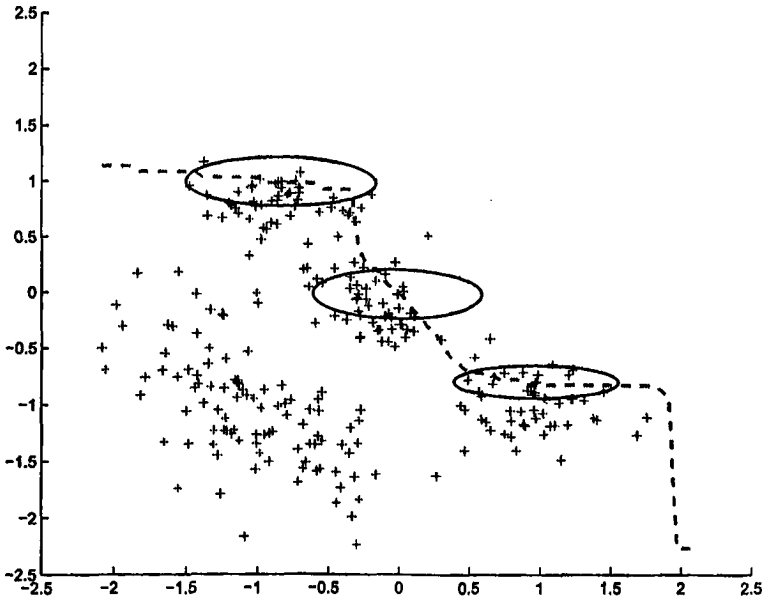
την ερμηνεία των συνιστωσών του PRBF σαν ομάδες προτύπων, θα θέλαμε να αποφύγουμε περιπτώσεις που ο ΕΜ έχει συγκλίνει σε μια λύση που μια συνιστώσα καλύπτει πολλές ομάδες ή μια ομάδα καλύπτεται από πολλές συνιστώσες. Ο τυπικός ΕΜ μπορεί να παγιδευτεί σε τέτοιες υποδεέστερες λύσεις, και αυτό εξαρτάται αποκλειστικά από την αρχικοποίησή του.

Για να αντιμετωπίσουμε αυτό το πρόβλημα, προτείνουμε μια αυξητική μέθοδο για το ΕΜ-στάδιο που ξεπερνά το πρόβλημα της αρχικοποίησης του ΕΜ. Κατά την διάρκεια αυτού του σταδίου, ξεκινώντας με μια συνιστώσα προσθέτουμε σταδιακά συνιστώσες στο δίκτυο, και στο δεύτερο στάδιο διασπάμε όλες τις συνιστώσες για να πάρουμε τον τελικό ταξινομητή.

6.4 Η Αυξητική Μέθοδος Εκπαίδευσης

Η προτεινόμενη αυξητική μέθοδος εφαρμόζεται στο πρώτο στάδιο της ιεραρχικής εκπαίδευσης. Όπως ήδη αναφέραμε, είναι λογικό να τοποθετήσουμε τις συνιστώσες σε περιοχές που περιέχουν πρότυπα πολλών κατηγοριών. Αυτή η τακτική αναμένουμε να δώσει μετά το στάδιο διάσπασης ένα δίκτυο με καλή επίδοση στην ταξινόμηση άγνωστων προτύπων.

Θεωρούμε ένα δίκτυο PRBF με M συνιστώσες κατά την διάρκεια του πρώτου σταδίου



Σχήμα 6.3: Η επιθυμητή τοποθέτηση των συνιστωσών πάνω στο όριο απόφασης (σχεδιασμένο με διακεκομμένη γραμμή) κατά το πρώτο στάδιο της ιεραρχικής εκπαίδευσης. Μια ακόλουθη διάσπαση των συνιστωσών θα δώσει μια ικανοποιητική λύση στο πρόβλημα της ταξινόμησης.

εκπαίδευσης. Προκειμένου να δημιουργήσουμε ένα δίκτυο με $M+1$ συνιστώσες η διαδικασία της προσθήκης εμπλέκει καθολική και τοπική αναζήτηση στων χώρο των παραμέτρων, για να ορίσει τις παραμέτρους της νέας συνιστώσας. Κατά την διάρκεια της καθολικής αναζήτησης, η μέθοδος εξετάζει ένα σύνολο από υποψήφιες περιοχές στον χώρο των προτύπων, και επιλέγει την καλύτερη με ένα κατάλληλο κριτήριο. Στην συνέχεια, για την τοπική αναζήτηση χρησιμοποιείται ο αλγόριθμος EM που προσαρμόζει τις παραμέτρους του νέου δικτύου με τις $M+1$ συνιστώσες. Η διαδικασία επαναληπτικής προσθήκης συνιστωσών ξεκινά με μια συνιστώσα και επαναλαμβάνεται μέχρι να ικανοποιηθεί κάποια συνθήκη τερματισμού.

6.4.1 Προσθήκη Συνιστώσας

Υποθέτοντας ένα δίκτυο με M συνιστώσες και διάνυσμα παραμέτρων Θ_M, η υπό συνθήκη κατανομή για κάθε κατηγορία k είναι $p(x|k;\Theta_M)$. Στην περίπτωση που προστίθεται μια νέα συνιστώσα $j = M+1$ με κατανομή $f_{M+1}(x)$, κάθε καινούρια κατανομή κατηγορίας $p(x|k;\Theta_{M+1})$ ορίζεται ως μια μίξη του τρέχοντος μοντέλου $p(x|k;\Theta_M)$ και της νέας

συνιστώσας $f_{M+1}(x)$:

$$p(x|k;\Theta_{M+1}) = (1-\alpha_k)p(x|k;\Theta_M) + \alpha_k f_{M+1}(x) \tag{6.15}$$

όπου α_k $(k = 1,\ldots,K)$ είναι οι συντελεστές μίξης της νέας συνιστώσας και $\alpha_k \in (0,1)$. Αυτή η προσέγγιση είναι ανάλογη με την αυξητικό αλγόριθμο Greedy-EM που έχει προταθεί στην [82] για την εκτίμηση κατανομών χωρίς επίβλεψη. Χρησιμοποιώντας την προηγούμενη εξίσωση, το δίκτυο που προκύπτει είναι πάλι PRBF. Η λογαριθμική πιθανοφάνεια $\mathcal{L}(\Theta_{M+1})$ του δικτύου με τις $M+1$ συνιστώσες είναι

$$\mathcal{L}(\Theta_{M+1}) = \sum_{k=1}^{K}\sum_{x\in X_k} \log\left\{(1-\alpha_k)p(x|k;\Theta_M) + \alpha_k f_{M+1}(x)\right\} . \tag{6.16}$$

Έστω PRBF(M) είναι το δίκτυο PRBF μετά την προσθήκη M συνιστωσών, και έστω PRBFsplit(M) το δίκτυο που προκύπτει μετά την διάσπαση των συνιστωσών του PRBF(M). Ο αυξητικός αλγόριθμος εκπαίδευσης που προτείνουμε συνοψίζεται στα ακόλουθα βήματα:

1. Θέτουμε $M := 1$. Αρχικοποιούμε το δίκτυο PRBF(1) με τη μια συνιστώσα, ως εξής:

$$\mu_1 = \frac{1}{|X|}\sum_{k=1}^{K}\sum_{x\in X_k} x \tag{6.17}$$

$$\Sigma_1 = \frac{1}{|X|}\sum_{k=1}^{K}\sum_{x\in X_k} (x-\mu_1)(x-\mu_1)^T \tag{6.18}$$

$$\pi_{1k} = 1,\ k = 1,\ldots,K. \tag{6.19}$$

2. Βρίσκουμε τις παραμέτρους της νέας συνιστώσας $f_{M+1}(x)$ και τους αντίστοιχους συντελεστές α_k, θεωρώντας τις κατανομές κατηγορίας $p(x|k;\Theta_M)$ σταθερές. Στην περίπτωση που δεν μπορεί να προστεθεί επιπλέον συνιστώσα τερματίζουμε την αυξητική διαδικασία και πηγαίνουμε στο βήμα 7.

3. Αρχικοποιούμε το δίκτυο με $M+1$ συνιστώσες μέσω της (6.15).

4. Εφαρμόζουμε τον αλγόριθμο EM στο δίκτυο μέχρι να συγκλίνει, για να πάρουμε το δίκτυο PRBF$(M+1)$.

5. Θέτουμε $M := M+1$.

6. Αν $M \leq M_{max}$, πηγαίνουμε στο βήμα 2.

7. Τπολογίζουμε το δίκτυο PRBFsplit(M) σύμφωνα με τις (6.10) ως (6.14).

Τα βήματα 1–6 αποτελούν το EM-στάδιο που κατασκευάζεται το PRBF, ενώ το βήμα 7 αντιστοιχεί στο Split-στάδιο κατασκευής του PRBFsplit. Είναι προφανές ότι η αυξητική διαδικασία τερματίζει είτε στο βήμα 2 (στην περίπτωση που δεν μπορούμε να εντοπίσουμε κατάλληλη θέση στα δεδομένα για την νέα συνιστώσα), είτε στο βήμα 6 αν έχει τοποθετηθεί ήδη ένας προκαθορισμένος μέγιστος αριθμός συνιστωσών.

6.4.2 Πού Τοποθετούμε τη Νέα Συνιστώσα;

Το βήμα 2 του αλγορίθμου είναι το πιο κρίσιμο κατά την προσθήκη της συνιστώσας, όπου καθορίζονται οι παράμετροι της νέας συνιστώσας με αναζήτηση μέσα σε ένα σύνολο υποψήφιων λύσεων. Μπορούμε να συνοψίζουμε αυτή την διαδικασία σε τρία βήματα:

2. (α΄) Δημιουργούμε ένα σύνολο από υποψήφιες συνιστώσες χρησιμοποιώντας μια τεχνική διαμέρισης των δεδομένων.

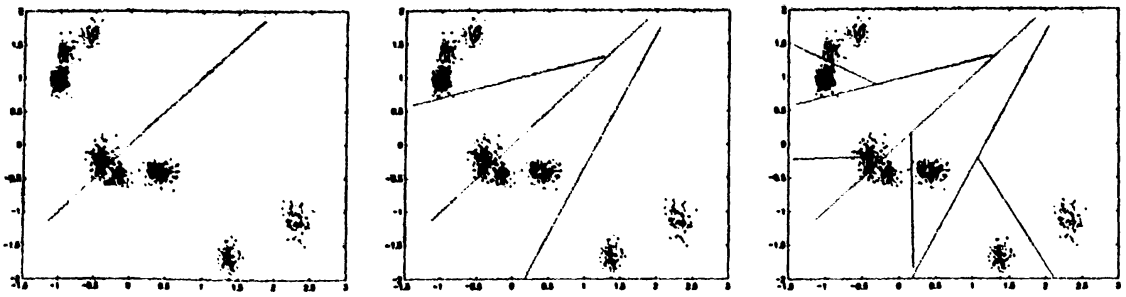
   (β΄) Προσαρμόζουμε τις παραμέτρους των υποψήφιων συνιστωσών.

   (γ΄) Χρησιμοποιούμε ένα κριτήριο επιλογής για να διαλέξουμε την συνιστώσα που θα προστεθεί στο δίκτυο.

Επειδή δεν είναι δυνατό να ορίσουμε απευθείας μια μοναδική καλή συνιστώσα για να προστεθεί, ορίζουμε ένα σύνολο από υποψήφιες αρχικές παραμετροποιήσεις, προσαρμόζουμε τις παραμέτρους χρησιμοποιώντας partial EM, και οι καλύτερες τιμές των παραμέτρων (μ, Σ, a_k) χρησιμοποιούνται στο βήμα 3 του αλγορίθμου.

Έστω M το τρέχον πλήθος συνιστωσών του δικτύου. Προκειμένου να δημιουργήσουμε το σύνολο των υποψήφιων αρχικών παραμετροποιήσεων διαμερίζουμε το σύνολο X σε M υποσύνολα βασισμένοι στις εκ των υστέρων πιθανότητες $P(j|x)$, έτσι για κάθε συνιστώσα j ορίζουμε το $S_j = \{x|P(j|x) > P(i|x), \forall i \neq j\}$. Οι πιθανότητες υπολογίζονται περιθωριοποιώντας τις ετικέτες των κατηγοριών:

$$P(j|x) = \sum_{k=1}^{K} P(j|x,k)P(k) \tag{6.20}$$



Σχήμα 6.4: Αναδρομική διαμέριση ενός τεχνητού συνόλου δεδομένων σε επικαλυπτόμενες περιοχές με χρήση του αλγορίθμου kd-tree. Όλες οι 14 διαμερίσεις που απεικονίζονται στα τρία γραφήματα λαμβάνονται υπόψη για τον προσδιορισμό των υποψήφιων παραμετροποιήσεων.

