



ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΚΑΙ ΔΙΟΙΚΗΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ SOM ΚΑΙ Η ΠΑΡΑΛΛΑΓΕΣ ΤΟΥΣ

Σακελλάρης Θεοχάρης

Επιβλέπων: Καρβέλης Πέτρος
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ

Ηράκλειο Κρήτης, Απρίλιος, 2026

NEURAL NETWORK SOM AND THEIR VARIATIONS

Εγκρίθηκε από τριμελή εξεταστική επιτροπή
Αρτα, 1/4/2026

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

- Επιβλέπων καθηγητής

Καρβέλης Πέτρος
Αναπληρωτής Καθηγητής

- Μέλος επιτροπής

Τζάλλας Αλέξανδρος
Αναπληρωτής Καθηγητής

- Μέλος επιτροπής

Γιαννακέας Νικόλαος
Αναπληρωτής Καθηγητής

© Σακελλάρης, Θεοχάρης, 2026.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Σακελλάρης, Θεοχάρης

Υπογραφή



- ΕΥΧΑΡΙΣΤΙΕΣ

Στην οικογένεια μου για την υποστήριξη που δείχνανε κατά την περίοδο διεξαγωγής της εργασίας. Καθώς και στο εκπαιδευτικό προσωπικό και την βοήθεια του στην εκπόνηση της εργασίας.

ΠΕΡΙΛΗΨΗ

Η παρακάτω εργασία έχει σαν σκοπό το να μαζέψει πληροφορίες πάνω στις παραλλαγές του αλγόριθμου SOM οποίος αρχικά προτάθηκε από τον Teuno Kohonen. Ο αλγόριθμος SOM είναι ένα δίκτυο χωρίς επίβλεψη και ξεκίνησέ σαν ένας αλγόριθμος ομαδοποίησής δεδομένων του οποίου η λειτουργία θυμίζει τον ανθρώπινο εγκέφαλο. Ο κυρίως τρόπος που το επιτυγχάνει αυτό είναι μέσω νευρώνων η οποίοι επηρεάζονται από τα δεδομένα με συνέπια να σχηματίζουν γειτονίες μεταξύ τους αναλόγως με την ομάδα δεδομένων που ανήκουν. Μια από της μεγάλες δυνάμεις του αλγορίθμου είναι η δυνατότητα να εμφανίσει τα αποτελέσματά του σε κατανοητά για τον χρήστη παραδείγματα. Ο συγκεκριμένος αλγόριθμος αρχικά γράφτηκε από τον Teuno Kohonen και με τον καιρό έχει αποκτήσει αρκετές παραλλαγές με σκοπό να ομαδοποίηση καλύτερα δεδομένα που έχουν να κάνουν με την ιατρική, γεωγραφία, εξόρυξη δεδομένων αλλά ακόμα και σχετικά μη αριθμητικά δεδομένα. Παρόλα αυτά ένα από τα κύρια ζητήματα των παραλλαγών του αλγόριθμου είναι η σταθερότητα του αλγορίθμου προς τα δεδομένα τα οποία εξετάζει αλλά και η υπολογιστική δύναμη που χρειάζεται, καθώς όμως και η ακρίβεια για τα δεδομένα τα οποία ομαδοποιούν κάτι το οποίο πολλές από της παραλλαγές έχουν σαν στόχο να βελτιώσουν. Σκοπός αυτής της εργασίας είναι η ομαδοποίηση αυτόν τον παραλλαγών, η επεξήγηση τους στο πως δουλεύουν αλλά και σε παραδείγματα χρήσης τους που χρησιμοποιήθηκαν. Καθώς και να δείξει της διαφορές μεταξύ του κανονικού som και μιας παραλλαγής του τον rsom μέσω σύγκρισης ομαδοποίησης τους σε διάφορα σετ δεδομένων.

ABSTRACT

The paper below has as target to collect information about the various variants on the SOM algorithm that was originally proposed by Teuno Kohonen. The SOM algorithm is a non supervised network and it begun as a clustering algorithm that was trying to mimic the human brain in the way it worked. The main way that it did that was via neurons that are affected by the data presented to them and that has as a consequence to form “neighborhoods” between them depending on the group of data that they match between them. One of the big strengths of the algorithm is its ability to show its results in an easy to understand format for the user. This particular algorithm was written by Teuno Kohonen and with time has gotten many variants that aim to cluster better all sorts kinds of data such as medicine, geography, data extraction and even non arithmetic data. Besides all that one of the main concerns of all variants is the stability of the algorithm towards the data that its working with and also the computing power it requires , as well as the accuracy of the data that it clusters something that most of the variants have as a target to improve upon the original algorithm. The aim of this paper is to cluster all these variants, to describe how they work and to show use cases where they were used at.

Keywords: Som, som variants, geo som ,arithmetic som, relational som

Περιεχόμενα

Contents

1 ΕΙΣΑΓΩΓΗ	13
1.1 Εισαγωγή στο που ανήκει ο αλγόριθμος SOM	13
1.2 Μη-εποπτευόμενη μάθηση	15
1.3 Σκοπός πτυχιακής εργασίας	19
1.4 διάρθρωση της εργασίας	19
2 Ο αλγόριθμος SOM	20
2.1 Πως λειτουργεί και η βασική ιδέα	20
2.2 Ο χάρτης του SOM	22
2.3 Γειτονίες νευρώνων	25
2.3.1 Συνάρτηση γειτνίασης	28
2.4 Ο αλγόριθμος του SOM	29
2.4.1 Batch αλγόριθμος	31
2.4.2 Umatrix	32
2.5 Εφαρμογές των αυτο-οργανόμενων χαρτών	33
3 Παραλλαγές SOM	35
3.1 Geo-enforced som	35
3.2 Spatial kangas som	35
3.3 GEO-SOM	36
3.4 Hierarchical soms	37
3.5 Σταθερή ανάθεση στον κανόνα χέσκες	38
3.6 SOFT TOPOGRAPHIC MAPPING –STM	38
3.7 Πιθανολογικά μοντέλα	39
3.7.1 Regularized EM	39
3.7.2 Variational EM	40

3.7.3 παραγωγική τοπογραφική χαρτογράφηση	40
3.8 Αλγόριθμος για ενδεχόμενα δεδομένα	40
3.9 Median SOM	41
3.10 Relational SOM.....	41
3.10.1 Fuzzy relational som	42
3.11 Kernel SOM	43
3.11.1 On-line kernel som	44
3.11.2 Bagged kernel som.....	45
4 Ανάλυση των δεδομένων και εξήγηση τους.....	46
4.1 iris dataset.....	46
4.2 Σετ υπολειμμάτων αμινοξέων	51
4.3 Παράδειγμα συνθετικού Δεδομένων ανομοιότητας δύο ομάδων	51
5 Αποτελέσματα	54
5.1 Λογισμικό που θα χρησιμοποιηθεί.....	54
5.2 Αποτελέσματα στο σετ δεδομένων iris	55
5.2.1 Μέγεθος χάρτη 5x5 normal som	55
5.2.2 Μέγεθος χάρτη 10x10 normal som	57
5.2.3 μέγεθος χάρτη 15x15 normal som.....	59
5.2.4 Μέγεθος χάρτη 10x10 normal som σε κυλινδρικό σχήμα	61
5.2.5 μέγεθος χάρτη 5x5 rsom σε επίπεδο σχήμα.....	64
5.2.6 μέγεθος 5x5 rsom σε κυλινδρικό σχήμα	67
5.2.7 Μέγεθος χάρτη 10x10 rsom.....	70
5.2.8 Μέγεθος 10x10 rsom σε κυλινδρικό σχήμα.....	72
5.2.9 Μέγεθος 14x14 rsom σε κυλινδρικό σχήμα.....	74
5.2.10 Μέγεθος χάρτη 14x14 rsom σε κυλινδρικό σχήμα	76
5.2.11 σύγκριση των καλύτερων αποτελεσμάτων.....	80
5.3 παράδειγμα σετ δεδομένων υπολειμμάτων αμινοξέων	82
5.3.1 Διαστάσεις χάρτη 4x4 επίπεδο σχήμα SOM	82
5.3.2 διαστάσεις χάρτη 5x5 σε επίπεδο σχήμα στον κανονικό som	84
5.3.3 Διαστάσεις 5x5 σε κυλινδρικό σχήμα στον κανονικό som	87
5.3.4 Διαστάσεις χάρτη 6x6 σε επίπεδο σχήμα στον κανονικό som	90

5.3.5 Διαστάσεις χάρτη 6x6 σε κυλινδρικό σχήμα σε κανονικό som	92
5.3.6 Διαστάσεις χάρτη 4x4 σε επίπεδο σχήμα σε rsom	95
5.3.7 Διαστάσεις χάρτη 4X4 σε κυλινδρικό σχήμα σε RSOM	97
5.3.8 Διαστάσεις χάρτη 5x5 σε επίπεδο σχήμα σε rsom	99
5.3.9 Διαστάσεις χάρτη 5x5 σε κυλινδρικό σχήμα σε rsom	101
5.3.10 Διαστάσεις χάρτη 6x6 σε επίπεδο σχήμα σε rsom	104
5.3.11 Διαστάσεις χάρτη 6x6 σε κυλινδρικό σχήμα σε rsom	106
5.3.12 Αποτελέσματα των παραδειγμάτων	108
5.4 Παράδειγμα dissimilarity data	111
5.4.1 Διαστάσεις χάρτη 5x5 σε επίπεδο σχήμα σε κανονικό som	111
5.4.2 Διαστάσεις χάρτη 5x5 σε κυλινδρικό σχήμα σε κανονικό som	112
5.4.3 Διαστάσεις χάρτη 10x10 σε επίπεδο σχήμα σε κανονικό som	114
5.4.4 Διαστάσεις χάρτη 10x10 σε κυλινδρικό σχήμα σε κανονικό som	115
5.4.5 Διαστάσεις χάρτη 17x17 σε επίπεδο σχήμα σε κανονικό som	116
5.4.6 Διαστάσεις χάρτη 17x17 σε κυλινδρικό σχήμα σε κανονικό som	118
5.4.7 Διαστάσεις χάρτη 5x5 σε επίπεδο σχήμα σε rsom	120
5.4.8 Διαστάσεις χάρτη 5x5 σε κυλινδρικό σχήμα σε rsom	121
5.4.9 Διαστάσεις χάρτη 10x10 σε επίπεδο σχήμα σε rsom	123
5.4.10 Διαστάσεις χάρτη 10x10 σε κυλινδρικό σχήμα σε rsom	124
5.4.11 Διαστάσεις χάρτη 17x17 σε επίπεδο σχήμα σε rsom	125
5.4.12 Διαστάσεις χάρτη 16x16 σε κυλινδρικό σχήμα σε rsom	126
5.4.13 Διαστάσεις χάρτη 17x17 σε κυλινδρικό σχήμα σε rsom	127
5.5.14 Σύγκριση των αποτελεσμάτων	128
6 Συμπεράσματα	131
• ΒΙΒΛΙΟΓΡΑΦΙΑ	132
appendix	138
appendix A κώδικας μετάτροπης iris	138
appendix B Κλάσεις που χρησιμοποιήθηκαν και τροποποιήσεις	139

1 ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή στο που ανήκει ο αλγόριθμος SOM

Ο αλγόριθμος SOM ανήκει σε ένα κλάδο της επιστήμης της πληροφορικής με όνομα μηχανική μάθηση . Ανήκει στο πεδίο της τεχνητής νοημοσύνης και ο βασικός σκοπός της είναι να αναγράφουν υπολογιστικά προγράμματα τα οποία μπορούν να μαθαίνουν. (47) Με την έννοια να “μαθαίνουν” μπορούμε να το εξηγήσουμε καλύτερα με το να πάρουμε ένα παράδειγμα μεταξύ μήλα και πορτοκάλια. Στο οποίο έχουμε σαν πρόβλημα να τα ξεχωρίσουμε μεταξύ τους. Αντί κάποιος προγραμματιστής να γράψει ένα πρόγραμμα στο οποίο εξηγεί ακριβώς και με μεγάλη ακρίβεια διάφορα παραδείγματα μεταξύ μήλων και πορτοκαλιών. Προγραμματίζει το πρόγραμμα ώστε να μάθει να ξεχωρίζει της διαφορές μεταξύ τους και να μην χρειάζεται μια μεγάλη λίστα παραδειγμάτων.(48) Πιο απλά το πρόγραμμα κάνει μια πρόβλεψη και αναλόγως την μάθηση και τα δεδομένα που έχουν χρησιμοποιηθεί για να τον εκπαιδεύσει αλλά και αν πολλές φορές η πρόβλεψη αυτή ήταν σωστή μαθαίνει να ξεχωρίζει τα δεδομένα. Η αρχή της μηχανικής μάθησης συνήθως συνδέεται με το όνομα του ψυχολόγου Frank Rosenblatt στο αεροναυτικό εργαστήριο του Κορνέλλ . Ο συγκεκριμένος επιστήμονας βασισμένος στις ιδέες και γνώσεις του ανθρώπινου νευρικού συστήματος δημιούργησε μαζί με μία ομάδα την μηχανή “perceptron” η οποία μπορούσε να αναγνωρίζει τα γράμματα της αλφάβητου. Η συγκεκριμένη μηχανή χρησιμοποιήθηκε σαν ένα πρωτότυπο για μοντέρνα συστήματα τεχνητής νοημοσύνης και το μοντέλο της μάθησης που χρησιμοποιούσε έμοιαζε αρκετά με μοντέλα μάθησης για ανθρώπους και ζώα. Η εξέλιξη αυτής της μηχανής έρχεται το 1972 από τον Novikoff ο οποίος καταφέρνει να αναπτύξει τον αλγόριθμο της μηχανής στο να καταλήγει την εκπαίδευση της σε ένα περιορισμένο αριθμό βημάτων μάθησης.(49) Έπειτα όμως το 1970 η επιστήμη παγώνει μιας και πολλές υποσχέσεις περί των δυνατοτήτων των μοντέλων δεν εκπληρώνονταν κάτι που οδήγησε και στο σταμάτημα της χρηματοδότησης της επιστήμης .Όμως πολύ επιστήμονες συνέχισαν χωρίς χρηματοδότηση τις έρευνες με την διαφορά να γίνεται το 1979 από τον Κουνιχίκο Φουκουσίμα(50). Αυτό που προτάθηκε ήταν ένα καινούργιο και μάλιστα από τα πρώτα, νευρωνικό δίκτυο το οποίο είχε σαν σκοπό να μπορεί να αναγνωρίζει πρόσωπα ανθρώπων. Αυτό το κατάφερε αναγνωρίζοντας την γεωμετρία του προσώπου κάποιου και του δόθηκε το όνομα “neocognitron”(51). Έπειτα το 1980-99 η επιστήμη πάλι ουσιαστικά παγώνει με τα πιο σημαντικά πράγματα εκείνη την περίοδο να είναι πάνω σε μια τεχνική ονομαστή “πίσω διάδοση” η αλλιώς “back propagation” η οποία γράφτηκε από τον Σέπλο Λινάινμαα είχε σαν λογική να χρησιμοποιηθούν τα λάθη που έγιναν κατά την εκπαίδευση στην εκπαίδευση. Το συγκεκριμένο πράγμα είναι κάτι που δεν χρησιμοποιήθηκε μέχρι και το 1985 όπου ο Rumelhart, Williams και Hinton την χρησιμοποίησαν και κατάφεραν να αποδώσουν κάποια ενδιαφέρων δεδομένα.(50) Έπειτα η εξέλιξη των υπολογιστών και καρτών γραφικών σε ταχύτητα βοήθησαν την επιστήμη να ανθίσει μιας και μπορούσαν αν υποστηριχθούν πιο βαριά μοντέλα από τα καινούργια συστήματα. Ένα από τα πράγματα τα οποία βοήθησαν πολύ στην εξέλιξη της μηχανικής μάθησης προς της εικόνες σίγουρα ήταν και το imagenet από την Φάιη-Φάιη Λι η οποία το 2019 έφτιαξε μια βάση δεδομένων με περισσότερες από 14 εκατομμύρια εικόνες η οποίες ήταν έτοιμες για εκπαίδευση από αλγορίθμους. Κάτι που ήταν δωρεάν και βοήθησε την επιστήμη να ανθήσει και που είναι διαθέσιμο ακόμα και σήμερα όχι μόνο για εκπαίδευση αλλά