όπου $P(k) = N_k/N$ η εκ των προτέρων πιθανότητα της κατηγορίας k. Για κάθε ένα από τα M σύνολα S_j, δημιουργείται ένα υποσύνολο από υποψήφιες συνιστώσες, χωρίζοντας τα δεδομένα με τον αλγόριθμο kd-tree [8]. Ένα kd-tree ορίζει μια αναδρομική διαμέριση του χώρου δεδομένων σε μη-επικαλυπτόμενες περιοχές. Πρόκειται για ένα δυαδικό δέντρο, όπου τα δεδομένα που σχετίζονται με ένα μη-τερματικό κόμβο χωρίζονται στα δυο χρησιμοποιώντας ένα υπερεπίπεδο για να ορίσουν τους απογόνους του κόμβου. Για την διαμέριση των δεδομένων ενός κόμβου χρησιμοποιούμε την ίδια προσέγγιση με την [52], όπου το υπερεπίπεδο ορίζεται να είναι κάθετο στην κύρια συνιστώσα του πίνακα συνδιακύμανσης των δεδομένων του κόμβου και να διέρχεται από το μέσο τους. Το Σχήμα 6.4 απεικονίζει τα στάδια της διαμέρισης για ένα τεχνητό σύνολο δεδομένων. Η διαδικασία διαμερισμού εφαρμόζεται αναδρομικά μέχρι να δώσει ένα δέντρο τεσσάρων επιπέδων, και χρησιμοποιούμε όλους τους κόμβους του (όχι μόνο τους τερματικούς) για να ορίσουμε τα επικαλυπτόμενα υποσύνολα S_j (14 υποσύνολα για κάθε συνιστώσα j). Ο δειγματικός μέσος και ο δειγματικός πίνακας συνδιακύμανσης κάθε υποσυνόλου αποτελούν τις υποψήφιες αρχικές παραμέτρους της συνιστώσας $M + 1$. Οι αρχικές τιμές των α_k ορίζονται ίσες με $\pi_{jk}/2$, για τα υποσύνολα που προέκυψαν από την διαμέριση του S_j.

Προκειμένου να προσαρμόσουμε περισσότερο τις παραμέτρους κάθε υποψήφιας συνιστώσας, εφαρμόζουμε τον partial EM που ανανεώνει μόνο τις παραμέτρους της νέας συνιστώσας ενώ οι υπόλοιπες παράμετροι του δικτύου μένουν σταθερές. Αυτή η περιορισμένη τοπική

βελτιστοποίηση είναι γρήγορη, συνήθως μια ή δύο επαναλήψεις αρκούν, και η συνιστώσα παραμένει κοντά στην αρχική της θέση. Έστω Θ_M το διάνυσμα παραμέτρων του PRBF(M) που θεωρείται σταθερό κατά την εκτέλεση του partial EM. Στο E-βήμα του partial EM υπολογίζουμε τις εκ των υστέρων πιθανότητες $P^{(t)}(j = M+1|x,k)$ χρησιμοποιώντας τις τρέχουσες εκτιμήσεις των $\alpha_k^{(t)}$, $\mu_{M+1}^{(t)}$ και $\Sigma_{M+1}^{(t)}$ σύμφωνα με την:

$$P^{(t)}(j = M+1|x,k) = \frac{\alpha_k^{(t)} f_{M+1}(x; \mu_{M+1}^{(t)}, \Sigma_{M+1}^{(t)})}{(1-\alpha_k^{(t)}) p(x|k;\Theta_M) + \alpha_k^{(t)} f_{M+1}(x; \mu_{M+1}^{(t)}, \Sigma_{M+1}^{(t)})}. \tag{6.21}$$

Στο M-βήμα ανανεώνουμε τις παραμέτρους της συνιστώσας σύμφωνα με τις:

$$\mu_{M+1}^{(t+1)} = \frac{\sum_{k=1}^{K} \sum_{x \in X_k} P^{(t)}(j = M+1|x,k)\, x}{\sum_{\ell=1}^{K} \sum_{x \in X_\ell} P^{(t)}(j = M+1|x,\ell)} \tag{6.22}$$

$$\Sigma_{M+1}^{(t+1)} = \frac{\sum_{k=1}^{K} \sum_{x \in X_k} P^{(t)}(j = M+1|x,k)(x - \mu_{M+1}^{(t+1)})(x - \mu_{M+1}^{(t+1)})^T}{\sum_{\ell=1}^{K} \sum_{x \in X_\ell} P^{(t)}(j = M+1|x,\ell)} \tag{6.23}$$

$$\alpha_k^{(t+1)} = \frac{1}{|X_k|} \sum_{x \in X_k} P^{(t)}(j = M+1|x,k),\ k = 1, \ldots, K. \tag{6.24}$$

Αν εφαρμόσουμε τον partial EM για κάθε υποψήφια συνιστώσα, παίρνουμε το σύνολο των υποψήφιων αρχικών παραμετροποιήσεων $\theta^l = \{\mu^l, \Sigma^l, a_k^l\}$ της συνιστώσας $M+1$. Μια από αυτές επιλέγεται σύμφωνα με την διαδικασία που ακολουθεί.

Έχουμε ήδη αναφέρει ότι θέλουμε η νέα συνιστώσα να τοποθετηθεί σε μια περιοχή του χώρου δεδομένων που περιέχει πρότυπα πολλών κατηγοριών. Ένας τρόπος για να ποσοτικοποιήσουμε τον βαθμό στον οποίο μια υποψήφια συνιστώσα έχει αυτή την ιδιότητα είναι να υπολογίσουμε την μεταβολή της λογαριθμικής πιθανοφάνειας για την κατηγορία k, που οφείλεται στην προσθήκη της υποψήφιας νέας συνιστώσας l με κατανομή $f_l = p(x;\theta^l)$ σύμφωνα με την (6.15). Έτσι ορίζουμε την μεταβολή $\Delta\mathcal{L}_k^l$ για την κατηγορία k ως:

$$\begin{aligned} \Delta\mathcal{L}_k^l &= \frac{1}{N_k}(\mathcal{L}_k(\Theta_{M+1}^l) - \mathcal{L}_k(\Theta_M)) \\ &= \frac{1}{N_k} \sum_{x \in X_k} \log\left\{1 - \alpha_k + \alpha_k \frac{p(x;\theta^l)}{p(x|k;\Theta_M)}\right\} \end{aligned} \tag{6.25}$$

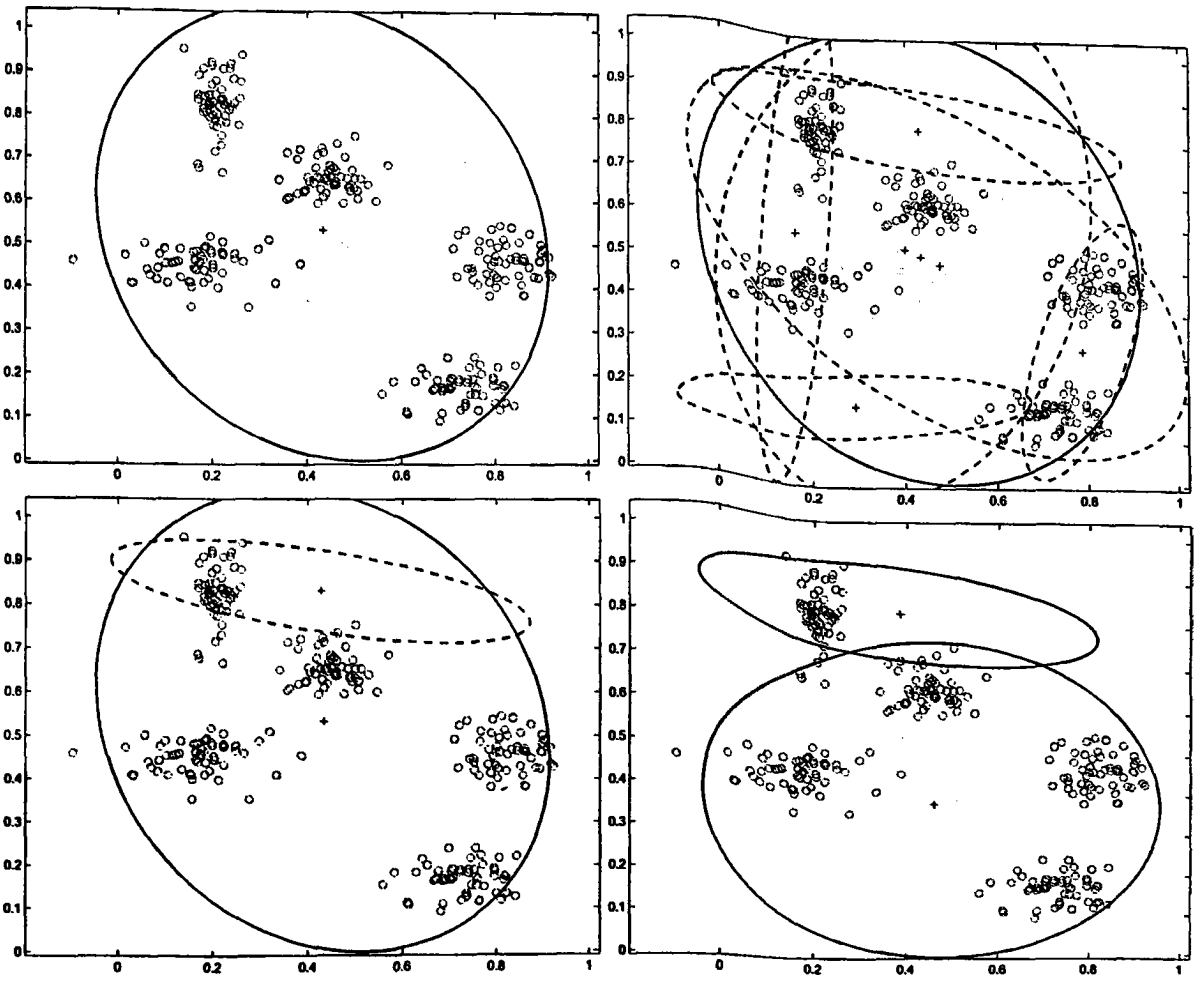
όπου $\Theta_{M+1}^l = \Theta_M \cup \theta^l$. Βασισμένοι στις τιμές $\Delta\mathcal{L}_k^l$, αναζητούμε ανάμεσα στις υποψήφιες συνιστώσες l, αυτές που όταν προστεθούν αυξάνουν την λογαριθμική πιθανοφάνεια σε δύο τουλάχιστον κατηγορίες. Τέτοιες υποψήφιες βρίσκονται σε περιοχές που περιέχουν πρότυπα

πολλών κατηγοριών. Προκειμένου να βρούμε την καλύτερη υποψήφια διατηρούμε τις συνιστώσες που αυξάνουν την λογαριθμική πιθανοφάνεια σε δύο τουλάχιστον κατηγορίες και απορρίπτουμε τις υπόλοιπες. Για κάθε συνιστώσα l που διατηρείται, προσθέτουμε τους θετικούς όρους $\Delta\mathcal{L}_k^l$ για να υπολογίσουμε την συνολική αύξηση της λογαριθμικής πιθανοφάνειας $\Delta\mathcal{L}_l$. Η υποψήφια l^* για την οποία μεγιστοποιείται η τιμή $\Delta\mathcal{L}_{l^*}$ προστίθεται στο τρέχον δίκτυο PRBF(M), εφόσον η τιμή αυτή είναι μεγαλύτερη από ένα κατώφλι (στα πειράματά μας το ορίσαμε ίσο με 0.01). Διαφορετικά θεωρούμε ότι η προσπάθεια να προσθέσουμε μια νέα συνιστώσα απέτυχε, και τερματίζουμε την EM-φάση του αλγορίθμου. Το κατώφλι που ορίσαμε αναφέρεται στην αύξηση της λογαριθμικής πιθανοφάνειας μετά από την προσθήκη μιας συνιστώσας. Τελικά, μετά από την προσθήκη πολλών συνιστωσών, η αύξηση τείνει στο μηδέν και η προσθήκη σταματά. Το κατώφλι που χρησιμοποιούμε το ορίσαμε εμπειρικά για να αποφασίσουμε πότε η αύξηση της πιθανοφάνειας είναι αμελητέα, και να αποφύγουμε την προσθήκη περιττών συνιστωσών. Μετά από πειραματισμό με πολλά σύνολα δεδομένων φτάσαμε στο συμπέρασμα ότι αλλαγές αυτής της κλίμακας στην πιθανοφάνεια δεν επηρεάζουν την επίδοση στην ταξινόμηση, και επίσης δεν οδηγούν σε πρόωρο τερματισμό της αύξησης του δικτύου.

Στο Σχήμα 6.5 απεικονίζεται η διαδικασία της προσθήκης συνιστώσας. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω φαίνεται το δίκτυο με μια συνιστώσα (βήμα 1), η μοναδική συνιστώσα και έξι υποψήφιες νέες συνιστώσες (βήμα 2(β)), η επιλεγμένη υποψήφια συνιστώσα (βήμα 2(γ)), και το δίκτυο με δύο συνιστώσες μετά από εφαρμογή του EM (βήμα 4).

Πρέπει να σημειωθεί ότι χάρη στην εφαρμογή του EM είναι δυνατό μια συνιστώσα που προστέθηκε να μετακινηθεί μακρυά από το όριο απόφασης όπου είχε αρχικά τοποθετηθεί. Αυτό μπορεί να συμβεί στην περίπτωση που υπάρχει μια περιοχή στον χώρο δεδομένων που περιέχει πρότυπα μιας κατηγορίας και δεν καλύπτεται επαρκώς από τις υπάρχουσες συνιστώσες. Ωστόσο αυτό δεν αποτελεί πρόβλημα για την μέθοδο, αφού η επόμενη συνιστώσα που θα προστεθεί είναι πολύ πιθανό ότι θα τοποθετηθεί στην ίδια αρχική θέση με την προηγούμενη, και θα παραμείνει εκεί.

Ολοκληρώνοντας, ένα πολύ ελκυστικό χαρακτηριστικό της προτεινόμενης μεθόδου είναι το ότι για την εκπαίδευση ενός δικτύου PRBF(M) με M συνιστώσες η μέθοδος κατασκευά-



Σχήμα 6.5: Η αυξητική διαδικασία προσθήκης των δύο πρώτων συνιστωσών του δικτύου PRBF. Οι υπάρχουσες συνιστώσες του δικτύου έχουν σχεδιαστεί με συνεχείς γραμμές και οι υποψήφιες συνιστώσες με διακεκομμένες.

Πίνακας 6.1: Τα χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιήθηκαν στα πειράματα.

	BLD	*PID*	*Iris*	*Veh*	*Glass*	*Wave*	*Wine*	*Thyr*
πρότυπα	345	768	150	846	214	5000	178	215
χαρακτηριστικά	6	8	4	18	9	21	13	5
κατηγορίες	2	2	3	4	6	3	3	3

ζει όλα τα ενδιάμεσα δίκτυα PRBF(m) με $m = 1, \ldots, M$ συνιστώσες. Έτσι οποιαδήποτε μέθοδος επιλογής μοντέλου μπορεί να υλοποιηθεί πολύ αποδοτικά, σε σύγκριση με την συμβατική μέθοδο εκπαίδευσης όπου ο αλγόριθμος EM πρέπει να εκτελεστεί επαναληπτικά για κάθε $m = 1, \ldots, M$. Ανάλογο μειονέκτημα έχουν και οι συμβατικοί μέθοδοι εκπαίδευσης του SVM, όπου χρειάζονται πολλές εκτελέσεις με διάφορες τιμές των υπερ-παραμέτρων (π.χ. για την ακτίνα ενός RBF πυρήνα) προκειμένου να καθοριστούν οι βέλτιστες τιμές των υπερ-παραμέτρων.

6.5 Πειράματα

Συγκρίναμε την προτεινόμενη αυξητική μέθοδο εκπαίδευσης (την καλούμε incremental PRBFsplit) με την ιεραρχική μέθοδο που προτάθηκε στην [78] (την καλούμε PRBFsplit) και με τον SVM ταξινομητή. Για την εκπαίδευση του SVM χρησιμοποιήσαμε την βιβλιοθήκη λογισμικού OSU SVM Classifier Matlab Toolbox version 3.0[1]. Εξετάσαμε οκτώ σύνολα δεδομένων από το UCI repository [11], συγκεκριμένα τα Bupa Liver Disorder (BLD), Pima Indian Diabetes (PID), Iris, Vehicles (Veh), Glass, Waveform (Wave), Wine και Thyroid (Thyr). Ο αριθμός των προτύπων, ο αριθμός των χαρακτηριστικών και ο αριθμός των κατηγοριών κάθε συνόλου συνοψίζονται στον Πίνακα 6.1.

Για κάθε σύνολο δεδομένων υπολογίσαμε το αναμενόμενο σφάλμα γενίκευσης με δια-

[1]Το λογισμικό είναι διαθέσιμο στο `http://www.ece.osu.edu/~maj/osu_svm`

σταυρωμένη επικύρωση, δηλαδή χωρίσαμε το σύνολο δεδομένων σε δέκα υποσύνολα και εκτελέσαμε το πείραμα δέκα φορές, χρησιμοποιώντας κάθε φορά ένα υποσύνολο για έλεγχο και τα υπόλοιπα εννιά για εκπαίδευση του δικτύου. Σε κάθε εκτέλεση του πειράματος εκπαιδεύσαμε εννιά δίκτυα για διαφορετικές τιμές του M και διαφορετικούς τύπους του πίνακα συνδιακύμανσης, χρησιμοποιώντας ένα από τα εννιά διαθέσιμα υποσύνολα για επικύρωση και τα υπόλοιπα για την αυξητική μέθοδο εκπαίδευσης. Επιλέξαμε τον συνδυασμό του M και του τύπου του πίνακα συνδιακύμανσης που είχαν το μικρότερο αναμενόμενο σφάλμα επικύρωσης, και τον ελέγξαμε με το υποσύνολο ελέγχου. Ακριβώς η ίδια διαδικασία επικύρωσης και τα ίδια υποσύνολα χρησιμοποιήθηκαν για την προσαρμογή των υπερ-παραμέτρων του SVM, και για την επιλογή μοντέλου κατά την ιεραρχική εκπαίδευση του PRBF. Το μοντέλο SVM με το οποίο πειραματιστήκαμε είχε RBF πυρήνες $K(x,y) = \exp\{-\gamma|x-y|^2\}$. Για προβλήματα με περισσότερες από δύο κατηγορίες χρησιμοποιήσαμε το σχήμα "1 προς 1" για να τα ανάγουμε σε προβλήματα δύο κατηγοριών, στην [1] γίνεται μια ανασκόπηση των σχετικών μεθόδων. Οι υπερ-παράμετροι που έπρεπε να ρυθμίσουμε ήταν το αντίστροφο εύρος γ του πυρήνα και το κόστος C παραβίασης περιορισμών. Για κάθε πείραμα, προκειμένου να καθορίσουμε τις τιμές των (C,γ) αρχικοποιήσαμε το $C=1$ και αναζητήσαμε μέσα στο σύνολο $\{2^s | s = -50,\ldots,50\}$ την τιμή του γ που ελαχιστοποιεί το σφάλμα επικύρωσης. Θέτοντας το γ στη βέλτιστη λύση, αναζητήσαμε στη συνέχεια την βέλτιστη τιμή του C μέσα στο σύνολο $\{2^s | s = -50,\ldots,50\}$. Προφανώς η μέθοδος αυτή δεν εξετάζει όλους τους πιθανούς συνδυασμούς των (C,γ) αλλά την επιλέξαμε για να έχουμε αποδεκτούς χρόνους εκτέλεσης. Πειραματιστήκαμε επίσης με την εξαντλητική αναζήτηση σε ένα πλέγμα του χώρου που ορίζουν τα (C,γ) και η επίδοση στην ταξινόμηση ήταν σχεδόν η ίδια, ενώ ο χρόνος εκτέλεσης ήταν 10–70 φορές μεγαλύτερος.

Ένα σημαντικό πρόβλημα που επηρεάζει την επίδοση του ταξινομητή PRBF είναι ο καθορισμός του τύπου του πίνακα συνδιακύμανσης. Χρησιμοποιήσαμε τρεις παραμετροποιήσεις του: συμμετρικός θετικά ορισμένος (τον καλούμε "πλήρη"), θετικός διαγώνιος (τον καλούμε "διαγώνιο") και θετικός διαγώνιος πίνακας με όλα τα στοιχεία της διαγωνίου ίδια (τον καλούμε "σφαιρικό"). Ο τύπος του πίνακα καθορίζει των αριθμό των παραμέτρων του δικτύου και το σχήμα των συνιστωσών για αυτό πρέπει να γίνει ένας συμβιβασμός. Ένας πλήρης πίνακας επιτρέπει συνιστώσες οποιασδήποτε μορφής, αλλά μπορεί να αυξήσει αδικαιολόγητα

την πολυπλοκότητα του δικτύου και να οδηγήσει σε υποδεέστερες λύσεις. Επιπλέον για μικρά σύνολα δεδομένων με πολλά χαρακτηριστικά, η χρήση του πλήρους πίνακα συνδιακύμανσης μπορεί να οδηγήσει σε συνιστώσες με πίνακες που έχουν μηδενικές ιδιοτιμές. Για να το αντιμετωπίσουμε, για κάθε επανάληψη της διασταυρωμένης επικύρωσης εκτελέσαμε τρεις φορές τον αλγόριθμο εκπαίδευσης χρησιμοποιώντας κάθε φορά διαφορετικού τύπου πίνακα συνδιακύμανσης. Με αυτό τον τρόπο κατασκευάσαμε κάθε φορά τρία δίκτυα, τα $PRBFsplit^{full}(m)$, $PRBFsplit^{diag}(m)$ και $PRBFsplit^{spherical}(m)$, και επιλέξαμε αυτό με την καλύτερη επίδοση στο σύνολο επικύρωσης. Είναι προφανές ότι για διαφορετικά υποσύνολα εκπαίδευσης τα δίκτυα που επιλέγουμε μπορεί να διαφέρουν ως προς τον αριθμό των συνιστωσών και τον τύπο του πίνακα συνδιακύμανσης. Όσο αφορά την ιεραρχική εκπαίδευση του PRBF, για κάθε τιμή του $m = 1, \ldots, M_{max}$ και κάθε τύπο πίνακα, εκτελέσαμε τον EM πέντε φορές με διαφορετικές αρχικοποιήσεις και επιλέξαμε το δίκτυο με το μικρότερο σφάλμα επικύρωσης για να προχωρήσουμε στη διάσπαση των συνιστωσών του. Τέλος πρέπει να αναφέρουμε ότι κατά την ιεραρχική και την αυξητική εκπαίδευση του PRBF ο μέγιστος αριθμός συνιστωσών ήταν $M_{max} = 30$ για όλα τα πειράματα. Αν και στις περισσότερες περιπτώσεις ο αυξητικός αλγόριθμος σταματούσε νωρίτερα, καθώς δεν μπορούσε να προσθέσει επιπλέον σημαντικές συνιστώσες.

Ο Πίνακας 6.2 συγκεντρώνει για όλα τα πειράματα την μέση τιμή και την τυπική απόκλιση του σφάλματος γενίκευσης, δηλαδή του ποσοστού των προτύπων σε κάθε υποσύνολο ελέγχου που ταξινομήθηκαν λάθος. Σημειώνουμε ότι έγινε η ίδια διαμέριση των δεδομένων σε δέκα υποσύνολα και για τις τρεις μεθόδους εκπαίδευσης. Είναι φανερό ότι η αυξητική μέθοδος εκπαίδευσης υπερτερεί της ιεραρχικής. Σε σύγκριση με το SVM, οι επιδόσεις τους είναι συγκρίσιμες. Για κάποια σύνολα δεδομένων η προτεινόμενη μέθοδος υπερτερεί, ενώ για κάποια άλλα το SVM δίνει καλύτερα αποτελέσματα. Έτσι δεν είναι δυνατό να βγάλουμε αξιόπιστα συμπεράσματα για την υπεροχή της μιας ή της άλλης μεθόδου όσον αφορά το σφάλμα γενίκευσης.

Ο Πίνακας 6.3 συγκεντρώνει πειραματικές μετρήσεις για την σύγκριση της προτεινόμενης μεθόδου με το SVM. Οι μετρήσεις αφορούν τον χρόνο εκτέλεσης και των αριθμών των συνιστωσών ή των διανυσμάτων που χρησιμοποιήθηκαν αντίστοιχα. Όσον αφορά τις συνιστώσες, η μέθοδός μας χρησιμοποιεί σημαντικά λιγότερες από τα διανύσματα στήρι-

Πίνακας 6.2: Η μέση τιμή (%) και η τυπική απόκλιση (σε παρένθεση) του σφάλματος γενίκευσης των τριών μεθόδων.

	PRBFsplit	*Incremental PRBFsplit*	*SVM*
BLD	31.3 (7.8)	28.4 (5.2)	30.5 (5.3)
PID	25.3 (4.6)	24.1 (5.8)	22.7 (4.6)
Iris	3.7 (5.2)	2.0 (3.2)	3.3 (4.7)
Veh	20.4 (3.6)	14.3 (5.17)	16.5 (5.5)
Glass	34.4 (10.3)	28.6 (8.2)	29.1 (9.5)
Wave	14.4 (1.5)	14.2 (1.9)	13.3 (1.2)
Wine	2.2 (2.8)	0.5 (1.75)	1.1 (3.5)
Thyr	7.0 (4.5)	4.6 (6.1)	5.5 (5.2)

ξης (support vectors) του SVM, αυτό οφείλεται στο ότι το μοντέλο που χρησιμοποιούμε βασίζεται στην ομαδοποίηση των προτύπων, και αποτελείται από ένα σχετικά μικρό αριθμό από μεγάλες περιοχές του χώρου προτύπων, που καθορίζονται από την περιοχή επιρροής κάθε συνιστώσας. Όσον αφορά τον χρόνο εκτέλεσης, τα αποτελέσματα εξαρτώνται από το πόσοι συνδυασμοί των υπερ-παραμέτρων (C, γ) του SVM δοκιμάζονται μέχρι να βρεθεί ο βέλτιστος. Οι χρόνοι είναι συγκρίσιμοι για τις δύο μεθόδους, όμως πιο διεξοδικοί τρόποι αναζήτησης των παραμέτρων μπορεί να αυξήσουν υπερβολικά τον χρόνο εκτέλεσης. Οι χρόνοι εκτέλεσης μετρήθηκαν για υλοποιήσεις των μεθόδων σε Matlab και την εκτέλεσή τους στον ίδιο προσωπικό υπολογιστή. Μια ποιοτική σύγκριση των μεθόδων γίνεται στην επόμενη ενότητα.

Πίνακας 6.3: Ο μέσος χρόνος εκτέλεσης (σε δευτερόλεπτα) και ο αριθμός των συνιστωσών/διανυσμάτων για την προτεινόμενη μέθοδο και το SVM.