και με εργασία για αξιολόγησης των αποτελεσμάτων των αλγορίθμων που μπορεί να το χρησιμοποιήσουν.(50)

Η μηχανική μάθηση είναι μια επιστήμη η οποία λόγω της φύσης της στηρίζεται αρκετά στην στατιστική. Αυτό συμβαίνει διότι τα αποτελέσματα δεν μπορούν να είναι μια απλοϊκή πρόβλεψη αντιθέτως θα πρέπει να είναι στατιστικώς σημαντικά. Η μηχανική μάθηση σαν επιστήμη έχει δει χρήσεις σε πολλά διαφορετικά σημεία εργασίας όπως για παράδειγμα σε εξόρυξη δεδομένων ,αναγνώρισης φωνής, σε παιχνιδομηχανές αλλά και σε εκπαίδευση κινήσεων ρομπότ.(47) Η αλγόριθμοι της μηχανικής μάθησης μπορούν να θεωρηθούν μη συμβατική σε σχέση με έναν κοινό αλγόριθμο. Αυτό προέρχεται από την φύση τους να αλλάζουν την αρχιτεκτονική τους αυτόματα αναλόγως με την εκπαίδευση που θα λάβουν ώστε να γίνουν καλύτερη στην διαχείριση της εργασίας που χρησιμοποιούνται. Ο αλγόριθμος έχει σαν σκοπό να βελτιστοποιήσει τον εαυτό του όχι μόνο για να διαχειρίζεται τα δεδομένα που του δίνονται για εκπαίδευση . Αλλά και να μπορεί να διαχειριστεί και άγνωστα δεδομένα τα οποία δεν έχει “ξαναδει”. Όταν αυτό επιτευχθεί με επιτυχία θεωρείται ότι ο αλγόριθμος είναι εκπαιδευμένος . Ένας ιδανικός αλγόριθμος μηχανικής μάθησης είναι αυτός ο οποίος μπορεί να μιμείται των τρόπο όπου ένας άνθρωπος ή άλλο ων μπορεί να σκέφτεται και να μαθαίνει από τα ερεθίσματα γύρω του. Να μπορεί όχι μόνο να μάθει από καινούργιες πληροφορίες αλλά και από τα λάθη του. Ο αλγόριθμος μπορεί να έχει πολλές παραμέτρους μάθησης η οποίες είναι αριθμητικές και ουσιαστικά πειράζοντας αυτές της αριθμητικές παραμέτρους καταφέρνει να “μαθαίνει” τα δεδομένα που του δίνονται. Αυτό μπορεί να γίνει και με πιο περίπλοκες αρχιτεκτονικές όπως για παράδειγμα να χρησιμοποιεί έναν ιστό με επίπεδα από τέτοιες μεταβλητές όπου επηρεάζουν η μια την άλλη με σκοπό να εντοπίζει καλύτερα μοτίβα στα δεδομένα και να μπορεί να μαντεύει πιο σωστά αποτελέσματα. (48)

Στο πεδίο της μηχανικής μάθησης υπάρχουν δύο κύριες και μια τρίτη κατηγορία που συνδυάζει της δύο προηγούμενες. Αυτές είναι η εποπτευόμενη μάθηση (supervised training) , η μη εποπτευόμενη μάθηση (unsupervised training) και μια τρίτη η μερικώς-εποπτευόμενη μάθηση(semi-supervised training). Ξεκινώντας με την εποπτευόμενη μάθηση ένα καλό παράδειγμα της είναι αυτό που δόθηκε στην αρχή του κεφαλαίου με τα μήλα και πορτοκάλια. Στην εποπτευόμενη μάθηση δίνουμε στον αλγόριθμο δεδομένα εκπαίδευσης στα οποία μαζί με τα δεδομένα βρίσκεται και η σωστή απάντηση για την κάθε καταχώρηση . Δηλαδή εάν δίνουμε στον αλγόριθμο ένα σετ δεδομένων με το χρώμα, σχήμα, βάρος, μέγεθος και άλλα χαρακτηριστικά για μήλα και πορτοκάλια θα του δίνουμε και ένα ακόμα το οποίο θα ήταν η απάντηση για το που ανήκει το κάθε παράδειγμα. Αυτό βασισμένο στην ικανότητα του μοντέλου να διαχειρίζεται κάθε μια από τις καταχωρήσεις στο σετ δεδομένων το κάνει να αλλάξει για να εξασφαλίσει ότι θα μπορεί να δουλεύει καλύτερα μετά από κάθε καταχώρηση ,με σκοπό να την αναγνωρίσει αν μια παρόμοια καταχώριση εμφανιζόταν ξανά αφού εκπαιδευτή. Φυσικά αυτό σημαίνει ότι ένα μοντέλο θα χρειαστεί πολλές καταχωρήσεις οι οποίες επηρεάζουν τις παραμέτρους του αλγόριθμου με σκοπό να μάθει την σωστή συμπεριφορά ώστε όταν του δοθεί ένα παράδειγμα που δεν έχει δει στα δεδομένα μάθησης του να καταφέρει να το αναγνωρίσει.(48)(47)

Η δεύτερη κατηγορία μηχανικής μάθησης είναι η μη εποπτευόμενη μάθηση. Όπου όπως και το όνομα συνιστά σε σχέση με την εποπτευόμενη μάθηση τα δεδομένα μάθησης δεν έχουν την

απάντηση στα δεδομένα μάθησης. Δηλαδή ο αλγόριθμος εκπαιδεύεται μόνο να αναγνωρίζει στα τυφλά μοτίβα στα δεδομένα. Για παράδειγμα αν αντί στο παράδειγμα με τα μήλα και πορτοκάλια δίνουμε στον αλγόριθμο όλα τα δεδομένα με πριν π.χ. χρώμα, βάρος, σχήμα, μέγεθος αλλά δεν δίνουμε την σωστή απάντηση ο αλγόριθμος θα έπρεπε να εκπαιδεύσει τον εαυτό του να ξεχωρίζει μεταξύ των δεδομένων χωρίς απαραίτητα να γνωρίζει τι είναι το καθένα αλλά απλώς να γνωρίζει ότι υπάρχουν διαφορετικές ομάδες. Ιδανικά ο αλγόριθμος θα μπορεί να αναγνωρίσει το πόσες ομάδες υπάρχουν στα δεδομένα και θα μπορεί να αναγνωρίσει και δεδομένα με συγκεκριμένες αλλαγές όπως για παράδειγμα, μια αλλαγή στο σχήμα από χτύπημα σε ένα πορτοκάλι ή μήλο.(48)(47)

Η τρίτη κατηγορία όπως προαναφέρθηκε είναι η μερικώς-εποπτευόμενη μάθηση η οποία ενώνει ουσιαστικά της δύο κατηγορίες. Δηλαδή δέχεται δεδομένα τα οποία σε μερικά δίνεται η απάντηση και σε μερικά όχι. Σκοπός αυτού του τρόπου μάθησης είναι να μιμηθεί πως στην φύση ένας οργανισμός μαθαίνει γύρω του μιας και πολλές φορές τα ερεθίσματα που δέχεται μπορεί να είναι τελείως άγνωστα. Γενικότερα υπάρχουν δύο πολύ σημαντικά πλεονεκτήματα που έχουν οι αλγόριθμοι της μηχανικής μάθησης. Μπορούν να κάνουν κοπιαστικές και επαναληπτικές εργασίες με μεγάλη αποδοτικότητα χωρίς να κουράζονται όπως ένας άνθρωπος. Επίσης πολλές φορές μπορούν να μάθουν δεδομένα τα οποία μπορεί να είναι όχι μόνο πιο περίπλοκα αλλά και να παρατηρήσουν πολύ πιο διακριτικά μοτίβα που μπορούν να κρύβονται σε δεδομένα που σε έναν άνθρωπο θα μπορούσαν να ξεφύγουν(48). Πιο συγκεκριμένα σε αυτήν την εργασία θα δοθεί βάση στην δεύτερη κατηγορία της μηχανικής μάθησης την μη εποπτευόμενη μάθηση. Για την οποία θα εξηγηθούν μερικά παραδείγματα παρακάτω καθώς και με πιο συγκεκριμένα τον αλγόριθμο που θα ασχοληθούμε τον SOM.

1.2 Μη-εποπτευόμενη μάθηση.

Συνεχίζοντας θα μιλήσουμε πιο συγκεκριμένα για την κατηγορία της μη-εποπτευόμενης μάθησης στην μηχανική μάθηση. Πιο συγκεκριμένα θα εξηγηθούν μερικές υποκατηγορίες της μη-εποπτευόμενης μάθησης. Καθώς επίσης και θα δοθούν κάποια παραδείγματα. Όπως προαναφέρθηκε Η μη-εποπτευόμενη μάθηση πρόκειται για αλγόριθμους όπου τα δεδομένα που τους δίνονται δεν συνοδεύονται με το που ανήκουν, της απαντήσεις δηλαδή στο σετ των δεδομένων. Αυτό φυσικά κάνει την διαχείριση των δεδομένων πιο δύσκολη και καθιστά το μόνο που μπορούμε να κάνουμε είναι να ομαδοποιήσουμε ή να οργανώσουμε την πληροφορία με κάποιο τρόπο. Για αυτό τον λόγο μια από τις κύριες υποκατηγορίες και χρήσεις είναι η ομαδοποίηση των δεδομένων. Ένας από τους πιο γνωστούς αλγόριθμους μη-εποπτευόμενης μάθησης ομαδοποίησης είναι ο k-means.

Ο k-means είναι ένας από τους πιο γνωστούς και απλούς αλγόριθμους ομαδοποίησης. Χρησιμοποιεί και λειτουργεί καλύτερα σε δεδομένα που σχηματίζουν σύνολα από ομάδες που έχουν δεδομένα κοντά το ένα με το άλλο. Στον συγκεκριμένο αλγόριθμο ο χρήστης επιλέγει έναν αριθμό k από ομάδες που ο θέλει ο αλγόριθμος να ομαδοποιήσει. Ο αλγόριθμος με την σειρά του δημιουργεί αντιπροσώπους με όνομα κερκοειδές. Η κερκοειδές έχουν όπως και το όνομα τους υποστηρίζει σκοπό να βρεθούν στο κέντρο των συνόλων ομάδων των δεδομένων μας. Αυτό συμβαίνει σταδιακά κατά την διάρκεια του αλγορίθμου αφού τα κερκοειδές τυχαίως

τοποθετηθούν στον χώρο με τα δεδομένα μας. Έπειτα τα κερκοειδές κουνιούνται με κάθε βήμα του αλγορίθμου προς το κέντρο των ομάδων των δεδομένων, κάτι που το οποίο μετριέται με κάποια νορμά αποστάσεις όπως η ευκλείδεια. Ο αλγόριθμός σταματάει όταν τα κερκοειδές σταματήσουν να κινούνται στον χώρο. Σκοπός του αλγορίθμου και το συνήθως καλύτερο αποτέλεσμα είναι να μειώσει την διασπορά των δεδομένων από τα κερκοειδή. Δηλαδή το άθροισμα των δεδομένων από κάθε ομάδα να είναι όσο πιο μικρό γίνεται.(52)(53). Αυτό φυσικά σημαίνει επίσης ότι το τελικό παράδειγμα του αλγορίθμου είναι πολύ εξαρτημένο από της αρχικές θέσης που δίνονται στα κερκοειδές.