	Incremental PRBFsplit		*SVM*	
	χρόνος	*συνιστώσες*	*χρόνος*	*διανύσματα*
BLD	33.8	4.5	39.7	224.0
PID	61.4	4.3	164.1	413.2
Iris	16.2	3.0	19.0	50.7
Veh	87.5	4.0	617.0	438.2
Glass	60.6	16.9	20.1	161.4
Wave	118.5	4.6	11357.1	2381.8
Wine	28.1	3.6	30.2	88.4
Thyr	45.3	3.8	19.1	69.4

6.6 Συμπεράσματα

Σε αυτή το κεφάλαιο παρουσιάστηκε μια αυξητική μέθοδο για την εκπαίδευση του δικτύου PRBF [22], που ξεπερνά το πρόβλημα της αρχικοποίησης του τυπικού αλγόριθμου EM και δίνει δίκτυα με ανώτερη γενικευτική ικανότητα. Ο προτεινόμενος αλγόριθμος βασίζεται στην προσεκτική τοποθέτηση νέων συνιστωσών σε περιοχές που είναι ενδιαφέρουσες από άποψη ταξινόμησης, δηλαδή σε περιοχές στις οποίες βρίσκονται πρότυπα πολλών κατηγοριών (το όριο απόφασης). Αυτές οι συνιστώσες στην συνέχεια διασπώνται για να προκύψουν υπο-συνιστώσες που αντιστοιχούν σε συγκεκριμένες κατηγορίες, επιτυγχάνοντας έτσι μια βελτιωμένη εκτίμηση των κατανομών των προτύπων κάθε κατηγορίας στην περιοχή ενδιαφέροντος. Τα πειραματικά αποτελέσματα δείχνουν ότι πρόκειται για μια μέθοδο ανταγωνιστική των SVM, που αξίζει να λαμβάνεται υπόψη κατά την κατασκευή ενός ταξινομητή.

Η προτεινόμενη αυξητική μέθοδος παρουσιάζει πολλές σημαντικές διαφορές με το SVM. Πρώτα από όλα, είναι μια στατιστική προσέγγιση που ακολουθεί το generative πρότυπο, σε αντίθεση με το discriminative πρότυπο που ακολουθούν το SVM και τα τυπικά νευρωνικά

μοντέλα (MLP και RBF). Επιπλέον μπορούμε να επισημάνουμε τις παρακάτω διαφορές:

- Το PRBF είναι εγγενώς κατάλληλο για προβλήματα πολλών κατηγοριών, ενώ το τυπικό SVM επιλύει προβλήματα δύο κατηγοριών και απαιτείται περαιτέρω ανάπτυξη για να επιλύσει προβλήματα περισσοτέρων κατηγοριών.

- Το PRBF ταξινομεί ένα πρότυπο x βασισμένο στις εκ των υστέρων πιθανότητες των κατηγοριών $P(k|x)$, έτσι μπορεί άμεσα να υπολογιστεί η αβεβαιότητα της απόφασης ή να χρησιμοποιηθούν οι πιθανότητες αυτές σε έναν πιθανοτικό κανόνα απόφασης.

- Το PRBF μπορεί απευθείας να δώσει μια κατάταξη των κατηγοριών βασισμένη στις αντίστοιχες εκ των υστέρων πιθανότητες.

- Το PRBF στηρίζει την απόφασή του σε μια εκτίμηση της κατανομής κάθε κατηγορίας στην περιοχή του προτύπου εισόδου, έτσι η κατηγοριοποίηση μπορεί εύκολα να ερμηνευθεί διαισθητικά. Από την άλλη, το SVM κατασκευάζει τα όρια απόφασης στον χώρο των πυρήνων (kernel space), έτσι η απόφαση του είναι πιο δύσκολο να ερμηνευθεί.

- Το PRBF περιγράφεται από ένα μικρό σύνολο περιοχών, που καθορίζονται από τις κανονικές συνιστώσες. Από την άλλη το SVM περιγράφεται από ένα πολύ μεγαλύτερο σύνολο διανυσμάτων στήριξης..

- Στην ιεραρχική εκπαίδευση του PRBF είναι πιο εύκολη και γρήγορη η επιλογή μοντέλου, αφού όλα τα μοντέλα δημιουργούνται διαδοχικά σε μια εκτέλεση του αλγορίθμου.

- Το PRBF είναι ένα μοντέλο που βασίζεται σε κατανομές, οπότε αναμένουμε να έχει μειωμένη απόδοση σε προβλήματα με πολύ μικρό αριθμό προτύπων και πολύ μεγάλο αριθμό χαρακτηριστικών, όπως συμβαίνει για παράδειγμα σε πολλά προβλήματα ταξινόμησης στην βιοπληροφορική (π.χ. ταξινόμηση της έκφρασης των γονιδίων).

- Το SVM δεν εμφανίζει αριθμητικές δυσκολίες παρόμοιες με αυτές του PRBF (π.χ. πίνακες διακύμανσης με μηδενικές ιδιοτιμές).

Υπάρχουν περιθώρια για περαιτέρω έρευνα, ιδιαίτερα στο θέμα του καθορισμού του συνόλου των υποψήφιων για προσθήκη συνιστωσών σε κάθε επανάληψη. Επίσης αξίζει να

εξεταστεί η χρήση Probabilistic Principal Component Analyzers [75] αντί για κανονικές συνιστώσες.

Κεφαλαιο 7

Ενεργητικη Μαθηση με το Νευρωνικο Δικτυο PRBF

Σε αυτό το κεφάλαιο παρουσιάζουμε μια μέθοδο ενεργητικής μάθησης για την εκπαίδευση του πιθανοτικού δικτύου RBF (PRBF) [21]. Όπως ήδη αναφέραμε στο Κεφάλαιο 2, το δίκτυο αυτό είναι μια ειδική περίπτωση του RBF, και αποτελεί μια γενίκευση της μίξης κανονικών κατανομών. Αρχικά προτείνουμε μια αυξητική μέθοδο για την εκπαίδευση του δικτύου με ημι-επίβλεψη που βασίζεται στον αλγόριθμο EM. Στην συνέχεια προτείνουμε μια μέθοδο ενεργητικής μάθησης που επαναληπτικά εφαρμόζει την διαδικασία μάθησης με ημι-επίβλεψη στα ταξινομημένα και τα αταξινόμητα δεδομένα, και στην συνέχεια επιλέγει ένα αταξινόμητο δεδομένο και ζητά να μάθει την κατηγορία του. Το κριτήριο επιλογής που προτείνουμε επιλέγει σημεία κοντά στο όριο απόφασης του τρέχοντος ταξινομητή, και διευκολύνει την αυξητική μέθοδο με ημι-επίβλεψη που επίσης εκμεταλλεύεται το όριο απόφασης. Τα αποτελέσματα της μεθόδου σε πειράματα με γνωστά σύνολα δεδομένων είναι ενθαρρυντικά.

7.1 Εισαγωγή

Η ενεργητική μάθηση ενός ταξινομητή αποτελεί ένα ιδιαίτερο πρόβλημα της μηχανικής μάθησης, όπου τα δεδομένα της εκπαίδευσης συλλέγονται ενεργά κατά την διάρκεια της εκπαίδευσης. Τα δεδομένα είναι διαθέσιμα σαν μια ροή ταξινομημένων προτύπων, αλλά η πληροφορία που μεταφέρουν καθορίζεται από τον ταξινομητή. Ο εκπαιδευόμενος ταξινομητής καθορίζει σε κάθε επανάληψη περιοχές ενδιαφέροντος στο χώρο δεδομένων, και ζητά δεδομένα εκπαίδευσης από αυτές τις περιοχές. Η σημασία την ενεργητικής μάθησης είναι ευρέως αποδεκτή, και στην [15] μελετάτε η βελτίωση της γενικευτικής ικανότητας του ταξινομητή καθώς ο αριθμός των ταξινομημένων προτύπων αυξάνεται. Πολλές σχετικές μέθοδοι έχουν προταθεί για το πρόβλημα. Στην [17] έχει προταθεί ένας αλγόριθμος για μια μίξη κανονικών κατανομών, που επιλέγει τα δεδομένα που ελαχιστοποιούν την διακύμανση του ταξινομητή. Στην [30] έχει προταθεί η ενεργητική μάθηση μιας επιτροπής από ταξινομητές, που επιλέγει δεδομένα για τα οποία τα μέλη της επιτροπής διαφωνούν. Με βάση την ίδια μέθοδο επιλογής, στην [56] έχει προταθεί η αξιοποίηση των διαθέσιμων αταξινόμητων προτύπων χρησιμοποιώντας τον EM για να βελτιωθεί η διαδικασία επιλογής, με στόχο την εκπαίδευση ενός απλοϊκού Bayesian (naive Bayes) ταξινομητή . Στην [86] εκπαιδεύουν Gaussian random fields και αρμονικές συναρτήσεις, και επιλέγουν δεδομένα βασισμένοι στο εκτιμώμενο αναμενόμενο σφάλμα ταξινόμησης.

Η εργασία μας επικεντρώνεται σε μια παραλλαγή του σεναρίου της ενεργητικής μάθησης που ονομάζεται ενεργητική μάθηση βασισμένη σε μια "δεξαμενή προτύπων" (pool-based), και έχει επίσης μελετηθεί στις [56, 86]. Σε αυτή την περίπτωση το σύνολο των ταξινομημένων και αταξινόμητων προτύπων είναι διαθέσιμο από την αρχή. Κατά την διάρκεια της εκπαίδευσης μας επιτρέπεται να ρωτάμε επαναληπτικά την κατηγορία ενός αταξινόμητου προτύπου, και να χρησιμοποιούμε την απάντηση για να βελτιώσουμε τον ταξινομητή. Στην πράξη αυτό το σενάριο είναι σημαντικό όταν το να ρωτήσεις έναν ειδικό του πεδίου κοστίζει ακριβά, όπως η διάγνωση ενός γιατρού, ή όταν ο όγκος των αταξινόμητων δεδομένων είναι τόσο μεγάλος που εμποδίζει την διεξοδική ταξινόμησή τους, όπως συμβαίνει συχνά στην ταξινόμηση κειμένων. Διαισθητικά το κίνητρο πίσω από την ενεργητική μάθηση που βασίζεται σε μια δεξαμενή προτύπων είναι το ότι τα αταξινόμητα πρότυπα μπορούν να βοηθήσουν στην κατασκευή ενός πιο λεπτομερούς στατιστικού μοντέλου παραγωγής των προτύπων. Έτσι

το πρόβλημα συνδέεται στενά με το πρόβλημα της μάθησης με ημι-επίβλεψη. Αλγόριθμοι για την μάθηση με ημι-επίβλεψη έχουν προταθεί για την μίξη κανονικών κατανομών [36, 74], καθώς και για το δίκτυο RBF [59]. Έχει λοιπόν διαπιστωθεί ότι τα αταξινόμητα πρότυπα αποκαλύπτουν χρήσιμη πληροφορία για την κατανομή των ταξινομημένων.

Στην συνέχεια θα μελετήσουμε την ενεργητική μάθηση που βασίζεται σε μια δεξαμενή προτύπων για το πιθανοτικό δίκτυο RBF. Όπως αναφέραμε είναι μια ειδική περίπτωση του δικτύου RBF, που υπολογίζει σε κάθε του έξοδο την κατανομή των προτύπων μιας κατηγορίας. Κατά την εκπαίδευσή του υιοθετούμε την ερμηνεία των συνιστωσών (συναρτήσεων βάσης) σαν ομάδες προτύπων, όπου κάθε ομάδα μπορεί να περιλαμβάνει πρότυπα κάθε κατηγορίας. Στο Κεφάλαιο 6 προτάθηκε μια αυξητική μέθοδος που βασίζεται στον EM για την εκπαίδευση του δικτύου PRBF με επίβλεψη. Στην συνέχεια θα την επεκτείνουμε για να καταλήξουμε σε μια αυξητική μέθοδο που βασίζεται στον EM για την εκπαίδευση του δικτύου PRBF με ημι-επίβλεψη. Σε αυτό μας διευκολύνει το γεγονός ότι κάθε συνιστώσα του PRBF περιγράφει την τοπική κατανομή των προτύπων, τα οποία μπορούν να είναι διαφόρων κατηγοριών. Έτσι για τα αταξινόμητα πρότυπα μπορούμε να περιθωριοποιήσουμε την μεταβλητή της κατηγορίας από τις εξισώσεις ανανέωσης του EM, και να χρησιμοποιήσουμε μαζί τα ταξινομημένα και τα αταξινόμητα πρότυπα για την εκτίμηση των παραμέτρων.

7.2 Μάθηση με Ημι-Επίβλεψη

Υποθέτουμε ένα σύνολο από ταξινομημένα πρότυπα $X = \{(x^n, y^n) | \, n = 1, \ldots, N\}$ και ένα σύνολο από αταξινόμητα πρότυπα $X_\emptyset = \{x^n | \, n = 1, \ldots, N_\emptyset\}$. Τα ταξινομημένα πρότυπα αποτελούνται από ένα τμήμα "εισόδου" $x \in \Re^d$ και ένα τμήμα "εξόδου" $y \in \{1, \ldots, K\}$, στην περίπτωση προβλημάτων με K κατηγορίες. Αυτό το τμήμα "εξόδου" αναθέτει μια κατηγορία στο πρότυπο, και στην περίπτωση των αταξινόμητων προτύπων λείπει. Έστω Ω η ένωση των συνόλων των ταξινομημένων και αταξινόμητων προτύπων, δηλαδή $\Omega = X \cup X_\emptyset$. Επιπλέον διαμερίζουμε το X σύμφωνα με τις "εξόδους" σε K ξένα μεταξύ τους σύνολα $X_k = \{(x^n, y^n) | \, y^n = k, n = 1, \ldots, N_k\}$ ένα για κάθε κατηγορία, τότε $\Omega = \bigcup_k X_k \cup X_\emptyset$.

Σύμφωνα με τον κανόνα απόφασης του Bayes, ένας ταξινομητής αναθέτει σε ένα άγνω-

στο πρότυπο x^* την κατηγορία k^* με την μέγιστη εκ των υστέρων πιθανότητα. Αν παραλείψουμε τον όρο της εκ των υστέρων πιθανότητας που εξαρτάται μόνο από το x^*, τότε

$$k^* = \arg\max_k p(x^*|k)p(k) \tag{7.1}$$

όπου $p(x|k)$ είναι η κατανομή δοθείσης της κατηγορίας k, και $p(k)$ είναι η εκ των προτέρων πιθανότητα της κατηγορίας. Για δύο κατηγορίες k και k', μέσω της εξίσωσης $p(x|k)p(k) = p(x|k')p(k')$ ορίζεται το όριο απόφασης του ταξινομητή, το οποίο διαμερίζει τον χώρο των προτύπων.