Ένας ακόμα σημαντικός αλγόριθμος ομαδοποίησης είναι ο expectation-maximization (EM). Ο συγκεκριμένος αλγόριθμος μοιάζει με τον kmeans στο πως τελικά ορίζει της ομάδες. Σε σχέση με τον kmeans ο αλγόριθμος EM θεωρείται ένας πιο επιεικής αλγόριθμος ομαδοποίησης. Ο λόγος αυτός είναι στον τρόπο με τον οποίο ομαδοποιεί τα δεδομένα αφού σε σχέση με τον kmeans όπου ένα δεδομένων ανήκει 100% στην ομάδα με το κοντινότερο κέντρο ο EM αντιθέτως ακολουθά την πιθανότητα του κάθε δεδομένου να ανήκει σε μια ομάδα η οποία πιθανότητα δεν είναι ποτέ 1 ή 0 αλλά κάτι στο ενδιάμεσο. Αρχικά όπως και με τον kmeans ο χρήστης πρέπει να δώσει έναν αριθμό k ομάδων και έπειτα δίνεται πάλι μια αρχική τυχαία τιμή με κέντρα αυτών των ομάδων. Ο αλγόριθμος μετά υπολογίζει της πιθανότητες του κάθε δεδομένο να ανήκει σε μια ομάδα εκ των οποίων σε κάθε βήμα ανανεώνουν τα κέντρα τους για να μεγιστοποιήσουν αυτές τις πιθανότητες. Η επαναλήψεις για τις ανανεώσεις των κέντρων συνεχίζουν μέχρι να ικανοποιήσουν μια τιμή που έχει δώσει ο χρήστης η οποία δεν είναι παρά η σύγκριση της αλλαγής της πιθανότητας των κέντρων με το προηγούμενο βήμα. Αφού η ανανεώσεις τελειώσουν το κάθε δεδομένο αναθέτετε σε κάθε ομάδα βασισμένο με την πιθανότητα του να ανήκει σε μια ομάδα. Φυσικά όπως και με τον kmeans η αρχικές θέσεις των κέντρων και όμως του αριθμού των ομάδων παίζει μεγάλο ρόλο στην ποιότητα του τελικού αποτελέσματος.
(52)(54)

Ένας ακόμα αλγόριθμος ομαδοποίησης ο οποίος μάλιστα θα μπορούσε να θεωρηθεί και μια εξέλιξη των προηγούμενων δύο είναι ο DBSCAN. Ο dbscan είναι ένας αλγόριθμος ο οποίος έχει σχεδιαστεί με σκοπό να μπορεί να αντιμετωπίζει “θόρυβο” δηλαδή απομονωμένα στοιχεία δεδομένων και να μην τα συμπεριλαμβάνει στην ομαδοποίηση ενώ ταυτόχρονα ομαδοποιεί με βάση την πυκνότητα των δεδομένων στον χώρο αντί με απλώς τις αποστάσεις των δεδομένων στον χώρο. Κάτι που του επιτρέπει να ομαδοποιεί δεδομένα τα οποία πολλές φορές μπορεί να έχουν εμφολευμένες ομάδες αναμεσα τους αλλά χωρίς να ανήκουν στην ομάδα. Αρχικά ο χρήστης χρειάζεται να εισάγει δύο τιμές, αυτές η τιμές είναι το eps το οποίο είναι η μέγιστη απόσταση ενός σημείου από ένα άλλο και η δεύτερη είσοδος είναι η minpts η οποία είναι το ποσά πολλά δεδομένα θα θέλουμε να υπάρχουν μέσα στην απόσταση eps. Σημαντικό είναι ότι σε αυτόν τον αλγόριθμο δεν ορίζεται ο αριθμός των ομάδων K μιας και το κάνει αυτόματα ο ίδιος ο αλγόριθμος. Ο αλγόριθμος έπειτα είναι έτοιμος να ξεκινήσει. Αρχικά βρίσκονται όλα τα δεδομένα που πληρούν της δύο παραπάνω εισόδους eps και minpts δηλαδή όλα τα δεδομένα που έχουν ένα συγκεκριμένο αριθμό δεδομένων γύρω τους σε απόσταση eps αυτά τα δεδομένα ονομάζονται πυρήνες η αλλιώς στα αγγλικά (core points). Εάν ένα δεδομένο βρίσκεται μέσα στην εμβέλεια των πυρήνων αλλά δεν είναι πυρήνες τα ίδια ονομάζονται οριακά σημεία (border point). Όποιο δεδομένο είναι μόνο του και δεν έχει κάποιο άλλο δεδομένο στην εμβέλεια του αγνοείται και θεωρείται θόρυβος. Έπειτα επιλέγεται ένας τυχαίος πυρήνας και του δίνεται η πρώτη ομάδα όπου με την σειρά για κάθε άλλο πυρήνα στην εμβέλεια του επίσης του δίνεται η πρώτη ομάδα μέχρι όλοι η πυρήνες μεταξύ τους να ανήκουν στην ομάδα. Όταν όλοι η πυρήνες από μια ομάδα ανήκουν στην πρώτη ομάδα τότε επιλέγεται ένας καινούργιος ανενεργός πυρήνας και του δίνεται η δεύτερη ομάδα κάτι το οποίο επαναλαμβάνεται μέχρι να μην υπάρχουν άλλοι κενοί πυρήνες. Εάν όμως ένα οριακό σημείο βρίσκεται μέσα σε μια ομάδα αυτό δεν την εξαπλώνει αλλά απλώς ανήκει σε αυτήν την ομάδα. Σε περίπτωση που ένα οριακό σημείο βρίσκεται ανάμεσα σε δύο ομάδες τότε θα ανήκει στην ομάδα που δημιουργήθηκε πρώτη. Με αυτόν τον τρόπο τα δεδομένα ομαδοποιούνται με βάση την πυκνότητα των δεδομένων χωρίς να λαμβάνουν υπόψη απομονωμένα δεδομένα δηλαδή τον θόρυβο(47)(55).

Ένα πολύ σημαντικό κομμάτι της μη-εποπτευόμενης εκπαίδευσης είναι επίσης και το πως θα διαχειριστούμε δεδομένα τα οποία βρίσκονται σε μεγάλες διαστάσεις. Δηλαδή σαν άνθρωποι έχουμε συνηθίσει να αντιλαμβανόμαστε πράγματα στον χώρο σε μέγιστο τρεις διαστάσεις. Δηλαδή σε πλάτος βάθος και ύψος και εννοείται ότι αυτός είναι και ένας τρόπος να αναπαραστήσουμε τα δεδομένα μας για να τα κατανοήσουμε πιο εύκολα. Όμως τα δεδομένα που παίρνουμε πολλές φορές μπορεί να είναι σε 4-5-6 ή και παραπάνω διαστάσεις. Σε αυτές τις περιπτώσεις μια λύση είναι να προσπαθήσουμε να μειώσουμε της διαστάσεις των δεδομένων. Μια από της τεχνικές για την μείωση των διαστάσεων των δεδομένων είναι η λεγόμενη ανάλυση κυρίων συνιστωσών η αλλιώς στα αγγλικά “PCA”(56). Η ανάλυση κυρίων συνιστωσών είναι μια στατιστική γραμμική τεχνική μείωσης διαστάσεων. Ουσιαστικά αναλύει ένα πίνακα δεδομένων ο οποίος δείχνει παρατηρήσεις οι οποίες είναι συνδεδεμένες μεταξύ τους. Με σκοπό να εξαγάγει σημαντικές πληροφορίες σε καινούριες ορθογώνιες μεταβλητές με όνομα κυρίες συνιστώσες. Αυτό καταφέρνει να αναδεικνύει μοτίβα ομοιότητας στα δεδομένα και να μπορούμε να δείξουμε τα δεδομένα μας σε μειωμένες διαστάσεις στον χάρτη. Για μια πιο περιληπτική εξήγηση στο πως δουλεύει η συγκεκριμένη τεχνική μπορεί κάποιος να αναφερθεί στο (57).

Μια πάρα πολύ μεγάλη υπό ομάδα στην κατηγορία της μη εποπτευόμενης μάθησης και η οποία επίσης που θα μελετήσουμε παραπάνω είναι τα νευρωνικά δίκτυα. Τα νευρωνικά δίκτυα η αλλιώς τεχνητά νευρωνικά δίκτυα έχουν βάση του μοντέλου τους το πως λειτουργεί ο ανθρώπινος εγκέφαλος. Είναι ένας κατευθυνόμενος γράφος ο οποίος ουσιαστικά είναι ένα σύστημα επεξεργασίας της πληροφορίας που δέχεται αλλά και επηρεάζεται από αλγόριθμους η οποίοι αλγόριθμοι των προσπελαύνουν. Ο γράφος είναι ένα σύνολο από πολλούς κόμβους η οποίοι συνδέονται μεταξύ τους. Κάθε ένας από αυτούς τους κόμβους λειτουργεί ανεξάρτητα από τους άλλους και χρησιμοποιεί μόνο δεδομένα που δέχεται σαν είσοδο και επεξεργασμένα που δίνει σαν έξοδο τα οποία καθοδηγούνται από την επεξεργασία του και των τιμών του. Ένα νευρωνικό δίκτυο συνήθως περιλαμβάνει 3 η και παραπάνω επίπεδα εκ των οποίων το πρώτο είναι το πεδίο εισόδου με τους κόμβους του, το επίπεδο εξόδου με τους κόμβους του και ανάμεσα τους έχουμε τους εσωτερικούς κρυμμένους κόμβους η οποίοι μπορεί να είναι παραπάνω από ένα επίπεδο. Ουσιαστικά ένα δεδομένο ξεκινάει από το επίπεδο εισόδου και στο επίπεδο εξόδου έχουμε το ποια θα είναι η πρόβλεψη του νευρωνικού δικτύου. Τα νευρωνικά δίκτυα έχουν την δυνατότητα να μαθαίνουν από προηγούμενες εφαρμογές έτσι ώστε σε περίπτωση που δοθεί μια ανεπαρκής λύση να δίκτυο τροποποιείται για να παραχθεί μια καλύτερη λύση στο πρόβλημα την επόμενη φορά που θα του ζητηθεί. Όμως ένα από τα μεγαλύτερα μειονεκτήματα είναι ότι είναι δύσκολο να κατανοηθούν από τους τελικούς χρήστες καθώς και το ότι λειτουργούν μόνο με αριθμητικά δεδομένα. Το νευρωνικό δίκτυο που θα ασχοληθούμε σε αυτήν την εργασία είναι η αυτό-οργανώμενη χάρτες. Ο αυτό-οργανώμενος χάρτης γνωστός και ως SOM και χάρτες κοχόνεν δημιουργήθηκε από τον Φιλανδό ερευνητή Τούβο Κόχονεν. Πρόκειται για την μετάφραση του ονόματος του στα αγγλικά Self organising maps, αλλά η συντομογραφία SOM είναι ο πιο διαδεδομένος όρος με τον οποίο αναφέρονται στους χάρτες. Δημιουργήθηκαν όπως αναφέρθηκε προηγουμένως από τον Τούβο Κόχονεν το 1981- 1982 (26) με σκοπό να μπορούμε να αντλήσουμε πληροφορίες από διαφόρων σε διαστάσεις δεδομένα είτε αυτά είναι σε μικρές σε διαστάσεις όπως δισδιάστατα η τρισδιάστατα αλλά είτε είναι σε μεγαλύτερες η διαστάσεις όπου θα είχαμε πρόβλημα στην κατανόηση τους. Είναι ένας αλγόριθμος ο οποίος χρησιμοποιείται από αρκετούς ερευνητές παγκοσμίως για της δυνατότητες του σαν ένας αλγόριθμος ομαδοποίησης ο οποίος διαχειρίζεται μεγάλα σε διαστάσεις δεδομένα ο οποίος με τους καιρούς έχει περάσει αρκετές παραλλαγές και βελτιώσεις. Παρακάτω θα εξηγηθεί καλύτερα η βασική ιδέα του αλγορίθμου, πως λειτουργεί ο αλγόριθμος SOM και για μερικές χρήσεις του.(27)(5)

Ένας από τους λόγους να χρησιμοποιήσει κάποιος ένα χάρτη som είναι η δυνατότητα του να διαχειρίζεται δεδομένα μεγάλων διαστάσεων. Όχι όμως μόνο αυτό αλλά και να μπορούν να απλοποιήσουν μετά αυτά τα δεδομένα σε δυσδιάστατη μορφή χωρίς να μεταλλάσσουν την τοπολογία των δεδομένων. Η αυτό-οργανώμενη χάρτες έχουν δει χρήση σε πολλά πεδία όπως ιατρική ,οικονομική αλλά και σε γεωγραφικά πεδία. Ο μεγάλος αριθμός παραλλαγών τους για να διαχειρίζονται πολλές διαφορετικές μορφές και τύπους δεδομένων είναι επίσης ένα πολύ μεγάλο πλεονέκτημα για αυτά. Σε αυτήν την εργασία θα εξηγηθούν μερικές από αυτές της παραλλαγές και θα δειχθούν και παραδείγματα μεταξύ μιας παραλλαγής αλλά καθώς και με τον κλασσικό som για να δείξουμε πως διαχειρίζονται τα δεδομένα αλλά και τα αποτελέσματα τους.

1.3 Σκοπός πτυχιακής εργασίας

Η συγκεκριμένη πτυχιακή εργασία έχει πολλαπλούς σκοπούς να πετύχει. Αρχικά την εξήγηση του τι είναι ένα som ,το πως λειτουργεί και περιπτώσεις χρήσης τους με σκοπό την καλύτερη κατανόηση του αλγορίθμου και των αποτελεσμάτων του .Σκοπός φυσικά είναι και η καλύτερη κατανόηση του αλγορίθμου για τον αναγνώστη για την συνέχεια της εργασίας .

Δευτερος σκοπός της εργασίας είναι να εξερευνήσει και να βρει παραλλαγές πάνω στον αλγόριθμο του som και να εξηγήσει πως αυτές λειτουργούν αλλά και πιθανές η και παρών χρήσης των παραλλαγών αυτών .

Τρίτως σκοπός είναι να δείξει ένα παράδειγμα μεταξύ του κανονικού αλγορίθμου και μιας παραλλαγής του αλγορίθμου ονόματος RSOM. Στο οποίο παράδειγμα θα δειχθούν παραδείγματα μεταξύ διάφορον σε διαστάσεις δεδομένων με σκοπό την ομαδοποίηση τους αλλά και στην παρατήρηση της συμπεριφοράς των δύο αλγορίθμων, της δυσκολίες και σύγκριση της ποιότητας των δεδομένων από το παράδειγμα αυτό. Αυτό που αποσκοπείτε να δειχθεί είναι το πόσο καλύτερα είναι τα αποτελέσματα που περνούμε μεταξύ των δύο αλγόριθμων τους.

1.4 διάρθρωση της εργασίας

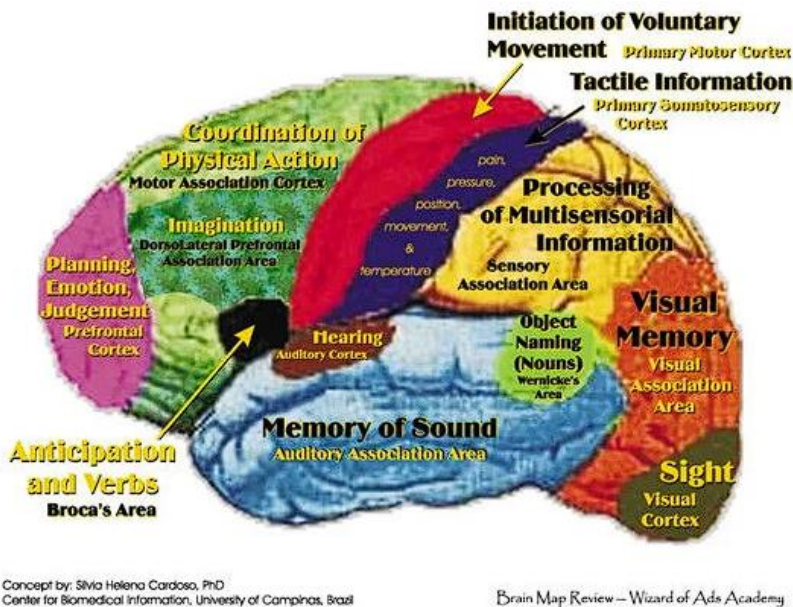
Στο πρώτο κεφάλαιο αναφέρεται ο σκοπός της πτυχιακής εργασίας καθώς και αμέσως μετά η διάρθρωση της. Στο δεύτερο κεφάλαιο παρουσιάζεται το τι είναι ένα δίκτυο som, το πως λειτουργεί αυτό μαζί με ορισμούς του που θα βοηθήσουν τον αναγνώστη να κατανοήσει το παρακάτω και κύριο κομμάτι της εργασίας αλλά και παραδείγματα εφαρμογής των δικτύων som που έχουν γίνει κατά καιρούς. Στο τρίτο κεφάλαιο εξηγούνται διάφορες παραλλαγές του som η οποίες έχουν δημιουργηθεί κατά καιρούς, μαζί με το πως λειτουργούν, τον σκοπό δημιουργίας τους , της εφαρμογές τους και παραδείγματα των αλγορίθμων τους. Στο τέταρτο κεφάλαιο δίνεται μια εξήγηση τον σετ δεδομένων που θα χρησιμοποιηθούν στο πρακτικό κομμάτι της εργασίας .

Στο πέμπτο κεφάλαιο παρουσιάζονται τα αποτελέσματα της υλοποίησης του πρακτικού κομματιού της εργασίας μεταξύ του κανονικού som και του Rsom χρησιμοποιώντας τα σετ που εξηγήθηκαν στον κεφάλαιο 4. Γίνεται αναφορά στο πακέτο λογισμικού που χρησιμοποιήθηκε χρησιμοποιώντας matlab το SOM toolbox καθώς και δείχνονται παραδείγματα τον αλγορίθμων που χρησιμοποιήθηκαν και περιγράφονται δυσκολίες που συναντήθηκαν. Στο κεφάλαιο 6 γίνεται αναφορά στα συμπεράσματα που έχουν προκύψει στην εργασία

2 Ο αλγόριθμος SOM

2.1 Πως λειτουργεί και η βασική ιδέα

Ο αυτό-οργανώμενος χάρτης SOM σαν ένα νευρωνικό δίκτυο έχει τις ρίζες του όπως και άλλα στην βιολογία. Όπως το όνομα συνιστά η ιδέα τους στηρίζεται στον εγκέφαλο των πλασμάτων ο οποίος είναι υπεύθυνος για οποιαδήποτε σκέψη απόφαση και αντίληψη των πλασμάτων με το περιβάλλον γύρω τους. Η ιδέα ονομάζεται εγκεφαλική χάρτες ή στα αγγλικά brain maps και φυσικά δεν είναι κάτι που έγινε μόνο από τους αυτό-οργανώμενους χάρτες μιας και υπήρχαν και προηγούμενα μοντέλα από το 1980. Στηρίζεται πάνω όπως προαναφέρθηκε στη βιολογία του εγκεφάλου η οποία είναι ότι διαφορετικά κομμάτια του εγκεφάλου είναι υπεύθυνα για διαφορετικά πράγματα.(εικόνα 1).



Concept by: Silvia Helena Cardoso, PhD
Center for Biomedical Information, University of Campinas, Brazil

Brain Map Review – Wizard of Ads Academy

(εικόνα 1)χάρτης εγκεφάλου χωρισμένος σε κομμάτια και ονομάζομενα για το τι είναι υπεύθυνα να επεξεργάζονται και να κάνουν από (28)

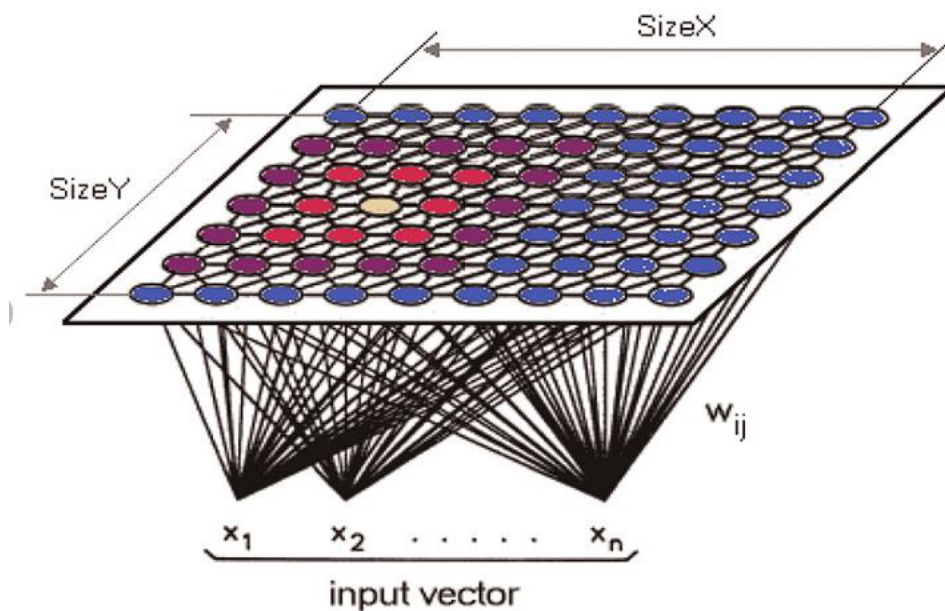
Δηλαδή ένα κομμάτι το οποίο είναι υπεύθυνο για την ακοή και κάποιο άλλο για ένα τελείως διαφορετικό ερέθισμα όπως η αφή αλλά μεταξύ τους συνεργάζονται για να χτίσουν μια εικόνα γύρω τους. Αρχικά είναι αξιότιμο να αναφερθεί ότι πιστευόταν ότι η συνάψεις του εγκεφάλου ή αλλιώς η χάρτες αυτή επιλεγόντουσαν τελείως γενετικά αλλά με τον καιρό βρέθηκε ότι η χάρτες στην πραγματικότητα δημιουργόντουσαν ανάλογα με της εμπειρίες του πλάσματος το οποίο διαβιούσε. (29)

Τα πρώτα μοντέλα τα οποία εμπνεύστηκαν από τα brain map είχαν δύο σημαντικό μειονεκτήματα το πρώτο ήταν ότι η τελική χάρτες η οποιοι παράγανε ήταν διαχωρισμένη. Δηλαδή δεν ήταν ένας ενιαίος χάρτης αλλά πολλά κομμάτια χαρτών διαχωρισμένα μεταξύ τους των οποίων τα αποτελέσματα είχαν τυχαία κατάταξη μεταξύ τους με αποτέλεσμα να μην υπάρχει κάποια σταθερή ενιαία κατάταξη μεταξύ τους. Το δεύτερο τους μειονέκτημα ήταν ότι δεν μπορούσαν να κλιμακωθούν σε μεγαλύτερες διαστάσεις για μεγαλύτερα δίκτυα , πάρα της μεγαλύτερης υπολογιστικής δύναμης που θα μπορούσε να διατηρηθεί. Πιστεύεται ότι το κομμάτι του παζλ το οποίο λειβόταν από τα πρώτα μοντέλα ήταν μια τιμή ελέγχου η οποία δεν θα επηρέαζε τα δεδομένα αλλά το ποσό παίζουν ρόλο τα δεδομένα που χρησιμοποιούνται. Αυτό στο εγkéφαλο γίνεται μέσω χημικών μορίων τα οποία συμπεριφέρονται σαν αγγελιοφόροι τα οποία γενικότερα στα μαθηματικά αγνοούνται αλλά θα μπορούσαν εάν κάποιος τα λάμβανε υπόψη να μεγαλώσουν τα μοντέλα και να μπορούν να κλιμακωθούν σε μεγαλύτερα δίκτυα.(29)

Ο πιο απλός τρόπος με τον οποίο εξηγείτε ο τρόπος με τον οποίο λειτουργούν η αυτό - οργανώμενη χάρτες είναι όπως παραπάνω εξηγήθηκε σαν ένας εγκέφαλος. Δηλαδή όπως ο ανθρώπινος εγκέφαλος θα συνδυάσει πολλά ερεθίσματα μαζί για να καταλάβει τι βρίσκεται στον χώρο γύρω του(30) . Για παράδειγμα έστω ότι κάποιος ακουμπούσε κάποιο σημείο του χεριού μας , το νευρικό μας σύστημα θα έστελνε σήμα στον εγκέφαλο μας για να μας ειδοποίηση για αυτό το ερέθισμα. Ταυτόχρονα εάν κάποιος ακουμπούσε ένα σημείο κοντινό σε αυτό πριν το σήμα που θα σταλθεί θα βρίσκεται στο ίδιο τμήμα του εγκέφαλου που χειρίζεται το ερέθισμα και κοντά στο προηγούμενο σήμα. Πιο ξεκάθαρα δυο νευρώνες η οποιοι βρίσκονται στο ίδιο τμήμα του εγκεφάλου και δηλαδή διαχειρίζονται τα ίδια ερεθίσματα πραγματοποιούν ομάδες μεταξύ τους όταν έχουν παρόμοια ερεθίσματα μεταξύ τους με αποτέλεσμα να σχηματίζουν ομάδες η αλλιώς χάρτες μεταξύ τους.(31) Όμως τι συμβαίνει όταν ένα ερέθισμα είναι πιο πολύπλοκο πως δηλαδή διαχειρίζεται κάτι ο οργανισμός μας όταν για παράδειγμα υπάρχει ακοή, όψη, και οσμή σε κάτι στον χώρο . Τότε θα χρησιμοποιηθούν όλα αυτά τα ερεθίσματα δηλαδή η ακοή, οσμή και όψη για να κατατοπίσει το ερέθισμα στον χώρο χρησιμοποιώντας τα σήματα που του δίνουν και συνδυάζοντας τα από τα διαφορετικά μέρη του εγκεφάλου που καταλήγουν. Έτσι με την ιδία λογική ο αλγόριθμος SOM χρησιμοποιεί πολλαπλούς νευρώνες η οποιοι στην “εκπαίδευση” τους θα μετατοπιστούν όσο πιο κοντά στα δεδομένα στον χώρο γίνεται για να δημιουργήσουν μεταξύ τους γειτονίες από νευρώνες μεταξύ τους. Ακολουθούν δηλαδή την ίδια λογική με τον εγκέφαλο χρησιμοποιούν τα δεδομένα ως ερεθίσματα στον χώρο και φτιάχνουν ομάδες η χάρτες με παρόμοια ερεθίσματα η αλλιώς δεδομένα έτσι ώστε να τα ομαδοποιήσουν.(32) Είναι επίσης αξιότιμο να αναφερθεί ότι τα ερεθίσματα που δέχεται ο εγκέφαλος και στη πραγματική ζωή δεν είναι ποτέ τέλεια η ακόμα και τα επιθυμητά . Πράγμα που όπως και η αυτό-οργανώμενη χάρτες έτσι και αυτός του καθιστά στο να λειτουργούν χωρίς επίβλεψη . Χρειάζεται δηλαδή να μάθει από το περιβάλλον όχι μόνο για τα ερεθίσματα αλλά και για την κατάλληλη απάντηση που αυτός θα δώσει σε αυτά.(30)

2.2 Ο χάρτης του SOM

Όπως προαναφέρθηκε λοιπόν το SOM είναι όπως ο εγκέφαλος, ένα νευρονικό δίκτυο το οποίο λειτουργεί χωρίς επίβλεψη. Αρχικά ο SOM πρόκειται για ένα τεχνητό πλέγμα από νευρώνες η οποίοι ανταγωνίζονται μεταξύ τους κατά την εκπαίδευσή τους. Το σχήμα του χάρτη η καλύτερα το μοντέλο θα μπορούσε κάποιος να το φανταστεί σαν δύο επίπεδα ένα είναι το επίπεδο εισόδου και το άλλο το επίπεδο εξόδου όπως φαίνεται και στην παρακάτω εικόνα(εικόνα 2).



(εικόνα 2) επίπεδο εισόδου και εξόδου σε διδιάστατο χάρτη SOM με $x_1, x_2, \dots, x_n$ να σηματοδοτούν τα δεδομένα και W_{ij} τους νευρώνες με i να είναι ένα διάνυσμα βάρους και j i τιμή δεδομένων που αναφερόμαστε.(33)

Σε αυτήν την εικόνα φαίνεται καθαρά το διάνυσμα εισόδου το οποίο βρίσκεται στο πρώτο επίπεδο. Στο συγκεκριμένο όπως και το όνομα συνιστά βρίσκονται τα δεδομένα τα οποία φορτώνουμε στον αυτό - οργανώμενο χάρτη . Τα δεδομένα μας μπορεί να είναι διαφόρων διαστάσεων και μεγέθους και αναπαρίστανται μαθηματικά με τον τύπο:

$$X = (x_1, x_2, \dots, x_n)$$

όπου n το διάνυσμα διάστασης και το δεδομένο που αναπαριστάται

Και φυσικά αναπαριστούν και εξηγούν τα δεδομένα μας και της ιδιότητες που αυτά μπορεί να

έχουν.

Μετά φαίνονται η συνάψεις από το επίπεδο εισόδου με τους νευρώνες του επιπέδου εξόδου. Η κάθε σύναψη ουσιαστικά εξηγείται από την τιμή w_{ij} η οποία βρίσκεται από δίπλα τους στην εικόνα και πρόκειται ουσιαστικά για το διάνυσμα βάρους ενός νευρώνα i προς μιας τιμής των δεδομένων μας j με w να είναι ο νευρώνας που αναφερόμαστε. Μαθηματικά η τιμή εξηγείται σαν:

$$w_i = [w_{i,1}, w_{i,2}, \dots, w_{i,n}]$$

Όπου n εννοείται το μέγεθος των δεδομένων μας στο επίπεδο εισόδου.(33)

Αυτές ουσιαστικά εκπροσωπούν της αποστάσεις η αλλιώς το πόσο μοιάζουν η νευρώνες στα δεδομένα που έχουμε .Η συγκεκριμένες αποστάσεις μπορούν να είναι διάφορες όπως για παράδειγμα:

ευκλιδιανές

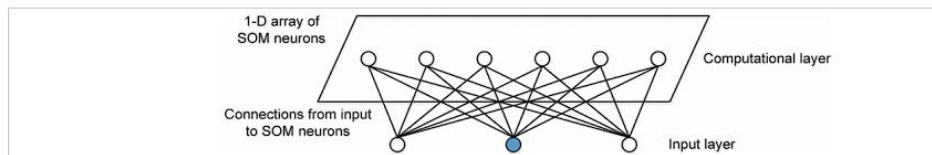
συσχέτισης

συνιμιτόνου κατεύθυνσης

απόστασης block η αλλιώς απόσταση μανχάταν(30)

Η πιο συνήθης απόσταση όμως είναι η ευκλιδιανή στο τετράγωνο.