Για να εκτιμήσουμε τις κατανομές κατηγοριών χρησιμοποιούμε το δίκτυο PRBF. Για είσοδο x η κατανομή κατηγορίας $p(x|k)$ είναι η k-οστή έξοδος ενός PRBF με J συνιστώσες (συναρτήσεις βάσης).

$$p(x|k) = \sum_{j=1}^{J} \pi_{jk}\, p(x|j) \tag{7.2}$$

Οι συντελεστές π_{jk} είναι μη-αρνητικοί και $\sum_j \pi_{jk} = 1$, ενώ οι συνιστώσες του είναι Gaussian

$$p(x|j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\{-\frac{1}{2}(x-\mu_j)^T(x-\mu_j)/\sigma_j^2\} \tag{7.3}$$

με κέντρο $\mu_j \in \Re^d$ και διακύμανση σ_j^2, υποθέτοντας σφαιρικό πίνακα συνδιακύμανσης. Προκειμένου να εκτιμήσουμε τις παραμέτρους του δικτύου

$$\theta = \{p(k), \pi_{jk}, \mu_j, \sigma_j |\ j = 1, \ldots, J,\ k = 1, \ldots, K\}$$

μεγιστοποιούμε την από κοινού πιθανοφάνεια, όπως στην [59]. Θεωρώντας ανεξάρτητες και όμοια κατανεμημένες παρατηρήσεις, η λογαριθμική από κοινού πιθανοφάνεια $\mathcal{L}$ των ταξινομημένων και αταξινόμητων προτύπων είναι

$$\begin{aligned} \mathcal{L} &= \log \prod_k \prod_{x \in X_k} p(x,k) \prod_{x \in X_\emptyset} p(x) \\ &= \sum_k \sum_{x \in X_k} \log p(k) \sum_j \pi_{jk} p(x|j) + \sum_{x \in X_\emptyset} \log \sum_k p(k) \sum_j \pi_{jk} p(x|j). \end{aligned} \tag{7.4}$$

Για την μεγιστοποίηση της $\mathcal{L}$ χρησιμοποιούμε τον αλγόριθμο EM [26]. Ο EM είναι ένας επαναληπτικός αλγόριθμος ο οποίος συγκλίνει σε ένα τοπικό μέγιστο της πιθανοφάνειας, και χρησιμοποιείται σε προβλήματα που υπάρχουν κρυμμένες μεταβλητές. Αυτές οι μεταβλητές καθορίζουν την λύση του προβλήματος, αν και δεν είναι παρατηρήσιμες. Στην περίπτωσή

μας οι κρυμμένες μεταβλητές καθορίζουν την συνιστώσα του δικτύου που παρήγαγε ένα πρότυπο, και την κατηγορία ενός αταξινόμητου προτύπου. Στην συνέχεια θα εξάγουμε τις εξισώσεις ανανέωσης του ΕΜ τις οποίες χρησιμοποιήσαμε [21].

Ορίζουμε μια κρυμμένη μεταβλητή $z^{(x)}$ για κάθε $x \in \Omega$ που αναθέτει στο πρότυπο μια κατηγορία και μια συνιστώσα. Κάθε $z^{(x)}$ είναι ένας δυαδικός $J \times K$ πίνακας, όπου $z_{jk}^{(x)} = 1$ αν στο x έχει ανατεθεί η k-οστή κατηγορία και η j-οστή συνιστώσα. Αυτή η ανάθεση είναι μοναδική, έτσι ώστε $\sum_j \sum_k z_{jk}^{(x)} = 1$. Επιπλέον για ένα ταξινομημένο πρότυπο (x, k) η αντίστοιχη $z^{(x)}$ είναι περιορισμένη έτσι ώστε $z_{j\ell}^{(x)} = 0$ για κάθε κατηγορία $\ell \neq k$ και για όλα τα j. Έτσι μια κρυμμένη μεταβλητή μπορεί να αναθέσει σε ένα ταξινομημένο πρότυπο οποιαδήποτε συνιστώσα, αλλά μόνο μια κατηγορία. Αυτό δεν ισχύει στην περίπτωση των αταξινόμητων προτύπων, στα οποία μπορεί να ανατεθεί οποιαδήποτε κατηγορία και οποιαδήποτε συνιστώσα. Δοθέντος του συνόλου των κρυμμένων μεταβλητών $Z = \{z^{(x)} | \, \forall x \in \Omega\}$, ορίζουμε την λογαριθμική πλήρη πιθανοφάνεια

$$\mathcal{Q} = \log \prod_{x \in \Omega} \prod_{k} \prod_{j} [p(k)\pi_{jk}p(x|j)]^{z_{jk}^{(x)}}. \tag{7.5}$$

Αν και δεν μπορούμε να υπολογίσουμε απευθείας την $\mathcal{Q}$, αφού εξαρτάται από τις άγνωστες τιμές των Z, μπορούμε όμως να υπολογίσουμε την αναμενόμενη τιμή της $\langle \mathcal{Q} \rangle$ ως προς την κατανομή των Z. Εφόσον η αναμενόμενη τιμή της $z_{jk}^{(x)}$ ισούται με την από κοινού εκ των υστέρων πιθανότητα $p(j, k|x)$ με την οποία το x έχει παραχθεί από την j-οστή συνιστώσα και ανήκει στην k-οστή κατηγορία, έπεται ότι

$$\langle \mathcal{Q} \rangle = \sum_{x \in \Omega} \sum_{k} \sum_{j} p(j, k|x) \log \{p(k)\pi_{jk}p(x|j)\}. \tag{7.6}$$

Ο αλγόριθμος ΕΜ επαναλαμβάνει δύο βήματα μέχρι να συγκλίνει. Στο Ε-βήμα υπολογίζει την αναμενόμενη τιμή της λογαριθμικής πλήρους πιθανοφάνειας $\langle \mathcal{Q} \rangle$, δοθείσης της τρέχουσας εκτίμησης για το διάνυσμα παραμέτρων θ. Στο Μ-βήμα εκτιμά το θ που μεγιστοποιεί την $\langle \mathcal{Q} \rangle$. Αυτή η διαδικασία συγκλίνει σε ένα τοπικό μέγιστο της λογαριθμικής από κοινού πιθανοφάνειας $\mathcal{L}$.

Πιο λεπτομερώς, στο Ε-βήμα υπολογίζουμε την $p(j, k|x)$ για κάθε $x \in \Omega$, $j \in \{1, \ldots, J\}$ και $k \in \{1, \ldots, K\}$ σύμφωνα με την

$$p(j, k|x) = p(j|k, x)p(k|x). \tag{7.7}$$

Αν το x είναι αταξινόμητο, τότε υπολογίζουμε τις $p(k|x)$ και $p(j|k,x)$ για κάθε κατηγορία k χρησιμοποιώντας το θεώρημα του Bayes

$$p(k|x) = \frac{p(x|k)p(k)}{\sum_\ell p(x|\ell)p(\ell)} \tag{7.8}$$

$$p(j|k,x) = \frac{\pi_{jk}p(x|j)}{\sum_i \pi_{ik}p(x|i)}. \tag{7.9}$$

Αν το x είναι ταξινομημένο, τότε εκμεταλλευόμαστε την πληροφορία για την κατηγορία του και θέτουμε

$$p(k|x) = \begin{cases} 1 & \text{αν } x \in X_k \\ 0 & \text{αν } x \notin X_k \end{cases} \tag{7.10}$$

και υπολογίζουμε την $p(j|k,x)$ παρομοίως

$$p(j|k,x) = \begin{cases} \frac{\pi_{jk}p(x|j)}{\sum_i \pi_{ik}p(x|i)} & \text{αν } x \in X_k \\ 0 & \text{αν } x \notin X_k \end{cases} \tag{7.11}$$

Στο Μ-βήμα μεγιστοποιούμε την $\langle Q \rangle$ ως προς το θ, δοθέντων των τρεχόντων εκτιμήσεων για τις από κοινού εκ των υστέρων πιθανότητες. Η λύση για κάθε $j \in \{1,\ldots,J\}$ και $k \in \{1,\ldots,K\}$ είναι

$$\mu_j = \frac{\sum_{x\in\Omega}\sum_k p(j,k|x)\,x}{\sum_{x\in\Omega}\sum_k p(j,k|x)} \tag{7.12}$$

$$\sigma_j^2 = \frac{1}{d}\frac{\sum_{x\in\Omega}\sum_k p(j,k|x)\,(x-\mu_j)^T(x-\mu_j)}{\sum_{x\in\Omega}\sum_k p(j,k|x)} \tag{7.13}$$

$$\pi_{jk} = \frac{\sum_{x\in\Omega} p(j,k|x)}{N_k + \sum_j \sum_{x\in X_\emptyset} p(j,k|x)} \tag{7.14}$$

$$p(k) = \frac{N_k + \sum_j \sum_{x\in X_\emptyset} p(j,k|x)}{N + N_\emptyset}. \tag{7.15}$$

Μια σημαντική πλευρά της εκπαίδευσης του δικτύου είναι η εκτίμηση του αριθμού των συνιστωσών που θα χρησιμοποιηθούν. Για να την αντιμετωπίσουμε υιοθετούμε την αυξητική προσέγγιση που περιγράψαμε στο Κεφάλαιο 6 για μάθηση με επίβλεψη, την οποία και τροποποιούμε κατάλληλα. Η αυξητική αυτή μέθοδος έχει δύο στάδια. Ξεκινάμε με ένα δίκτυο που έχει μόνο μια συνιστώσα, της οποίας οι παράμετροι υπολογίζονται εύκολα από τα στατιστικά του συνόλου δεδομένων. Στο πρώτο στάδιο της μεθόδου προσθέτουμε επαναληπτικά νέες συνιστώσες, μέχρι να φτάσουμε στην επιθυμητή πολυπλοκότητα. Στη συνέχεια, στο δεύτερο στάδιο διασπάμε όλες τις συνιστώσες προκειμένου να αυξήσουμε την επίδοση του δικτύου. Τα δυο στάδια παρουσιάζονται πιο αναλυτικά στην συνέχεια.

7.2.1 Προσθήκη Συνιστωσών

Δοθέντος ενός δικτύου με M συνιστώσες μπορούμε να κατασκευάσουμε ένα δίκτυο με $M+1$ συνιστώσες. Αν η δοθείσα κατανομή κατηγορίας είναι $p(x|k)$, τότε η προσθήκη μιας Gaussian συνιστώσας $q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2)$ έχει ως αποτέλεσμα την $\hat{p}(x|k)$ σύμφωνα με την:

$$\hat{p}(x|k) = (1 - \alpha_k)\, p(x|k) + \alpha_k\, q(x) \tag{7.16}$$

όπου α_k είναι η εκ των προτέρων πιθανότητα με την οποία η συνιστώσα q παράγει πρότυπα της κατηγορίας k. Ωστόσο πρέπει να εκτιμήσουμε τα α_k, το κέντρο μ_q και την διακύμανση σ_q^2 της q. Για αυτό αναζητούμε παραμέτρους τέτοιες ώστε η q να βρίσκεται κοντά στο όριο απόφασης. Η καλή εκτίμηση των κατανομών κατηγορίας κοντά στο όριο απόφασης είναι πολύ σημαντική για την επίδοση του ταξινομητή.

Ακολουθώντας την προσέγγιση του Κεφαλαίου 6, καταφεύγουμε στην μέθοδο kd-tree [8] για τον διαμερισμό του συνόλου των προτύπων. Χρησιμοποιούμε μόνο τα ταξινομημένα πρότυπα, τα οποία και χωρίζουμε σε M υποσύνολα

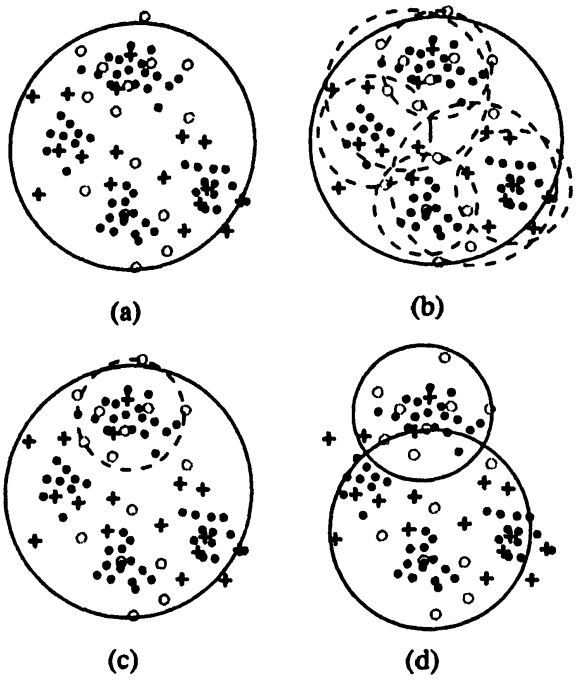
$$X_j = \{(x,k) |\, (x,k) \in X,\, p(j|k,x) > p(i|k,x), \forall i \neq j\}$$

ένα για κάθε κόμβο. Κατόπιν χωρίζουμε πάλι κάθε ένα από τα X_j σε έξι υποσύνολα κατασκευάζοντας ένα kd-tree. Η μέση τιμή και η διακύμανση κάθε υποσυνόλου αποτελούν τις υποψήφιες εκτιμήσεις των μ_q και σ_q^2. Οι εκτιμήσεις των αντιστοίχων εκ των προτέρων κατανομών είναι $\alpha_k = \pi_{jk}/2$. Διαμερίζοντας κάθε συνιστώσα δημιουργούμε $6M$ σύνολα από υποψήφια διανύσματα $\theta_q = \{\alpha_k, \mu_q, \sigma_q^2\}$, έτσι πρέπει να επιλέξουμε το πιο κατάλληλο σύμφωνα με κάποιο κριτήριο.