Τέλος είναι το τελευταίο επίπεδο και αυτό που ενδιαφέρει τον χρήστη το επίπεδο εξόδου στο οποίο βρίσκονται η νευρώνες. Το συγκεκριμένο επίπεδο όπως προαναφέρθηκε πρόκειται για ένα πλέγμα από νευρώνες το οποίο μπορεί να είναι διάφορον διαστάσεων. Όπως μονοδιάστατο , δισδιάστατη αλλά και σε τρεις διαστάσεις αν και δεν συνηθίζονται. Παραπάνω φαίνεται ένα δυσδιάστατο επίπεδο εξόδου (εικόνα 2) και παρακάτω στην (εικόνα 3) ένα μονοδιάστατο δίκτυο SOM.



(εικόνα 3) μονοδιάστατο δίκτυο SOM όπου στο κάτω μέρος βλέπουμε δεδομένα στο επίπεδο εισόδου να συνδέονται με συνάψεις με τους νευρώνες στο επίπεδο εξόδου(30)

Ουσιαστικά αυτό που παίρνουμε στο επίπεδο εξόδου είναι η τοπολογία του χάρτη νευρώνων που εκπαιδεύσαμε στα δεδομένα. Η οποία μας δείχνει την ομοιότητα των νευρώνων μας με τα δεδομένα . Η ομοιότητα ή απόσταση των νευρώνων παίρνεται από τον τρόπο εκπαίδευσης του SOM. Δηλαδή το πως η νευρώνες σχηματίζουν την τοπολογία του χάρτη. Ένας απλός τρόπος για να εξηγήσει κάποιος τον τρόπο εκπαίδευσης του SOM είναι ο εξής.

Αρχικά οι νευρώνες τοποθετούνται σε τυχαία σημεία στον χάρτη η αλλιώς επίπεδο εξόδου. Έπειτα ανταγωνίζονται μεταξύ τους στο να “μοιάσουν” όσο πιο πολύ μπορούν στα δεδομένα στο επίπεδο εισόδου ο νευρώνας που μοιάζει περισσότερο στα δεδομένα ονομάζεται BMU από της αγγλικές λέξης best matching unit η αλλιώς νικητής νευρώνας.

Κατόπιν η νευρώνες συνεργάζονται μεταξύ τους για να φτιάξουν ομάδες η αλλιώς γειτονίες με βάση τον νικητή νευρώνα .(30)

Αυτός είναι ένας απλός τρόπος να εξηγήσει κάποιος το πως λειτουργεί ο SOM και το πως φτάνει στα αποτελέσματα που φτάνει.

2.3 Γειτονίες νευρώνων.

Όπως έχει αναφερθεί αρκετές φορές ο πιο απλός τρόπος να εξήγηση κάποιος το τι είναι μια γειτονιά νευρώνων είναι ο εξής. Μια γειτονιά από νευρώνες είναι μια ομάδα από νευρώνες η οποίοι μοιάζουν περισσότερο σε ένα δεδομένο X. Η συγκεκριμένη γειτονιά από νευρώνες είναι κάτι που προσδιορίζεται από τον χρήστη για το πόσο μεγάλη θα είναι αυτή και δεν υπάρχει κάποια συγκεκριμένη η απόλυτη απόσταση μιας και η ποιότητα του τελικού πεδίου εξόδου διαφέρει ανάλογα με τα δεδομένα μας και της μεταβλητές που θα της δώσουμε. Πέρα από αυτό φυσικά κάποιος θα αναρωτιέται το πως μοιάζει μία γειτονιά νευρώνων και τη σχήμα μπορεί αυτή να πάρει.

Τα σχήματα της γειτονίας εξαρτώνται φυσικά από δύο σημαντικούς παράγοντες 1 η ακτίνα της γειτονίας και 2 η διάστασης του δικτύου SOM . Όπως αναφέρθηκε προηγουμένως ο τοπογραφικός χάρτης των νευρώνων μας μπορεί να είναι από μια έως και σε τρεις διαστάσεις. Αυτό σημαίνει ότι το σχήμα των γειτονιών των νευρώνων που θα έχουμε σαν αποτέλεσμα από το δίκτυο θα μπορεί να είναι ανάμεσα σε :

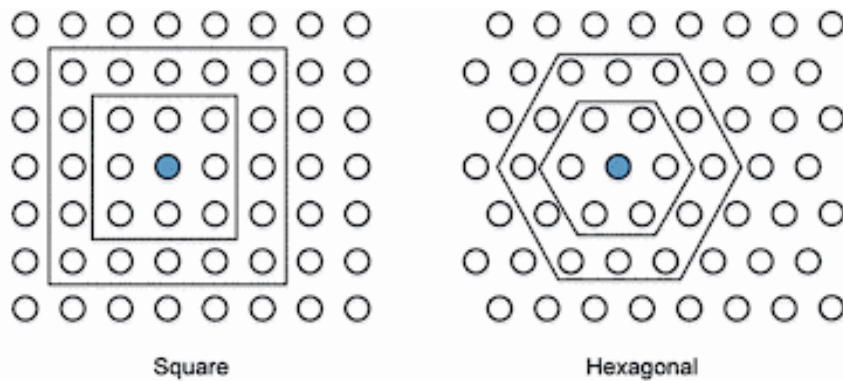
Ένα γραμμικό σχήμα το οποίο θα αναπαριστούσε έναν μονοδιάστατο χάρτη SOM συγκεκριμένα στην εικόνα παραπάνω βλέπουμε και μια γειτονιά ακτίνας 1 και 2 με τον μπλε νευρώνα να είναι ο νικητής νευρώνας.(εικόνα 4)



Linear

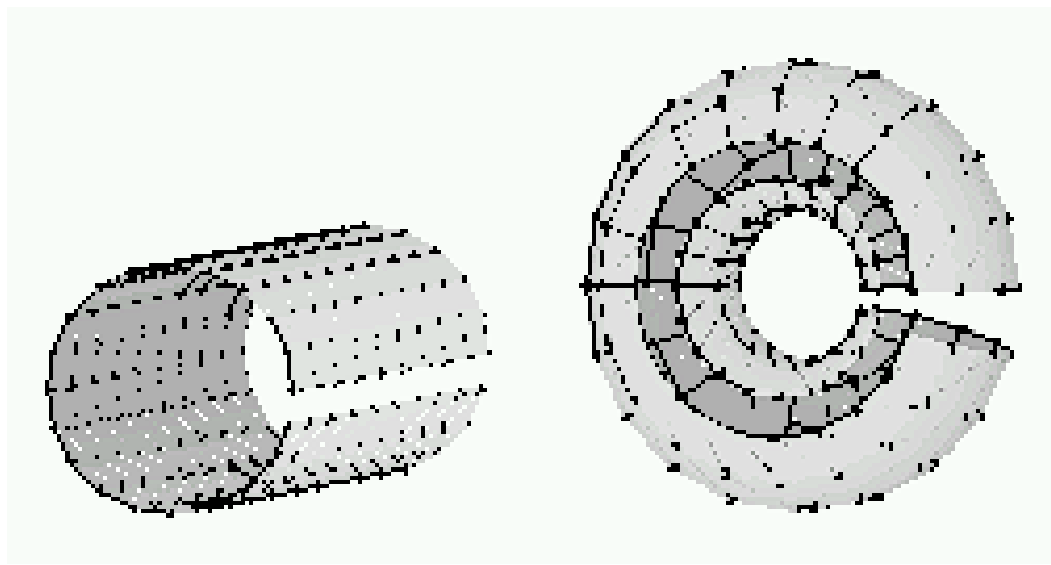
(εικόνα 4). μονοδιάστατο πλέγμα τοπολογίας SOM όπου με μπλε σηματοδοτείται ο νικητής νευρώνας ο οποίος έχει δύο παραδείγματα γειτονιών , στο πρώτο εσωτερικό όριο σηματοδοτείται μια ακτίνα γειτονίας 1 ενώ στο δεύτερο εξωτερικό όριο μια ακτίνα γειτονίας 2(30)

Στην πιο συνηθισμένη τοπολογία δύο διαστάσεων από της οποίες χρησιμοποιείτε είτε σε εξαγωνικό σχήμα αλλά και τετραγωνικό σχήμα με γειτονίες ακτίνας 1 και 2 και με τον μπλε νευρώνα ως τον νικητή νευρώνα (εικόνα 5).

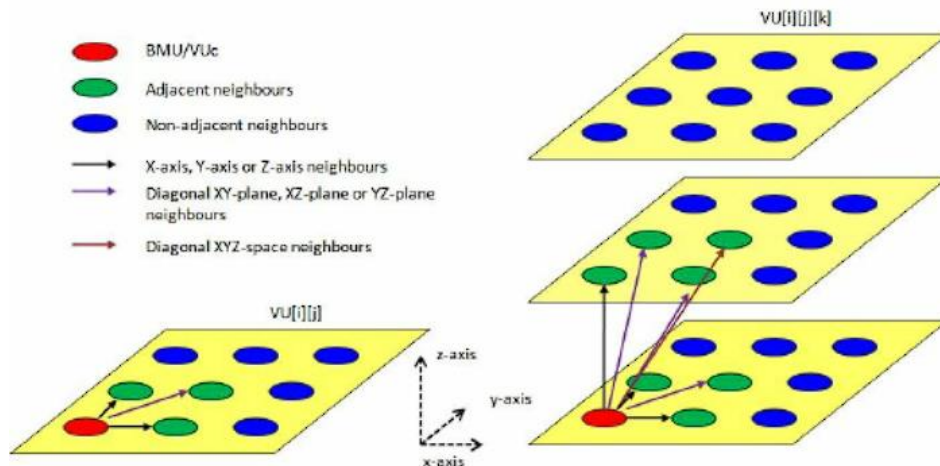


(εικόνα 5) δυοδιάστατο πλέγμα τοπολογίας SOM με δύο διαφορετικές τοπολογίες χάρτη , Τετραγωνική τοπολογία στα αριστερά και εξαγωνική τοπολογία στα δεξιά. Επίσης δείχνονται παραδείγματα γειτονιών όπου με μπλε σηματοδοτείται ο νικητής νευρώνας ο οποίος έχει δύο παραδείγματα γειτονιών , στο πρώτο εσωτερικό όριο σηματοδοτείται μια ακτίνα γειτονίας 1 ενώ στο δεύτερο εξωτερικό όριο μια ακτίνα γειτονίας 2 (30)

Αλλά και σε τρεις διαστάσεις σε σχήμα κυλίνδρου ή ενός τοροειδούς σχήμα (εικόνα 6) αν και η συγκεκριμένη τοπολογία δεν συνηθίζεται και το δυσδιάστατο πλέγμα ήδη λειτουργεί καλά για σετ δεδομένων τα οποία είναι σε μεγάλες διαστάσεις. Στην (εικόνα 7) επίσης μπορούμε να δούμε λίγο καλύτερα πως υπολογίζονται η διάσταση στον τρισδιάστατο χώρο.



(Εικόνα 6) τρεις διαστάσεις πλέγματος τοπολογίας του τελικού χάρτη SOM, σχήμα κύλινδρου SOM στα αριστερά και τοροηδές στα δεξιά (34)



Εικόνα 7. 7 τριδιάστατο πλέγμα χάρτη som, δείχνεται δίκτυο των νευρώνων του som σε τρεις διαστάσεις όπου με κόκκινο σηματοδοτούμε τον νικητή νευρώνα, με πράσινο τους διπλανούς νευρώνες και με μπλε τους πιο απομακρυσμένους νευρώνες σε ακτίνα 1 από τον νικητή νευρώνα. Στα αριστερά βλέπουμε μόνο ένα επίπεδο του πλέγματος ενώ στα δεξιά βλέπουμε για το ίδιο παράδειγμα 3 επίπεδα του χάρτη. Με μαύρες γραμμές σηματοδοτούνται οι γείτονες που είναι μόνο σε μια διάσταση μακριά, με μοβ σε 2 διαστάσεις μακριά και με κόκκινο σε 3 διαστάσεις μακριά. (35)

2.3.1 Συνάρτηση γειτονιάσης

Μεγάλο ρόλο στην οργάνωση των γειτονιών για τους νευρώνες παίζει επίσης η συνάρτηση γειτονιάσης. Είναι ουσιαστικά ο τρόπος με τον οποίο η νευρώνες αλληλοεπιδρούν μεταξύ τους στο πλέγμα και είναι ο κεντρικός ρόλος του πως διαμορφώνεται η τοπολογία του χάρτη στο πλέγμα. Μάλιστα χωρίς αυτήν ο αλγόριθμος δεν είναι πάρα ένας απλός σταδιακά αυξανόμενος kmeans αλγόριθμος. (36) Αυτό που προσπαθούμε να πέτυχουμε με την συγκεκριμένη συνάρτηση είναι μια συνάρτηση που σταδιακά θα καταλήξει στο 0 έτσι ώστε να σταθεροποιήσουμε την κίνηση των νευρώνων στον χώρο ώστε να σχηματίσουν τις τελικές ομάδες. Μια από τις πιο συνηθισμένες συνάρτησης γειτονιάσης είναι η συνάρτηση γκάους. (36)

Είναι σημαντικό επίσης να αναφερθεί ότι ο αλγόριθμος είναι αρκετά ισχυρός στο θέμα της συνάρτησης γειτνίασης και γενικότερα τα αποτελέσματα τα οποία θα εμφανίσει δεν διαφέρουν πολύ μεταξύ διαφορετικών αλγορίθμων. Φυσικά αυτό εννοείται ότι προϋποθέτει ότι η μεταβλητή μάθησης μηδενίζεται στο τέλος της εκπαίδευσης του αλγορίθμου αλλά και ότι η ακτίνα είναι μεγαλύτερη το 0 .Μίας και αν η ακτίνα είναι 0 η μονή γειτονιά που μπορεί να σχηματιστεί είναι μεταξύ των ίδιων των νευρώνων.