Χρησιμοποιούμε το κριτήριο που εισάγαμε στο προηγούμενο κεφάλαιο, οπότε υπολογίζουμε την αλλαγή στην λογαριθμική πιθανοφάνεια $\Delta\mathcal{L}_k^{(q)}$ της κατηγορίας k μετά την προσθήκη της q

$$\begin{aligned} \Delta\mathcal{L}_k^{(q)} &= \frac{1}{N_k} \left(\log \hat{p}(x|k) - \log p(x|k)\right) \\ &= \frac{1}{N_k} \sum_{x \in X_k} \log \left\{ 1 - \alpha_k + \alpha_k \frac{q(x)}{p(x|k)} \right\}. \end{aligned} \tag{7.17}$$

Διατηρούμε αυτά τα θ_q που αυξάνουν την λογαριθμική πιθανοφάνεια δύο τουλάχιστον κατηγοριών, και απορρίπτουμε τα υπόλοιπα. Για κάθε θ_q που κρατήσαμε, προσθέτουμε τους



Σχήμα 7.1: Η προσθήκη των δύο πρώτων συνιστωσών. Οι συνιστώσες του δικτύου σχεδιάστηκαν με συνεχείς γραμμές, και οι υποψήφιες συνιστώσες με διακεκομμένες. Οι κουκίδες αναπαριστούν τα ταξινόμητα πρότυπα σε ένα πρόβλημα δύο κατηγοριών.

θετικούς όρους $\Delta\mathcal{L}_k^q$ για να υπολογίσουμε την συνολική αύξηση της λογαριθμικής πιθανοφάνειας $\Delta\mathcal{L}_q$. Το υποψήφιο q^* που έχει την μέγιστη τιμή $\Delta\mathcal{L}_{q^*}$ καθορίζει τις παραμέτρους της νέας συνιστώσας που θα προσθέσουμε στο δίκτυο, αν αυτή η μέγιστη τιμή είναι μεγαλύτερη από ένα προκαθορισμένο κατώφλι (σε όλα τα πειράματά μας χρησιμοποιήσαμε την τιμή 0.01). Διαφορετικά θεωρούμε ότι η προσπάθεια προσθήκης μιας νέας συνιστώσας απέτυχε.

Μετά από την πετυχημένη προσθήκη μιας συνιστώσας, εφαρμόζουμε τον αλγόριθμο εκπαίδευσης με ημι-επίβλεψη, τον οποίο περιγράψαμε προηγουμένως. Αυτή η διαδικασία μπορεί να επαναληφθεί μέχρι να προσθέσουμε τον απαιτούμενο αριθμό συνιστωσών στο δίκτυο. Στο Σχήμα 7.1 απεικονίζεται η προσθήκη των δύο πρώτων συνιστωσών ενός δικτύου. Το αρχικό δίκτυο με μια μοναδική συνιστώσα απεικονίζεται στο Σχήμα 7.1(a). Οι έξι υποψήφιες συνιστώσες και η επιλεγμένη συνιστώσα απεικονίζονται στο Σχήμα 7.1(b) και στο Σχήμα 7.1(c) αντίστοιχα. Στο Σχήμα 7.1(d) απεικονίζεται το δίκτυο μετά την εφαρμογή του ΕΜ για μάθηση με ημι-επίβλεψη.

7.2.2 Διάσπαση Συνιστωσών

Μετά το στάδιο της προσθήκης συνιστωσών, μπορεί να υπάρχουν συνιστώσες τοποθετημένες σε περιοχές όπου οι κατηγορίες επικαλύπτονται. Προκειμένου να αυξήσουμε την γενικευτική ικανότητα του δικτύου διασπάμε τέτοιες συνιστώσες. Σε αυτό το στάδιο της εκπαίδευσης χρησιμοποιούμε και τα ταξινομημένα και τα αταξινόμητα πρότυπα. Για κάθε συνιστώσα υπολογίζουμε τις από κοινού εκ των υστέρων πιθανότητες $p(j,k|x)$, και διαπιστώνουμε αν η συνιστώσα είναι υπεύθυνη για πρότυπα πολλών κατηγοριών. Αν $\sum_{x\in\Omega} p(j,k|x) > 0$, τότε την αφαιρούμε από το δίκτυο και προσθέτουμε μια ξεχωριστή συνιστώσα για την k-οστή κατηγορία. Έτσι τελικά κάθε συνιστώσα είναι υπεύθυνη για πρότυπα μόνο μιας κατηγορίας. Διασπώντας την συνιστώσα $p(x|j)$, η συνιστώσα που προκύπτει για την κατηγορία k είναι μια Gaussian $p(x|j,k)$ με κέντρο μ_{kj}, διακύμανση σ^2_{kj} και συντελεστή στάθμισης π_{jk}. Αυτές οι παράμετροι εκτιμώνται σύμφωνα με τις:

$$\mu_{kj} = \frac{\sum_{x\in\Omega} p(j,k|x)\, x}{\sum_{x\in\Omega} p(j,k|x)} \tag{7.18}$$

$$\sigma^2_{kj} = \frac{1}{d}\frac{\sum_{x\in\Omega} p(j,k|x)\,(x-\mu_{kj})^T(x-\mu_{kj})}{\sum_{x\in\Omega} p(j,k|x)} \tag{7.19}$$

$$\pi_{jk} = \frac{\sum_{x\in\Omega} p(j,k|x)}{N_k + \sum_j \sum_{x\in X_\emptyset} p(j,k|x)}. \tag{7.20}$$

Συνεπώς η κατανομή κατηγορίας εκτιμάται ως:

$$p(x|k) = \sum_j \pi_{jk} p(x|j,k). \tag{7.21}$$

Στην περίπτωση ενός συνόλου εκπαίδευσης όπου όλα τα πρότυπα είναι ταξινομημένα, η πιθανοφάνεια κάθε κατηγορίας αυξάνεται μετά την διάσπαση, όπως έχει αποδειχθεί στην [78]. Ωστόσο στην μάθηση με ημι-επίβλεψη δεν μπορούμε να εγγυηθούμε ότι η διάσπαση αυξάνει την από κοινού πιθανοφάνεια.

7.3 Αλγόριθμος Ενεργητικής Μάθησης

Στη προηγούμενη ενότητα περιγράψαμε μια αυξητική μέθοδο για την εκπαίδευση του PRBF χρησιμοποιώντας ταξινομημένα και αταξινόμητα πρότυπα, Στην συνέχεια ενσωματώνουμε

αυτή την μέθοδο σε έναν αλγόριθμο ενεργητικής μάθησης, όπου επαναληπτικά επιλέγουμε ένα αταξινόμητο πρότυπο και ζητάμε να μάθουμε την κατηγορία του. Αφού μας δοθεί η απάντηση, προσθέτουμε το ταξινομημένο πια πρότυπο στο σύνολο εκπαίδευσης και εκπαιδεύουμε πάλι το δίκτυο. Το κρίσιμο σημείο είναι η επιλογή ενός προτύπου που ωφελεί όσο το δυνατό περισσότερο τον ταξινομητή μας. Προτείνουμε την επιλογή ενός προτύπου που βρίσκεται κοντά στο όριο απόφασης. Με αυτό τον τρόπο διευκολύνουμε την επαναληπτική προσθήκη συνιστωσών πάνω στο όριο απόφασης, όπως περιγράψαμε προηγουμένως.

Ως κριτήριο για την επιλογή του κατάλληλου προτύπου προτείνουμε τον λόγο των εκ των υστέρων πιθανοτήτων των κατηγοριών. Για κάθε αταξινόμητο πρότυπο $x \in X_\emptyset$ υπολογίζουμε τις εκ των υστέρων πιθανότητες $p(k|x)$ για κάθε κατηγορία, και στην συνέχεια βρίσκουμε τις δυο κατηγορίες με τις μεγαλύτερες τιμές:

$$\kappa_1^{(x)} = \arg\max_k p(k|x),\ \kappa_2^{(x)} = \arg\max_{k \neq \kappa_1^{(x)}} p(k|x). \tag{7.22}$$

Επιλέγουμε να ρωτήσουμε την κατηγορία του $\hat{x}$ που έχει τον μικρότερο λόγο μεταξύ των δυο μεγαλύτερων εκ των υστέρων πιθανοτήτων των κατηγοριών:

$$\hat{x} = \arg\min_{x \in X_\emptyset} \log \frac{p(\kappa_1^{(x)}|x)}{p(\kappa_2^{(x)}|x)}. \tag{7.23}$$

Με αυτό τον τρόπο επιλέγουμε ένα αταξινόμητο πρότυπο που βρίσκεται κοντά στο όριο απόφασης του τρέχοντος ταξινομητή. Παρατηρήστε ότι σύμφωνα με την (7.1) ταξινομούμε τα πρότυπα στην κατηγορία με την μέγιστη εκ των υστέρων πιθανότητα. Οπότε για κάποιο x στο όριο απόφασης ισχύει ότι $p(\kappa_1^{(x)}|x) = p(\kappa_2^{(x)}|x)$. Συνεπώς αν ένα πρότυπο προσεγγίζει το όριο απόφασης μεταξύ δύο κατηγοριών, τότε ο αντίστοιχος λογαριθμικός λόγος των εκ των υστέρων πιθανοτήτων τείνει στο μηδέν.

Συνοψίζοντας την μέθοδο που παρουσιάσαμε, προτείνουμε τον ακόλουθο αλγόριθμο για ενεργητική μάθηση βασισμένη σε μια "δεξαμενή προτύπων":

1. Είσοδος: Το σύνολο X των ταξινομημένων προτύπων, το σύνολο $X_\emptyset$ των αταξινόμητων προτύπων, και ένα εκφυλισμένο δίκτυο $PRBF_{J=1}$ με μια συνιστώσα.

2. Για $s = 0, \ldots, S-1$

(α΄) Προσθέτουμε μια συνιστώσα στο δίκτυο $PRBF_{J+s}$ για να σχηματίσουμε
το $PRBF_{J+s+1}$.

(β΄) Εφαρμόζουμε τον ΕΜ για μάθηση με ημι-επίβλεψη του $PRBF_{J+s+1}$.

3. Για $s = 0, \ldots, S$

(α΄) Διασπάμε τις συνιστώσες του $PRBF_{J+s}$ για να σχηματίσουμε το $PRBF_{J+s}^{split}$.

4. Επιλέγουμε το δίκτυο $PRBF_{J^*}^{split} \in \{PRBF_J^{split}, \ldots, PRBF_{J+S}^{split}\}$ που μεγιστοποιεί την από κοινού πιθανοφάνεια.

5. Ορίζουμε το τρέχον δίκτυο: $PRBF_J = PRBF_{J^*}$.

6. Αν το $X_\emptyset$ είναι κενό πηγαίνουμε στο βήμα 7, αλλιώς

(α΄) Επιλέγουμε ένα αταξινόμητο πρότυπο $\hat{x}$ σύμφωνα με την (7.23), και ρωτάμε την κατηγορία του $\hat{y}$.

(β΄) Ανανεώνουμε τα σύνολα: $X = X \cup \{(\hat{x}, \hat{y})\}$ και $X_\emptyset = X_\emptyset \setminus \{\hat{x}\}$.

(γ΄) Πηγαίνουμε στο βήμα 2.

7. Έξοδος: Διασπάμε τις συνιστώσες του $PRBF_J$ για να σχηματίσουμε το τελικό δίκτυο $PRBF_J^{split}$.

Σε όλα μας τα πειράματα χρησιμοποιούμε $S = 1$, έτσι προσπαθούμε να προσθέσουμε μια μόνο συνιστώσα σε κάθε επανάληψη της ενεργητικής μάθησης.

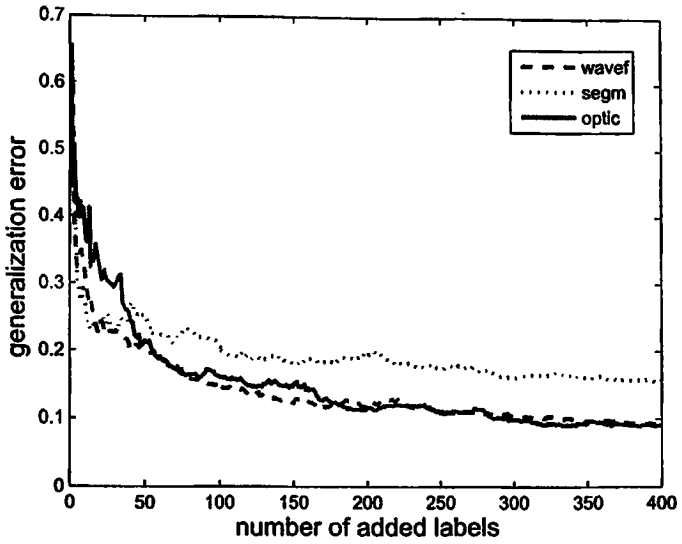
7.4 Πειράματα

Για την πειραματική αξιολόγηση της μεθόδου μας χρησιμοποιήσαμε τρία σύνολα, διαθέσιμα από το UCI repository [11]. Το πρώτο σύνολο είναι το segmentation, που αποτελείται από 2310 πρότυπα με 19 χαρακτηριστικά που ανήκουν σε 7 κατηγορίες. Το δεύτερο σύνολο είναι το waveform, που αποτελείται από 5000 πρότυπα με 21 χαρακτηριστικά που ανήκουν σε 3

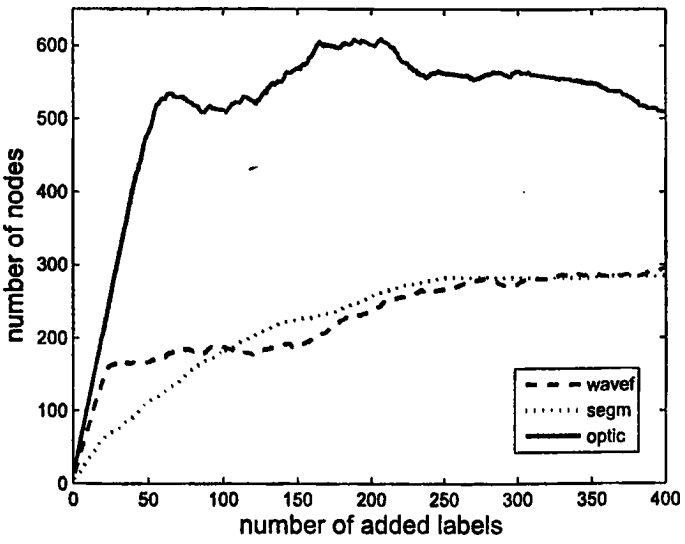
κατηγορίες. Το τελευταίο σύνολο είναι το optical digits, που αποτελείται από 5620 πρότυπα με 62 χαρακτηριστικά που ανήκουν σε 10 κατηγορίες. Όλα τα σύνολα κανονικοποιήθηκαν, έτσι ώστε τα χαρακτηριστικά τους να έχουν μέση τιμή μηδέν και τυπική απόκλιση μονάδα. Σε όλα τα πειράματα εφαρμόσαμε τον προτεινόμενο αλγόριθμο ξεκινώντας με 50 τυχαία επιλεγμένα ταξινομημένα πρότυπα, και επιλέξαμε ενεργητικά άλλα 400. Κάθε πείραμα επαναλήφθηκε πέντε φορές, και υπολογίσαμε το αναμενόμενο σφάλμα γενίκευσης σε ένα ξεχωριστό σύνολο ελέγχου που περιείχε το 10% του αρχικού συνόλου. Στο Σχήμα 7.2 εμφανίζεται το αναμενόμενο σφάλμα γενίκευσης και στο Σχήμα 7.3 ο αναμενόμενος αριθμός συνιστωσών του δικτύου, μετά από την απόκτηση κάθε νέου ταξινομημένου προτύπου. Τα αποτελέσματα είναι πολύ ικανοποιητικά, καθώς το σφάλμα γενίκευσης σχεδόν υποδιπλασιάστηκε σε όλες τις περιπτώσεις μετά από την προσθήκη 50 ταξινομημένων προτύπων. Μετά από την προσθήκη 300 προτύπων το σφάλμα είχε σχεδόν συγκλίνει, και η προσθήκη επιπλέον προτύπων πρόσφερε αμελητέα βελτίωση. Μετά από την προσθήκη 400 ταξινομημένων προτύπων το αναμενόμενο σφάλμα για το σύνολο segmentation ήταν 0.156, για το σύνολο waveform 0.091 και για το σύνολο optical digits 0.089. Ο αριθμός των συνιστωσών συνέκλινε αργότερα από το σφάλμα, αλλά τελικά και αυτός σταθεροποιήθηκε. Ο μέσος αριθμός συνιστωσών μετά από την προσθήκη 400 ταξινομημένων προτύπων για το σύνολο segmentation ήταν 285.2, για το σύνολο waveform 294.6 και για το σύνολο optical digits 509.

7.5 Συμπεράσματα

Προτείναμε έναν αλγόριθμο για την ενεργητική μάθηση του ταξινομητή PRBF. Εξηγάγαμε ένα EM αλγόριθμο για μάθηση με ημι-επίβλεψη, και μια αυξητική παραλλαγή του που προσθέτει συνιστώσες στο δίκτυο σειριακά. Χρησιμοποιήσαμε αυτή την μέθοδο για να εκτιμήσουμε τις κατανομές των προτύπων δοθέντων των κατηγοριών για ενεργητική μάθηση βασισμένη σε μια "δεξαμενή προτύπων". Τα πειραματικά αποτελέσματα είναι ενθαρρυντικά, και μικρές τροποποιήσεις της μεθόδου μπορεί να βελτιώνουν επιπλέον την επίδοσή της. Για παράδειγμα θα μπορούσαμε να ρωτήσουμε την κατηγορία για μια ομάδα από αταξινόμητα



Σχήμα 7.2: Το αναμενόμενο σφάλμα γενίκευσης του δικτύου για ενεργητική μάθηση βασισμένη σε μια "δεξαμενή προτύπων".



Σχήμα 7.3: Ο αναμενόμενος αριθμός συνιστωσών του δικτύου για ενεργητική μάθηση βασισμένη σε μια "δεξαμενή προτύπων".

πρότυπα, πριν επιχειρήσουμε να προσθέσουμε μια νέα συνιστώσα. Ένα άλλο ενδιαφέρον θέμα είναι η πολυπλοκότητα του τελικού δικτύου. Αυτό οφείλεται στο ότι σε κάθε επανάληψη η μέθοδος επιχειρεί να προσθέσει μια νέα συνιστώσα. Τελικά οι περισσότερες έχουν πολύ μικρό συντελεστή στάθμισης, ωστόσο θα μπορούσαμε να χρησιμοποιήσουμε ένα σύνολο επικύρωσης για καλύτερη επιλογή μοντέλου.

Η μελλοντική έρευνα μπορεί να εστιαστεί σε μια πιο διεξοδική μελέτη της μεθόδου, και στην εξέταση πολλών ζητημάτων, με πιο σημαντικό την σύγκριση με άλλες μεθόδους επιλογής για την ενεργητική μάθηση των κατηγοριών. Επίσης μπορεί να μελετηθεί το πρόβλημα της ανίχνευσης πρωτοτυπίας (novelty detection), καθώς σχετίζεται άμεσα με την ενεργητική μάθηση. Σε αυτό το πρόβλημα ο ταξινομητής έχει την ικανότητα να εντοπίζει και να μαθαίνει καινούριες κατηγορίες χωρίς να γνωρίζει την ύπαρξή τους από πριν. Φαίνεται ότι η παραπάνω μεθοδολογία μπορεί να επεκταθεί ώστε να καλύψει και τέτοιες περιπτώσεις.

Κεφαλαιο 8

Συμπερασματα

Στην παρούσα διατριβή μελετήσαμε προβλήματα που εμφανίζονται κατά την ταξινόμηση και την ομαδοποίηση, στα οποία δώσαμε λύσεις ανάγοντάς τα σε προβλήματα εκτίμησης κατανομών. Συγκεκριμένα στο πρόβλημα ταξινόμησης υιοθετήσαμε την generative προσέγγιση, οπότε εκτιμήσαμε την κατανομή των προτύπων δοθείσας της κατηγορίας και χρησιμοποιήσαμε τον κανόνα του Bayes για να πάρουμε αποφάσεις. Τα μοντέλα που μελετήσαμε ήταν το πιθανοτικό δίκτυο ακτινικών συναρτήσεων βάσης (PRBF) για τα προβλήματα ταξινόμησης, και η μίξη κανονικών κατανομών για τα προβλήματα ομαδοποίησης. Για την εκπαίδευσή τους βασιστήκαμε στον αλγόριθμο EM και στην variational Bayesian μέθοδο που ακολουθεί την ίδια φιλοσοφία. Σε όλα τα προβλήματα δώσαμε ιδιαίτερη έμφαση στην επιλογή μοντέλου (αριθμός συνιστωσών), αφού είναι το μεγαλύτερο πρόβλημα που έχουν όλα τα παραμετρικά μοντέλα. Οι λύσεις που προτείναμε ακολουθούσαν μια αυξητική προσέγγιση, δηλαδή την σταδιακή προσθήκη συνιστωσών ξεκινώντας από μια αρχική. Επιπλέον για την ομαδοποίηση μελετήσαμε και την επιλογή χαρακτηριστικών, ένα δύσκολο πρόβλημα που έχει αναδειχθεί πρόσφατα.

Ανακεφαλαιώνοντας, στο Κεφάλαιο 3 προτείναμε μια αυξητική μέθοδο για την επιλογή μοντέλου και την εκπαίδευση μιας μίξης, που βασίζεται στη variational Bayesian μέθοδο. Στο Κεφάλαιο 4 προτείναμε μια variational Bayesian μέθοδο για την εκπαίδευση μιας μίξης που αντιμετωπίζει ταυτόχρονα το πρόβλημα της επιλογής χαρακτηριστικών και της επιλογής μοντέλου. Στο Κεφάλαιο 5 προτείναμε μια μέθοδο για την σταδιακή κατάτμηση

εικόνων με μίξη κανονικών κατανομών, που επιλέγει με ενεργό τρόπο νέα εικονοστοιχεία για να βελτιώσει την κατάτμηση. Στο Κεφάλαιο 6 προτείναμε μια αυξητική μέθοδο για την εκπαίδευση του δικτύου PRBF, η οποία προσθέτει επαναληπτικά νέες συνιστώσες κοντά στο όριο απόφασης. Τέλος στο Κεφάλαιο 7 προτείναμε μια μέθοδο ενεργητικής μάθησης για την αυξητική εκπαίδευση του PRBF. Η μέθοδος βασίζεται στον αλγόριθμο EM με ημι-επίβλεψη, και επιλέγει ενεργά πρότυπα κοντά στο όριο απόφασης.

Όσο αφορά θέματα για περαιτέρω έρευνα πιστεύουμε ότι οι προσεγγίσεις και οι μεθοδολογίες που διέπουν τα θέματα που εξετάζει η διατριβή αποδείχτηκαν πολύ αποτελεσματικές, και να αξίζει αξιοποιηθούν περισσότερο. Αναφέρουμε πρώτα κάποια θέματα που προέκυψαν στα προβλήματα μάθησης χωρίς επίβλεψη. Όσο αφορά την αυξητική variational Bayesian μέθοδο όπως την προτείναμε στο Κεφάλαιο 3 για την μίξη κανονικών κατανομών, η έρευνα θα μπορούσε να εστιάσει στην αναζήτηση διαφορετικών τρόπων για τον ορισμό της εκ των προτέρων πιθανότητας του πίνακα ακριβείας, και ενδεχόμενη χρήση ενός ιεραρχικού μοντέλου (hierarchical prior) για τις εκ των προτέρων πιθανότητες, δηλαδή εισαγωγή εκ των προτέρων πιθανότητας στις παραμέτρους της εκ των προτέρων πιθανότητας. Παρόμοιο ιεραρχικό μοντέλο εκ των προτέρων πιθανοτήτων θα πρέπει να εξεταστεί και για την variational Bayesian μέθοδο επιλογής χαρακτηριστικών του Κεφαλαίου 4, ειδικά χρησιμοποιώντας μια εκ των προτέρων πιθανότητα για τη σημαντικότητα των χαρακτηριστικών. Στην ίδια μέθοδο ανοικτό είναι και το πρόβλημα της χρήσης ενός πλήρους πίνακα συνδιακύμανσης για τις συνιστώσες, ώστε να απαλλαγούμε από την περιοριστική υπόθεση των ανεξάρτητων χαρακτηριστικών. Εξετάζοντας την μέθοδο για κατάτμηση εικόνων θεωρούμε ότι η ενσωμάτωσή της σε εφαρμογές ομαδοποίησης συνόλου εικόνων και αναζήτησης εικόνων σε βάσεις θα μπορούσε να προσφέρει πλεονεκτήματα στον τομέα της ταχύτητας και της μοντελοποίησης. Ιδιαίτερο ενδιαφέρον παρουσιάζει και το πρόβλημα της επιλογής χαρακτηριστικών για την κατάτμηση εικόνων με βάση χαρακτηριστικά χαμηλού επιπέδου, όπως το χρώμα και η υφή.

Ενδιαφέροντα ζητήματα προκύπτουν όμως και στα προβλήματα μάθησης με επίβλεψη. Όσο αφορά την αυξητική μάθηση του PRBF θα είχε ενδιαφέρον η εφαρμογή της σε δίκτυα με διαφορετικές συνιστώσες, ιδιαίτερα με διακριτές κατανομές ή με συνδυασμό διακριτών και συνεχών κατανομών, αφού πολύ συχνά στην πράξη αντιμετωπίζουμε προβλήματα με συνεχή

και διακριτά χαρακτηριστικά ταυτόχρονα. Παραμένοντας όμως στα συνεχή, θα ήταν ενδιαφέρουσα η μελέτη δικτύων με συνιστώσες από Factor Analyzers για να αντιμετωπιστούν τα προβλήματα που έχουν οι κανονικές συνιστώσες, όταν δεν προκύπτει θετικά ορισμένος πίνακας συνδιακύμανης. Επίσης η εφαρμογή της variational Bayesian μεθόδου για την εκπαίδευση ενός δικτύου PRBF θα αποτελούσε μια ευκαιρία για συγκρίσεις με την αυξητική μέθοδο που προτείναμε, για καλύτερη κατανόηση της χρησιμότητας του ορίου απόφασης ως κριτηρίου για την τοποθέτηση νέων συνιστωσών, και ίσως για βελτιωμένα κριτήρια. Επιπλέον η μέθοδος ενεργητικής μάθησης του PRBF θα μπορούσε να διαμορφωθεί κατάλληλα για προβλήματα ανίχνευσης outlier ή νέων άγνωστων κατηγοριών στα πρότυπα. Τέλος όλες οι προτεινόμενοι μέθοδοι μπορούν να χρησιμοποιηθούν και να δοκιμαστούν σε πραγματικά προβλήματα από διάφορα επιστημονικά πεδία για τα οποία υπάρχουν διαθέσιμα σύνολα εκπαίδευσης.