2.4 Ο αλγόριθμος του SOM

Ξέροντας λοιπόν τι είναι ο SOM , η λογική του και πως δουλεύει είναι η ώρα να εξηγηθεί το πως δουλεύει ο αλγόριθμος του SOM . Ποιά είναι τα βήματα που παίρνει για να φτάσει στο επιθυμητό αποτέλεσμα. Αλλά και το πως λειτουργεί ο batch αλγόριθμος που παρόλο ότι θεωρείται μια παραλλαγή του αλγορίθμου είναι μια από της πιο σημαντικές τους η οποία είναι η ρίζα σε πολλές άλλες.

Πριν ξεκινήσει κάποιος την εκπαίδευση του αλγορίθμου πρέπει φυσικά όπως αναφέρθηκε προηγουμένως να οριστούν η αρχικές θέσης των νευρώνων πάνω στον χάρτη .Υπάρχουν φυσικά διάφοροι τρόποι και λογικές στο πως κάποιος μπορεί να το κάνει αυτό αλλά η πιο διαδεδομένες είναι η εξής:

- 1:Να δοθούν τελείως τυχαίες τιμές η οποίες δεν έχουν καμία σχέση δηλαδή είναι τελείως ανεξάρτητες από τα δεδομένα εισόδου τα οποία δίνουμε στον χάρτη.
- 2:Η χρήση τυχαίων σημείων από τα δεδομένα μας για να ορίσουμε της αρχικές τιμές των νευρώνων
- 3:Μια αρχικοποίηση των τιμών των αλγορίθμων η οποία προσπαθεί να μιμηθεί την κατανομή των δεδομένων μας στον χώρο.

Είναι άξιο να αναφερθεί ότι όπως και η συνάρτηση γειτνίασης ο αλγόριθμος είναι γέρος και τα αποτελέσματα μας δεν θα επηρεαστούν δραματικά από την επιλογή του πως θα διαμοιραστούν η νευρώνες στον χώρο.

Ξεκινώντας με τον κλασσικό αλγόριθμο, γνωστός και ως sequential training όπου στα ελληνικά μεταφράζεται ως διαδοχική εκπαίδευση και είναι ο πρώτος αλγόριθμος εκπαίδευσης για τον SOM. Αρχικά πρέπει να επιλεχτεί το μέγεθος του χάρτη των νευρώνων που θα χρησιμοποιήσουμε. Όπως προαναφέρθηκε αυτό είναι κάτι του οποίου η επιλογή στηρίζεται πάνω στον χρήστη .Έπειτα πρέπει να γίνει η επιλογή των βαρών που θα έχουν η νευρώνες .Αυτό δεν είναι παρά η αρχικές θέσης των νευρώνων η οποίοι όπως προαναφέρθηκε μπορούν να διαμοιραστούν στον χώρο με διάφορους τρόπους . Για το παράδειγμα αυτό θα θεωρηθεί ότι το μίσημα των τιμών είναι με τυχαίες τιμές ανεξάρτητες από το πλήθος των τιμών εισόδου. Έπειτα πρέπει να επιλεχτεί η συνάρτηση γειτνίασης η οποία θα χρησιμοποιήσουμε .Όπως

είπαμε προηγουμένως η συνάρτηση γειτωνίας είναι υπεύθυνη για την αυτοοργάνωση του χάρτη των νευρώνων και είναι μια τιμή μεταξύ 0 και 1 όπου 1 είναι τιμές μακριά μεταξύ του νευρώνα και των δεδομένων και πιο κοντά στο 0 για τιμές που είναι κοντά μεταξύ τους.

Τότε τα βήματα του κλασσικού αλγόριθμου για τον SOM θα είναι:

- 1: επιλογή ενός τυχαίου δεδομένο από το σετ δεδομένων
- 2: Υπολογισμός όλων το αποστάσεων μεταξύ του δεδομένου και όλων των νευρώνων για την εύρεση του νικητή νευρώνα
- 3: Ανανέωση όλων των αποστάσεων των νευρώνων
- 4: επανάληψη των βημάτων μέχρι τα κριτήρια που θέσαμε να ικανοποιηθούν συνήθως τα κριτήρια αυτά είναι ένας συγκεκριμένος αριθμός επαναλήψεων που έχει επιλεγεί. Με μεγαλύτερους αριθμούς επαναλήψεων να μας δίνουν καλύτερα αποτελέσματα μεν αλλά να χρειάζονται περισσότερο χρόνο εννοείται και φυσικά η ποιότητα του αλγορίθμου ακολούθη το σχήμα μιας εκθετικής μείωσης πράγμα που κάνει της πολλές επαναλήψεις μη αρκετά κερδοφόρες από έναν αριθμό και μετά. Επίσης πρέπει να φροντίσουμε ότι στις επαναλήψεις αυτές η μεταβλητή μάθησης και η ακτίνα γειτωνίας μειώνονται προς το 0 για να έχουμε ένα πιο σταθερό χάρτη με λιγότερα λάθη και για να είμαστε σίγουροι ότι ο χάρτης θα συμπίψει μεταξύ του.

Όπως επίσης προηγουμένως αναφέρθηκε η απόσταση η οποία χρησιμοποιείτε είναι ευκλιδική στο τετράγωνο αλλά δεν είναι ο κανόνας να χρησιμοποιηθεί μόνο αυτή.

Αξιοτιμο επίσης να αναφερθεί ότι κατά την διάρκεια της εκπαίδευσης μπορούν να παρατηρηθούν δυο κεντρικές “φάσης” η οποίες το δίκτυο ακολουθεί.

Η πρώτη από αυτές είναι η φάση της αυτό-οργάνωσης η οποία βρίσκεται στην αρχή της εκπαίδευσης. Η συγκεκριμένη φάση γίνεται όταν ακόμα η μεταβλητές η οποίες οργανώνουν των χάρτη έχουν μεγάλες τιμές δηλαδή στην αρχή της εκπαίδευσης και είναι η στιγμή οπου ο χάρτης παίρνει το σχήμα των δεδομένων μας ενώ ταυτόχρονα μειώνεται η ακτίνα γειτωνίας και η μεταβλητή εκπαίδευσης.

Η δεύτερη από αυτές είναι η φάση της σύμπτυξης του χάρτη η αλλιώς της ρύθμισης του χάρτη για ακρίβεια . Η συγκεκριμένη φάση βρίσκεται προς το τέλος της εκπαίδευσης και πρόκειται για της μικρότερες κινήσεις των νευρώνων στον χάρτη με σκοπό να βελτιώσουν το σχήμα που έχουν πάρει με τα δεδομένα που είχαν . Συνηθίζεται επίσης σε αυτήν την φάση η ακτίνα γειτωνίας και μεταβλητή εκπαίδευσης να μένουν σταθερές σαν τιμές μιας και ήδη είναι πολύ μικρές. Ο συγκεκριμένος αλγόριθμος είναι επίσης κατάλληλος για να τρέξει σε ρύθμιση on-line.

(36)(39)(38)(37)(30)

2.4.1 Batch αλγόριθμος

Ο batch αλγόριθμος είναι μια πολύ σημαντική παραλλαγή του κανονικού αλγορίθμου.

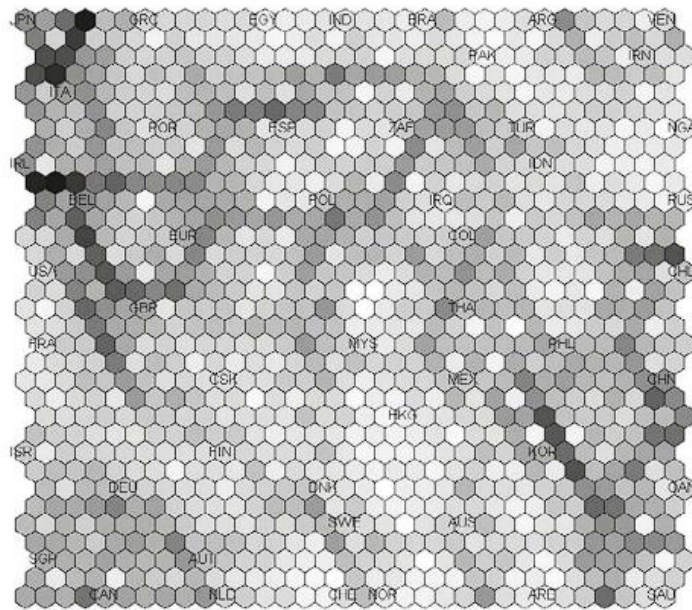
Θεωρείται μια από της πιο σημαντικές αλλαγές του μιας και λήγει σημαντικά προβλήματα στην ταχύτητα του αλγορίθμου κατά την εκπαίδευση αλλά και στην σταθερότητα των αποτελεσμάτων του. Για παράδειγμα ένα σημαντικό πρόβλημα του κλασσικού αλγορίθμου είναι ότι λόγω του ότι περνάμε 1 την φορά τα δεδομένα μας καθιστά τον αλγόριθμο να γίνεται δραματικά πιο αργός αναλόγως με το μέγεθος των δεδομένων μας. Επίσης σε περίπτωση που ένα διαφορετικό δεδομένο επιλεχτεί αν τρέξουμε τον αλγόριθμο δυο διαφορετικές φορές τότε θα έχουμε διαφορετικούς χάρτες για τα ίδια δεδομένα. Αυτά τα βασικά προβλήματα έρχεται να λύσει ο batch αλγόριθμος χρησιμοποιώντας μια διαφορετική λογική για την εκπαίδευση των δεδομένων. Η διαφορά μεταξύ των δυο αλγορίθμων είναι στο ότι στον batch som αντί να ανανεώνουμε τα βάρη του κάθε νευρώνα για κάθε ένα διαφορετικό δεδομένο 1 την φορά . Τώρα βάζουμε όλα τα δεδομένα από το σετ δεδομένων και υπολογίζουμε των νικητή νευρώνα με όλα τα δεδομένα ταυτόχρονα και έπειτα γίνεται οποιαδήποτε ανανέωση του χάρτη νευρώνων και των βαρών τους. Έτσι καταφέρνουμε μια πιο σταθερή μορφή του χάρτη και πιο γρήγορη όμως λόγω του ότι χρειαζόμαστε είδη έτυμο το σετ δεδομένων ο συγκεκριμένος αλγόριθμος δεν είναι κατάλληλος για να τρέξει σε ρύθμιση on-line. Με λίγα λόγια η διαφοράς μεταξύ των δυο αλγορίθμων βήμα προς βήμα είναι:

- 1: αντί για επιλογή ενός δεδομένου από το σετ ,φόρτωση όλων των δεδομένων ταυτόχρονα
- 2: αντί για εύρεση ενός νικητή νευρώνα, εύρεση όλων των νικητών νευρώνων στα δεδομένα μας
- 3:αντί για ανανέωση νευρώνων για ένα δεδομένο , ανανέωση νευρώνων για όλα τα δεδομένα
- 4:επανάληψη των βημάτων 2 και 3 μέχρι τα κριτήρια για σταμάτημα που θέσαμε να ικανοποιηθούν.

Η αλγόριθμοι μεταξύ τους φυσικά δεν έχουν μεγάλες διαφορές αλλά είναι πιο πολύ μια διαφορετική λογική στο πως διαχειρίζονται τα δεδομένα που τους δίνονται. Σε σχέση με τον κανονικό αλγόριθμο ο batch som θεωρείται μια ντετερμινιστική προσέγγιση στο πρόβλημα και θεωρείται μια καλύτερη επιλογή σε βιομηχανικές χρήσης λόγω της σταθερότητας των αποτελεσμάτων που δίνει. Βέβαια σε σχέση με τον κανονικό SOM ο οποίος είναι πιο γερός στο πως θα γίνει η αρχική διαμοίραση των νευρώνων ο batch SOM είναι πιο επιρρεπής σε αυτήν.(5)(30)

2.4.2 Umatrix

Είναι σημαντικό επίσης να μιλήσει κάποιος για έναν από του πιο διαδεδομένους τρόπους αναπαράστασης των αποτελεσμάτων μας με ένα από τα πιο δυνατά εργαλεία , τα umatrices η αλλιώς umatrix. Το umatrix είναι μια μέθοδος για εικονοποίηση των αποτελεσμάτων από τον Άλφρεντ ουλτς και Kraaijveld et al. Σκοπός τους είναι να δείξουν της απόστασης μεταξύ των νευρώνων σε ένα δισδιάστατο χάρτη προσθέτοντας παραπάνω θέσης η οποίες δεν επηρεάζουν τους νευρώνες του χάρτη για παράδειγμα εάν ένας χάρτης νευρώνων έχει διαστάσεις 18x17 το umatrix θα έχει διάσταση 35x33 τους νευρώνες και την εκπαίδευση τους. Όπως και με το επίπεδο εξόδου εμφανίζονται σε εξάγωνα αλλά και κυβικές εμφανίσεις και είναι ένας από τους καλύτερους τρόπους να δείξει κάποιος δεδομένα μεγάλων διαστάσεων σε κατανοητή προς τον χρήστη μορφή. Παρακάτω (εικόνα 8) είναι ένα παράδειγμα ενός umatrix το οποίο δείχνει την οικονομική κατάσταση 50 χωρών όπου χρωματισμένα με λευκό είναι η κοντινότερες αποστάσεις μεταξύ νευρώνων και με μαύρο η μακρύτερες.(37)



(εικόνα 8) Σχεδιάγραμμα umatrix ενός χάρτη som με τις οικονομικές καταστάσεις διαφορετικών 50 χωρών (37)

2.5 Εφαρμογές των αυτο-οργανόμενων χαρτών

Η αυτό-οργανώμενη χάρτες έχουν αρκετές χρήσεις κυρίως σαν εξαιρετική αλγόριθμοί ομαδοποίησης . Λόγω του ότι είναι αρκετά δυνατή αλγόριθμοι μερικές εφαρμογές που ο ίδιος ο kohonen είχε προτείνει ήταν σε χρήσης για μηχανική όραση, ανάλυση εικόνας, ανάλυση ομιλίας, στην ρομποτική, χημεία, φυσική , στην ιατρική αλλά και φυσικά για την επεξεργασία δεδομένων προβλημάτων τεχνητής νοημοσύνης και φυσικά σε αλλά πολλά. Πολλές από αυτές της εφαρμογές έχουν ήδη δει αρκετή χρήση και μερικές από αυτές αναγράφονται παρακάτω.

Μια από της πρώτες εφαρμογές των χαρτών kohonen ήταν σε φωνητικούς χάρτες. Η φωνητική χάρτες χρησιμοποιούνται για αναγνώριση ομιλίας το οποίο γίνεται κυρίως με την αναγνώριση φωνημάτων. Χρησιμοποιείτε ένας φωνητικός χάρτης σαν αυτόν που είχε δημιουργήσει ο Τούβο Κόχονεν για το φιλανδικό φωνητικό σύστημα. Για την αναπαράσταση του χρησιμοποιήθηκαν βραχυχρόνια φάσματα συχνοτήτων τα οποία υπολογιζόντουσαν με ταχύ μετασχηματισμό fourier 256 σημείων. Από τις 256 φασματικές συνιστώσες δημιουργείτε με ομαδοποίηση των καναλιών ένα διάνυσμα $x(t)$ 15 διαστάσεων όπου είναι και η είσοδος στο δίκτυο som.

Ένα μέρος που έχουν δει αρκετή χρήση οι χάρτες SOM είναι σε δορυφορικές εικόνες. Όταν γίνεται ανάλυση σε τέτοιου είδους εικόνες τα δεδομένα μας είναι τα pixel της εικόνας που αναλύουμε και γενικότερα κατηγοριοποιούνται . Αυτό είναι κάτι που γενικότερα είναι αρκετά αργό, ακριβό και με μεγάλες πιθανότητες λαθών. Οπότε ένας χάρτης SOM θα μπορούσε να εκπαιδευτεί σε αυτές της εικόνες από της οποίες μέσα υπάρχουν κατηγοριοποιημένα pixel και μη και μετά να εισαχθούν ταμπέλες σε δεδομένα που έχουν κατηγοριοποιηθεί. Συγκεκριμένα μια τέτοια χρήση έχει γίνει η οποία αφορούσε την ανακλαστικότητα ωκεανών όπου ένα 20×20 σε μέγεθος προπαμπίλιστικό SOM εκπαιδεύετε σε 43 χιλιάδες εξιδιάστατα δεδομένα. Από τα οποία το καθένα δείχνει ένα pixel εικόνας . Κάποιος μετά θα δώσει ταμπέλες σε μερικά από αυτά τα pixel τα οποία θα χρησιμοποιηθούν για να δώσουν ταμπέλες στους νευρώνες SOM . Τα αποτελέσματα αυτού του αλγορίθμου ήταν να δείξει σε αρκετές εικόνες χρησιμοποιούνται αρκετά εφόδια για να αφαιρεθούν τα σύννεφα από της εικόνες και το πως θα ήταν καλύτερη ιδέα να ξαναγίνει επεξεργασία εικόνων με πολλά σύννεφα. (36)

Μια άλλη εφαρμογή των αυτο-οργανόμενων χαρτών είναι σε σημασιολογικούς χάρτες. Η σημασιολογική χάρτες έχουν σκοπό την ομαδοποίηση των λέξεων με βάση την σημασία τους. Χρησιμοποίησαν 10000 τυχαίες προτάσεις η οποίες είχαν σαν λεξικό 15 προεπιλεγμένες λέξεις όπου 3 επιλεγόντουσαν για την πρόταση. Κάθε λέξη αντικαταστάθηκε με ένα 7διαστατο διανυσματικό κωδικό .κατόπιν εκπαιδεύτηκε ένα som δίκτυο 10×15 νευρώνων και 7×3 εισόδων . Έπειτα για κάθε διάνυσμα κωδικό δημιουργήθηκε ένα διάνυσμα $x(i)$ αποτελούμενο από 3 κομμάτια. την μέση τιμή του κωδικού της προηγούμενης λέξης , τον κωδικό πολλαπλασιασμένο με ένα θετικό συντελεστή κλίμακας και την μέση τιμή του κωδικού της επόμενης λέξης. Αυτά τα κομμάτια ονομαζόταν context και επέτρεπαν στο δίκτυο να δει της λέξης ως ένα οργανικό κομμάτι της πρότασης. Το αποτέλεσμα είναι ένας χάρτης ο οποίος ομαδοποιεί της λέξεις όχι με βάση των συντακτικό τους ρόλο αλλά επιπλέον με σημασιολογικά χαρακτηριστικά.

Μια ακόμη πολύ σημαντική εφαρμογή του som είναι η ομαδοποίηση λέξεων με βάση της σημασιολογικές τους κατηγορίες. Κάτι που χρησιμεύει πολύ σε μηχανές αναζήτησης μια μέθοδος για την στατιστική επεξεργασία των κειμένων είναι ο σχηματισμός ιστογραμμάτων που εμφανίζονται στο κείμενο. Θέλουμε δηλαδή να δείξουμε την συχνότητα εμφάνισης των λέξεων στο κείμενο μας. Για βοήθημα χρησιμοποιούμε ένα σημασιολογικό χάρτη ο οποίος έχει σκοπό να περιγράψει το προφίλ του κειμένου βασισμένο στο ιστόγραμμα του σημασιολογικού χάρτη μας. Η λογική είναι ότι παρόμοια κείμενα που μιλούν για παρόμοια ή και το ίδιο θέμα θα έχουν παρόμοια εμφάνιση των ίδιων σημασιολογικών ομάδων.

Την λογική αυτή ακολούθησε η μέθοδος websom από τον Τίμο Χόνκελα. Η μέθοδος χρησιμοποιεί δυο χάρτες som ο πρώτος από αυτούς ομαδοποιεί της λέξεις του κειμένου εισόδου σε σημασιολογικές κατηγορίες των λέξεων. Ο δεύτερος χάρτης δέχεται σαν είσοδο τα ιστογράμματα των κατηγοριών λέξεων των κειμένων τα οποία ομαδοποιεί με βάση αυτά τα ιστογράμματα. Με λίγα λόγια η αναζήτηση κειμένων με websom δεν γίνεται με λέξεις κλειδιά όπως παραδοσιακά γίνεται αλλά μέσω πλοήγησης στον δισδιάστατο χάρτη των κειμένων.

3 Παραλλαγές SOM

Η χάρτες SOM έχουν κατά καιρό λάβει παραλλαγές με σκοπό την καλύτερη ομαδοποίηση δεδομένων συγκεκριμένης κατηγορίας. Παραδείγματος για τοπογραφία αλλά ακόμα και για μη αριθμητικά δεδομένα. Η κατηγορίες αυτές έχουν αναπτυχθεί για την καλύτερη ομαδοποίηση δεδομένων ακόμα και σε προβλήματα που ο κανονικός som θα μπορούσε να δουλέψει. Σκοπός τους είναι το να χρησιμοποιηθεί η κύρια φιλοσοφία του αλγόριθμου som. Η πιο κύριες κατηγορίες φαίνεται να χωρίζονται σε παραλλαγές του SOM για ομαδοποίηση αριθμητικών δεδομένων και σχεσιακών η οποίες είναι η εξής.

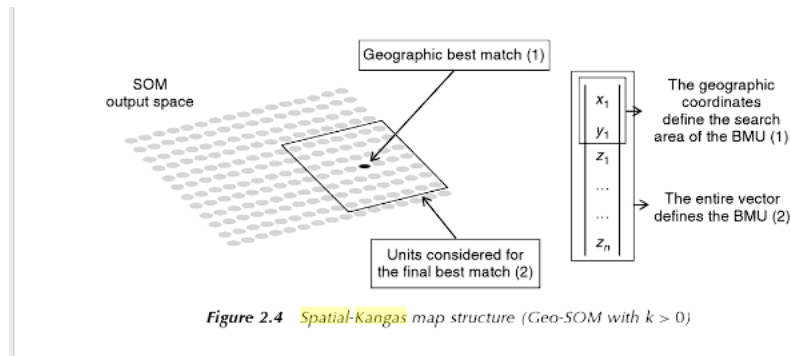
3.1 Geo-enforced som

Το geo-enforced som είναι μια παραλλαγή του αλγόριθμου που του επιτρέπει να δέχεται γεωγραφικά δεδομένα και να τα χρησιμοποιεί με την ίδια βαρύτητα με άλλα διαφορετικά δεδομένα.(1) Ο τρόπος που το κάνει αυτό είναι ακολουθώντας την λογική των εν μέρη παραλλαγών στις οποίες έχουμε ως σκοπό όπως αναφέρθηκε προηγούμενος να μετατρέψουμε κάποιες τιμές ώστε να είναι το ίδιο σημαντικές, να έχουν την ίδια βαρύτητα μεταξύ τους. Η μεγάλη αλλαγή σε αυτόν τον αλγόριθμο δεν βρίσκεται στον αλγόριθμο του SOM αλλά στο πως διαχειριζόμαστε τα δεδομένα μας πριν τα φορτώσουμε (1) . Δηλαδή το πως θα προεπεξεργαστούμε τα δεδομένα τα οποία θα χρησιμοποιήσουμε. Αυτή η μεταβλητή μια επιλογή από τον χρήστη και ο σκοπός μας στο τέλος της επεξεργασίας είναι να μπορούμε να χρησιμοποιήσουμε οποιαδήποτε μεταβλητή που μας δείχνει ένα σημείο στον χώρο μαζί με της κοινωνικό-οικονομικές μεταβλητές που θα χρησιμοποιήσουμε. Παρόλο που το συγκεκριμένο παράδειγμα δεν πειράζει των κώδικα του SOM είναι μια σημαντική βάση για τους επομένους 2 αλγορίθμους. Όμως γενικότερα το συγκεκριμένο παράδειγμα είναι ένας τρόπος να πρη-επεξεργαστούμε τα δεδομένα μας για να μπορέσουμε να τα χρησιμοποιήσουμε(2)

3.2 Spatial kangas som

Το Spacial kangas SOM είναι μια παραλλαγή του SOM την οποία εισάγουμε μια καινούργια μεταβλητή την λεγομένη μεταβλητή γεωγραφικής ανοχής. Ο τρόπος που λειτουργεί αυτή η παραλλαγή είναι σε δυο κύριες φάσεις. Στην πρώτη φάση αναζητείτε μια γεωγραφική γειτονιά όπου θα γίνει η αναζήτηση για τον νικητή νευρώνα. Αυτή η γεωγραφική γειτονιά έχει ως κέντρο την καλύτερη γεωγραφική αντιστοιχία και χρησιμοποιείτε σαν το κέντρο της αναζήτησης μας. Έπειτα περνάει στην επόμενη φάση στην οποία χρησιμοποιεί στην επιλογή του νικητή νευρώνα όχι μόνο την καλύτερη γεωγραφικά αντιστοιχία αλλά όλους τους νευρώνες γύρω από αυτήν. Το μέγεθος αυτής της αναζήτησης δίνεται από την προηγούμενως αναφερόμενη μεταβλητή γεωγραφικής ανοχής . Ουσιαστικά αυτό που κάνει είναι να επιλεγεί μόνο μια ομάδα από τους νευρώνες σαν υποψήφιους νικητές νευρώνες . Αυτό ουσιαστικά καταλήγει στο να διαχωρίζει την αναζήτηση σε δύο κομμάτια το πρώτο αναλαμβάνει οι γεωγραφικές συνθήκες να είναι αποδεκτές και έπειτα εφόσον έχουμε αποδεκτές συνθήκες να βρούμε πιο συγκεκριμένες πληροφορίες στην αναζήτηση στην μορφή του νικητή νευρώνα που επιλέγετε μέσα από την

ομάδα αυτή. Στην εικόνα (εικόνα 9) παρακάτω μπορούμε να δούμε ένα παράδειγμα ενός spatial-kangas χάρτη. Όπου αυτό που μας δείχνει είναι στο κέντρο του τετράγωνου την καλύτερη γεωγραφικά αντιστοιχία και μέσα στα όρια τους νευρώνες που θα χρησιμοποιηθούν στην αναζήτηση του νικητή νευρώνα.(3)(4)



(Εικόνα 9) Παράδειγμα χάρτη spatial-kangas στο κέντρο της τετράγωνης ομάδας βρίσκεται ο νευρώνας με την καλύτερη γεωγραφική αντιστοιχία και τα όρια σηματοδοτούν το πεδίο αναζήτησης για τον νικητή νευρώνα.(3)

3.3 GEO-SOM

Ο geo-som είναι μια ένωση των δύο παραπάνω παραδειγμάτων σε ένα. Η λογική του ξεκινάει με την σκέψη να χρησιμοποιηθεί ο πρώτος κανόνας της γεωγραφίας. Ο οποίος είναι “ τα πάντα σχετίζονται με τα πάντα, αλλά τα πράγματα που είναι πιο κοντά μεταξύ τους σχετίζονται μεταξύ τους περισσότερο (Ουάλντο Τομπλερ 1970)(1). Στην περίπτωση του αλγορίθμου αυτό μπορεί να μεταφραστεί ως η νευρώνες που είναι γεωλογικά πιο κοντά στο μοτίβο των δεδομένων μας θα σχετίζονται παραπάνω με τα δεδομένα και θα πρέπει να ψαχθούν με βάση αυτό. Το geo-som δεν είναι κάτι ιδιαίτερα διαφορετικό από αυτά που έχουμε εξηγήσει ως τώρα αλλά η ένωση τους. Είναι δηλαδή προ επεξεργασμένα δεδομένα τα οποία ελέγχουν χρησιμοποιώντας την μεταβλητή γεωγραφικής ανοχής που εξηγήθηκε παραπάνω το ποσό κοντά θέλουμε η νευρώνες να είναι γεωγραφικά στα δεδομένα(3). Αυτό φυσικά σημαίνει ότι εάν ορίσουμε αυτή την μεταβλητή στο 0 τότε ο νικητής νευρώνας θα είναι πάντα αυτός που είναι γεωγραφικά πιο κοντά στα δεδομένα που αναζητάμε. Κάτι το οποίο είναι κάτι που θυμίζει έναν υπερχάρτη το οποίο είχε προταθεί από τον Κοχόνεν το 1991 στον οποίο μόνο ορισμένα από τα δεδομένα που εισάγουμε χρησιμοποιούνται για να βρούμε τον νικητή νευρώνα(1). Πιο απλά η αναζήτηση του geo som λειτουργεί σε δύο κύριες φάσεις η οποίες θυμίζουν τον τρόπο τον spatial-kangas map. Φάση 1 να βρούμε τον πιο γεωγραφικά κοντινό νευρώνα στο δεδομένο που αναζητούμε. Φάση 2 να βρούμε τον νικητή νευρώνα σε μια ακτίνα αναζήτησης που έχουμε ορίσει ως μεταβλητή γεωγραφικής

ανοχής(2). Αυτό που τελικά καταφέρνει ο geo som είναι να υποχρεώσει τα δεδομένα αναζήτησης και τους νευρώνες να είναι κοντά στον χώρο μεταξύ του με σκοπό να φτιάξει ομάδες κοντά μεταξύ τους ακολουθώντας έτσι τον πρώτο κανόνα της γεωγραφίας.(1)Παρακάτω δίνεται ένα παράδειγμα της βασικής ιδέας του αλγορίθμου.