BIBLIOGRAPHY

[1] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

[2] E. Alpaydin. *Introduction to Machine Learning.* MIT Press, 2004.

[3] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, second edition, 1984.

[4] G. B. Arfken and H. J. Weber. *Mathematical Methods For Physicists.* Elsevier, sixth edition, 2005.

[5] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 21–30. Morgan Kaufmann, 1999.

[6] H. Attias. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

[7] H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 2003.

[8] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[9] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41(3–4):561–575, 2003.

[10] C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[11] C. L. Blake and C. J. Merz. UCI repository of machine learning databases. [Online]. Available: `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 1998.

[12] W. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994.

[13] P. Carbonetto, N. de Freitas, P. Gustafson, and N. Thompson. Bayesian feature weighting for unsupervised learning, with application to object recognition. In *9th International Conference on Artificial Intelligence and Statistics*, 2003.

[14] M. A. Carreira-Perpinan and Williams C. K. I. On the number of modes of a Gaussian mixture. In *4th International Conference, Scale Space 2003*, pages 625–640. Springer-Verlag, 2003.

[15] V. Castelli and T. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16:105–111, 1995.

[16] S. Chretien and A. O. Hero. Kullback proximal algorithms for maximum likelihood estimation. *IEEE Trans. on Information Theory*, 46(5):1800–1810, 2000.

[17] D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[18] C. Constantinopoulos and A. Likas. Efficient training algorithms for the probabilistic RBF network. In *Proc. 3rd Hellenic Conference on Artificial Inteligence, SETN 2004*, pages 183–190. Springer, 2004.

[19] C. Constantinopoulos and A. Likas. Active Bayesian mixture learning for image modeling and segmentation using low level features. In *Proc. Machine Learning for Signal Processing 2006*, page to appear, 2006.

[20] C. Constantinopoulos and A. Likas. An active learning approach for training the probabilistic RBF classification network. In *Proc. ICINCO 2006, Workshop on Artificial Neural Networks and Intelligent Information Processing*, page to appear, 2006.

[21] C. Constantinopoulos and A. Likas. Active learning with the probabilistic RBF classifier. In *Proc. 16th International Conference on Artificial Neural Networks*, page to appear, 2006.

[22] C. Constantinopoulos and A. Likas. An incremental training method for the probabilistic RBF network. *IEEE Transactions on Neural Networks*, to appear, 2006.

[23] C. Constantinopoulos, M. K. Titsias, and A. Likas. Bayesian feature and model selection for Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6), 2006.

[24] A. Corduneanu and C. M. Bishop. Variational Bayesian model selection for mixture distributions. In *Artificial Intelligence and Statistics 2001*, pages 27–34. Morgan Kaufmann, 2001.

[25] C. Cortes and V. Vapnik. Support–Vector networks. *Machine Learning*, 20(3):273–297, 1995.

[26] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[27] J. Dy and C. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5:845–889, 2004.

[28] B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Chapman and Hall, first edition, 1981.

[29] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[30] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by comitee algorithm. *Machine Learning*, 28:133–168, 1997.

[31] J. H. Friedman and J. J. Meulman. Clustering objects on subsets of attributes. *Journal of Royal Statistical Society*, 66(4):815–849, 2004.

[32] K. Fukumizu, S. Akaho, and S. Amari. Critical lines in symmetry of mixture models and its application to component splitting. In *Advances in Neural Information Processing Systems 15*, pages 865–872. MIT Press, 2003.

[33] Z. Ghahramani. An introduction to Hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15:9–42, 2001.

[34] Z. Ghahramani and M. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press, 2000.

[35] Z. Ghahramani and M. Beal. Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems 13*, pages 507–513. MIT Press, 2001.

[36] Z. Ghahramani and M. Jordan. Supervised learning from incomplete data via an EM approach. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*. Morgan Kaufmann, 1994.

[37] J. Goldberger, S. Gordon, and H. Greenspan. Unsupervised image-set clustering using an information theoretic framework. *IEEE Transactions on Image Processing*, 15(2):449–458, 2006.

[38] P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrica*, 82:711–732, 1995.

[39] R. J. Hathaway. Another interpretation of the EM algorithm for mixture distributions. *Statistics and Probability Letters*, 4:53–56, 1986.

[40] D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. Kluwer, 1998.

[41] A. O. Hero and A. J. Fessler. Convergence in norm for alternating Expectation-Maximization (EM) type algorithms. *Statistica Sinica*, 5(1):41–54, 1995.

[42] P. D. Hoff. Model-based subspace clustering. *Bayesian Analysis*, 1(2):321–344, 2006.

[43] G.-B. Huang, P. Saratchandran, and N. Sundararajan. An efficient sequential learning algorithm for growing and pruning RBF (GAP-RBF) networks. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 34(6):2284–2292, 2004.

[44] G.-B. Huang, P. Saratchandran, and N. Sundararajan. A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation. *IEEE Trans. Neural Networks*, 16(1):57–67, 2005.

[45] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

[46] T. Jaakkola. Tutorial on variational approximation methods. In M. Opper and D. Saad, editors, *Advanced mean field methods: theory and practice*, pages 129–160. MIT Press, 2000.

[47] A. K. Jain, R. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Tarnsactions on Pattern Analysis and Machine Intelligence*, 22(1):4–38, 2000.

[48] M. I. Jordan, Z. Ghahramani, Jaakkola T. S., and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. Kluwer, 1998.

[49] V. Kadirkamanathan and M. Niranjan. A function estimation approach to sequential learning with neural networks. *Neural Computation*, 5(6):954–975, 1993.

[50] S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous multivariate distributions*, volume 1. John Wiley & Sons, second edition, 2000.

[51] M. H. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using a mixture model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004.

[52] A. Likas, N. A. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, vol. 36:451–461, 2003.

[53] J. S. Liu, J. L. Zhang, M. J. Palumbo, and C. E. Lawrence. Bayesian clustering with variable and transformation selections. *Bayesian Statistics*, 7:249–276, 2003.

[54] Z. K. Mao and G. Huang. Neuron selection for RBF neural network classifier based on data structure preserving criterion. *IEEE Trans. Neural Networks*, 16(6):1531–1540, 2005.

[55] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th International Conference on Computer Vision*, volume 2, pages 416–423, 2001.

[56] A. K. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In Jude W. Shavlik, editor, *Proc. 15th International Conference on Machine Learning*. Morgan Kaufmann, 1998.

[57] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. John Wiley & Sons, 1997.

[58] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2000.

[59] D. Miller and H. Uyar. Combined learning and use for a mixture model equivalent to the RBF classifier. *Neural Computation*, 10:281–293, 1998.

[60] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[61] R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–370. Kluwer, 1998.

[62] Y. Oyang, S. Hwang, Y. Ou, C. Chen, and Z. Chen. Data classification with radial basis function networks based on a novel kernel density estimation algorithm. *IEEE Trans. Neural Networks*, 16(1):225–236, 2005.

[63] N. Pal and S. Pal. A review of image segmentation techniques. *Pattern Recognition*, 26:1277–1294, 1993.

[64] J. Park and Sandberg J. W. Universal approximation using radial basis function network. *Neural Computation*, 3(2):246–257, 1991.

[65] J. Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225, 1991.

[66] T. Poggio and F. Girosi. Network for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

[67] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[68] S. Richardson and P. Green. On Bayesian analysis of mixtures with unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59(4):731–792, 1997.

[69] T. J. Sanjay-Gopal, S. Hebert. Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm. *IEEE Transactions on Image Processing*, 7(7):1014–1028, 1998.

[70] G. Sfikas, C. Constantinopoulos, A. Likas, and N. P. Galatsanos. An analytic distance metric for Gaussian mixture models with application in image retrieval. In *Proc. 15th International Conference on Artificial Neural Networks*, volume 2, pages 835–840, 2005.

[71] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[72] P. Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.

[73] D. F. Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.

[74] S. Tadjudin and A. Landgrebe. Robust parameter estimation for mixture model. *IEEE Trans. Geoscience and Remote Sensing*, 38:439–445, 2000.

[75] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[76] M. K. Titsias and A. Likas. A probabilistic RBF network for classification. In *Proceedings of International Joint Conference on Neural Networks*, volume 4, pages 238–243. IEEE, 2000.

[77] M. K. Titsias and A. Likas. Shared kernel models for class conditional density estimation. *IEEE Trans. Neural Networks*, 12(5):987–997, 2001.

[78] M. K. Titsias and A. Likas. Mixture of experts classification using a hierarchical mixture model. *Neural Computation*, 14(9):2221–2244, 2002.

[79] M. K. Titsias and A. Likas. Class conditional density estimation using mixtures with constrained component sharing. *IEEE Trans. Pattern Anal. and Machine Intell.*, 25(7):924–928, 2003.

[80] N. Ueda and Z. Ghahramani. Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15(10):1223–1241, 2002.

[81] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. SMEM algorithm for mixture models. *Neural Computation*, 12(9):2109–2128, 2000.

[82] N. A. Vlassis and A. Likas. A Greedy-EM algorithm for Gaussian mixture learning. *Neural Processing Letters*, 15:77–87, 2002.

[83] J. C. F. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11:95–103, 1983.

[84] L. Yingwei, N. Sundararajan, and P. Saratchandran. A sequential learning scheme for function approximation using minimal radial basis function neural networks. *Neural Computation*, 9(2):461–478, 1997.

[85] Z. Zhang, C. Chen, J. Sun, and K. L. Chan. EM algorithms for Gaussian mixtures with split-and-merge operation. *Pattern Recognition*, 36(9):1973–1983, 2003.

[86] X. Zhu, J. Lafferty, and Z. Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. 20th International Conference on Machine Learning*, 2003.

ΒΙΟΓΡΑΦΙΚΟ

Ο Κωνσταντίνος Κωνσταντινόπουλος πήρε το Πτυχίο Πληροφορικής και το Μεταπτυχιακό Δίπλωμα Ειδίκευσης στην Πληροφορική από το Τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων, το 2000 και το 2002 αντίστοιχα. Από το 2002 είναι Υποψήφιος Διδάκτορας του Τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων.

Τα ερευνητικά του ενδιαφέροντα περιλαμβάνουν προβλήματα της μηχανικής μάθησης, τα νευρωνικά δίκτυα, τις μικτές κατανομές, και προσεγγιστικές μεθόδους κατά Bayes, καθώς και εφαρμογές της μηχανικής μάθησης σε προβλήματα επεξεργασίας εικόνας.

Άρθρα σε διεθνή περιοδικά με σύστημα κριτών:

- Constantinopoulos, C. and Titsias, M. K. and Likas, A. Bayesian Feature and Model Selection for Gaussian Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (6), 2006.

- Constantinopoulos, C. and Likas, A. An Incremental Training Method for the Probabilistic RBF Network. *IEEE Transactions on Neural Networks*, to appear, 2006.

Άρθρα σε πρακτικά συνεδρίων με σύστημα κριτών:

- Constantinopoulos, C. and Likas, A. Active Learning with the Probabilistic RBF Classifier. *Proc. 16th International Conference on Artificial Neural Networks*, to appear, 2006.

- Constantinopoulos, C. and Likas, A. Active Bayesian Mixture Learning for Image Modeling and Segmentation Using Low Level Features. *Proc. Machine Learning for Signal Processing 2006*, to appear, 2006.

- Constantinopoulos, C. and Likas, A. An Active Learning Approach for Training the Probabilistic RBF Classification Network. *Proc. ICINCO 2006, Workshop on Artificial Neural Networks and Intelligent Information Processing*, to appear, 2006.

- Sfikas, G. and Constantinopoulos, C. and Likas, A. and Galatsanos, N. P. An Analytic Distance Metric for Gaussian Mixture Models with Application in Image Retrieval. *Proc. 15th International Conference on Artificial Neural Networks*, volume 2, pages 835–840, 2005.

- Constantinopoulos, C. and Likas, A. Efficient Training Algorithms for the Probabilistic RBF Network. *Proc. 3rd Hellenic Conference on Artificial Inteligence, SETN 2004*, Springer, pages 183–190, 2004.

- Constantinopoulos, C. and Titsias, M. K. and Likas, A. A Bayesian Regularization Method for the Probabilistic RBF Network. *Proc. 2nd Hellenic Conference on Artificial Inteligence, SETN 2002*, Springer, pages 337–345, 2002.

- Frossyniotis, D. and Vrettos, S. and Stafylopatis, A. and Constaninopoulos, C. and Fotiadis, D. I. and Likas, A. and Potamias, G. and Naka, A. and Tzimas, Th. and Michalis, L. K. An Intelligent System for the Early Diagnosis of Coronary Artery Disease. *Neural Networks and Expert Systems in Medicine and Healthcare, NNESMED 2001*, 2001.

Τεχνική αναφορά:

- Constantinopoulos, C. and Likas, A. Bayesian Gaussian Mixture Learning Based on Variational Component Splitting. Computer Science Deptartment, University of Ioannina, TR 2005-24, 2005.