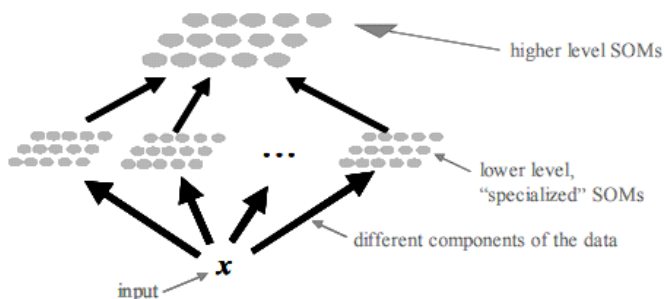
Αρχικά γίνεται η αρχικοποίησή των μεταβλητών που θα χρησιμοποιήσουμε. Αυτές είναι φυσικά η μεταβλητή γεωγραφικής ανοχής και φυσικά οι επαναλήψεις του αλγορίθμου. Επίσης εδώ πέρα ορίζουμε και το μέγεθος του χάρτη νευρώνων, την ακτίνα γειτνίασης, το πόσο μεγάλο θα είναι το βήμα ανανέωσης των νευρώνων καθώς όμως και τον ορισμό μιας λογικής μεταβλητής true/false η οποία θα χρησιμοποιηθεί για να ανάγκαση νευρώνες να μείνουν σε συγκεκριμένες γεωλογικές τοποθεσίες.. Επίσης επεξεργαζόμαστε τα δεδομένα όπως αναφέραμε στον geo enforced-som με σκοπό να μετατρέψουμε τις μεταβλητές που μας δείχνουν κάποιο σημείο στον τόπο στο να είναι ίσες σε βάρος μεταξύ τους.

Μετά περνάμε στο κομμάτι της εκπαίδευσης όπου μπαίνουμε στον πρώτο βρόχο επανάληψης ο οποίος θα τρέχει για όσο έχουμε αποφασίσει προηγούμενος στην αρχικοποίηση. Έπειτα μπαίνουμε στον επόμενο βρόχο ο οποίος θα ελέγξει της γεωγραφικές αποστάσεις όλων των νευρώνων με το δεδομένο που εξετάζουμε. Με σκοπό να βρούμε τον νευρώνα ο οποίος είναι γεωγραφικά πιο κοντά σε αυτό. Έπειτα θα επιλέξουμε τον νικητή νευρώνα εξετάζοντας μόνο τους νευρώνες που είναι μέσα στην ακτίνα της γεωγραφικής ανοχής που επιλέξαμε προηγούμενος με κέντρο τον νικητή γεωγραφικά νευρώνα.

Έπειτα προχωράμε στην κίνηση των νευρώνων στην οποία εξετάζουμε εάν η λογική μεταβλητή που ορίσαμε που κλειδώνει γεωγραφικά τους νευρώνες είναι true τότε η νευρώνες κινούν μόνο τα βάρη τα οποία έχουν να κάνουν με της μεταβλητές που δεν είναι η γεωγραφικές μεταβλητές αλλιώς γίνεται κανονική μετακίνηση όλων των βαρών και επαναλαμβάνουμε μέχρι να τελειώσουν η επαναλήψεις. Για μια πιο αναλυτική περιγραφή του αλγορίθμου μπορεί κάποιος να αναφερθεί στο (3).

3.4 Hierarchical soms

Τα hierarchical soms έχουν την λογική ότι πρώτα μικρά σε επίπεδο και πιο εξειδικευμένα SOM αναλύουν την πληροφορία και βγάζουν συμπεράσματα τα οποία περνάνε σε ένα μεγαλύτερο επιπέδου SOM για να βγάλει μια τελική συνοψισμένη πληροφορία (4). Στην γεωγραφική μεριά η χρήση των hierarchical som μας επιτρέπει να δημιουργήσουμε θεματικές ομαδοποίησής στα μικρότερα επίπεδα και να περάσουμε τα συμπεράσματα τους στο μεγαλύτερο επίπεδο μαζί με τα γεωγραφικά δεδομένα. Κάτι που μας επιτρέπει ένα δίκαιο συναγωνισμό νευρώνων σε διαφορετικά σε φύση δεδομένα και που μπορεί να είναι αρκετά πολύτιμο εφόσον έχουμε την δυνατότητα να ενώσουμε αυτά τα SOM δυναμικά μεταξύ τους(1). Με αυτόν τον τρόπο έχουμε την δυνατότητα να μελετήσουμε συγκεκριμένα δεδομένα και να τα ομαδοποιήσουμε με μεγαλύτερη ακρίβεια. Παρακάτω στην (εικόνα 10) μπορούμε να δούμε ένα παράδειγμα ενός hierarchical som.



(Εικόνα 10) Παράδειγμα hierarchical som όπου ένα δεδομένο X βλέπουμε χωρίζεται σε πολλά κομμάτια και ομαδοποιείται από διαφορετικά μικρότερα SOM τα οποία χρησιμοποιούνται σαν δεδομένα για ένα μεγαλύτερο κύριο som(1)

3.5 Σταθερή ανάθεση στον κανόνα χέσκες

Η συγκεκριμένη παραλλαγή στον αλγόριθμο SOM έγινε με σκοπό να λύσει την απουσία της μεταβλητής ενέργειας για τον συνεχή κανόνα μάθησης του Τούβο Κόχονεν (6). Η λύση που δόθηκε ήταν μια παραλλαγή στην μεταβλητή του νικητή νευρώνα από τον Τομ Χέσκες όπου του επιτρέπει να παίρνει την συνεχή κλίση από την μεταβλητή ενέργειας. (5)

Με τον κανόνα του Χέσκες η μεταβλητή της ενεργείας είναι συνεχής και γραμμική το οποίο κάνει τον συγκεκριμένο αλγόριθμο μια γραμμική καταφατική διαδικασία στην μεταβλητή ενέργειας.(5)

3.6 SOFT TOPOGRAPHIC MAPPING –STM

Ο αλγόριθμος soft topographic maps η αλλιώς STM είναι μια παραλλαγή του som η οποία δημιουργεί ένα χάρτη από νευρώνες η οποίοι είναι οργανωμένοι σε ένα ορθογώνιο πλέγμα το οποίο ορίζει την σχέση τους με τους γείτονες τους. (7) Η ιδέα του αλγόριθμου είναι ότι δημιουργεί έναν χάρτη από ένα σετ νευρώνων η μοντέλων οργανωμένα σε ένα ορθογώνιο πλέγμα το οποίο καθορίζει της γειτονικές σχέσεις τους(8). Ο αλγόριθμος είναι επίσης σε θέση να εργαστεί με δεδομένα των οποίων τα χαρακτηριστικά είναι ανάμοια μεταξύ τους σε αντίθεση με τον κλασικό SOM.Ο αλγόριθμος χρησιμοποιείται κυρίως για την εικονικοποίηση των δεδομένων ,ομαδοποίηση και στην αναγνώριση μοτίβων και έχει χρησιμοποιηθεί κυρίως σε βιολογικές και ιατρικές έρευνες. Παρακάτω εξηγείτε περιληπτικά η βασική ιδέα του αλγορίθμου.(7)(8)

Αρχικά ξεκινάμε με την αρχικοποίηση μερικών σημαντικών τιμών του αλγορίθμου όπως του πίνακα αναζήτησης για την εξίσωση γειτνίασης η οποία είναι η ίδια με τον κανονικό som, ο υπολογισμός του πίνακα ανομοιότητας των δεδομένων που θα χρησιμοποιήσουμε . Καθώς και σημαντικές μεταβλητές που θα χρησιμοποιηθούν στον αλγόριθμο όπως για παράδειγμα το κόστος εκχώρησης των διανυσμάτων των δεδομένων και της τιμές του β το οποίο μας δείχνει το πόσο πολύ εξομάλυνση έχει γίνει στην αρχική εξίσωση ενέργειας .Η β είναι υπεύθυνη για πόσο μεγάλη θα είναι η εκπαίδευση του αλγορίθμου οπότε επιλέγετε και μια τελική τιμή β_{final} που θα είναι μεγαλύτερη από την πρώτη τιμή β και θα λειτουργήσει σαν όριο στον αλγόριθμο. Έπειτα της αρχικοποίησης του αλγορίθμου προχωράμε στο κομμάτι εκπαίδευσης του αλγορίθμου το οποίο αποτελείται από δυο κυρίους βρόχους. Από τον πρώτο βρόχο ο οποίος περιέχει έναν expectation-maximization αλγόριθμο ο οποίος επαναλαμβάνεται μέχρι να τελειώσει ο δεύτερος βρόχος ο οποίος επαναλαμβάνεται όσο η τιμή β είναι μικρότερη από την β_{final} που ορίσαμε στην αρχικοποίηση του αλγορίθμου. Για μία πιο αναλυτική εξήγηση του κώδικα του αλγορίθμου STM μπορεί κάποιος να αναφερθεί στο (7).

3.7 Πιθανολογικά μοντέλα

Η βασική ιδέα των πιθανολογικών μοντέλων είναι να χρησιμοποιήσουν ένα πιθανολογικό πλαίσιο για την επιλογή του νικητή νευρώνα(5). Ο τρόπος που το κάνουν αυτό είναι επιβάλλοντας ένα περιορισμένο μίγμα γκαουσιανών κατανομών με τέτοιο τρόπο που να μιμούνται ένα πλέγμα SOM. Λόγω αυτόν τον παραλλαγών ο τελικός αλγόριθμος καταλήγει να διαφέρει αρκετά από τον κλασικό SOM (9).Μερικά σημαντικά πιθανολογικά μοντέλα είναι τα εξής.

3.7.1 Regularized EM

Βασισμένος σε προηγούμενη δουλειά του στην μεταβλητές ενέργειας για το SOM ο Χέσκες στο (46) μας έδειξε ότι μπορούμε να δούμε το batch som σαν ένα αλγόριθμο expectation maximization (EM) .Ωστόσο το αντίθετο της λογαριθμικής πιθανότητας μας οδηγεί σε k-means κβαντικά λάθη . Για την αποφυγή αυτόν τον λαθών τα δεδομένα μας κανονικοποιούνται με τρόπο όπου τιμωρούν πρωτότυπα τα οποία δεν σέβονται προηγούμενα δεδομένα και τη δομή τους. Ο τελικός αλγόριθμος αφοτου εφαρμόσουμε της αρχές του αλγορίθμου EM είναι ένας που μοιάζει αρκετά με τον SOM σε ρύθμιση batch.(9)

3.7.2 Variational EM

Το variational EM επιβάλλει τον περιορισμό στο κρυφό επίπεδο των μεταβλητών. Αυτό το κάνει εφαρμόζοντας μια τυχαία κατανομή q στις κρυφές μεταβλητές του προβλήματος. Η κρυφές μεταβλητές είναι αυτές που εκχωρούμε η οποίες αντιστοιχούν κάθε σημείο πληροφορίας με ένα κομμάτι του μίγματος (9). Η ισότητα της ολοκληρωμένης λογαριθμικής πιθανότητας του μοντέλου μίγματος είναι ίση με τρία κομμάτια. Την πλήρη πιθανότητα ολοκληρωμένη πάνω από της κρυφές μεταβλητές, την εντροπία της τυχαίας κατανομής και την απόκλιση Kullback-Leibler, μεταξύ της τυχαίας κατανομής q και της μεταγενέστερης κατανομής των κρυμμένων μεταβλητών γνωρίζοντας τα σημεία πληροφορίας. Αυτή η ισότητα επιτρέπει σε κάποιον να παράγει τον αλγόριθμο EM της μεταγενέστερης κατανομής των κρυμμένων μεταβλητών εφόσον γνωρίζουμε την ότι τα σημεία πληροφορίας μπορούν να υπολογιστούν ακριβώς. (9)

3.7.3 παραγωγική τοπογραφική χαρτογράφηση

Η παραγωγική τοπογραφική χαρτογράφηση είναι ένα πιθανολογικό μοντέλο το οποίο έχει σαν κύριο σκοπό να υπολογίσει την πυκνότητα και γενικότερα τον διασκορπισμό των δεδομένων μας στο χώρο. Είναι βασισμένο σε ομοιόμορφη κατανομή σε σταθερό πλέγμα το οποίο χαρτογραφείτε μέσω συγκεκριμένου ομαλού μη γραμμικού στο χώρο της πληροφορίας. (10) Ο συγκεκριμένος αλγόριθμος έχει τον περιορισμό του στον χώρο της διασκορπισμένης πληροφορίας διότι τα κέντρα της γκαουσιανής κατανομής υπολογίζονται μέσω ομαλής μη γραμμικής χαρτογράφησης. (5)

3.8 Αλγόριθμος για ενδεχόμενα δεδομένα

Ο αλγόριθμος som για ενδεχόμενα δεδομένα είναι ένας από τους τρόπους στους οποίους ο αλγόριθμος som έχει μετατραπεί για να μπορεί να διαχειρίζεται δεδομένα τα οποία δεν είναι αριθμητικά. (5) Γνωστός και ως KORRESP μιμείται κάτι που ονομάζεται Factorial Correspondence Analysis ή στα ελληνικά ανάλυση παραγοντικής αντιστοιχίας για να ενώσει της διαστάσεις των δεδομένων και μετά να συνεχίσει όπως ένας κανονικός SOM αλγόριθμος. (5) (11) Πιο συγκεκριμένα ο αλγόριθμος δέχεται για εισαγωγή δεδομένων ένα πίνακα δεδομένων διαστάσεων $m \times n$ τον οποίο επεκτείνει σε $(m+n) \times (n+m)$ ενώνοντας πρώτα την πιο πιθανή με βάση την ανάλυση αντιστοιχίας κολώνα και έπειτα κάνοντας το αντίθετο για της γραμμές. Το αποτέλεσμα είναι ένας πεταμένος πίνακας ο οποίος περιέχει δυο τύπους κολώνων της φόρμας η μάθηση μετά γίνεται όπως τον κανονικό som αλγόριθμο όπου τα δεδομένα που εισάγονται γίνονται εναλλάξ με τα παρακάτω κριτήρια. Αρχικά ο νικητής νευρώνας βρίσκεται σε βάζοντας της γραμμές που μετατοπιστήκανε από τον κανόνα της παραγοντικής αντιστοιχίας ενώ η ενημέρωση των γειτόνων χρησιμοποιούν όλο το διάνυσμα βάρους. Έπειτα γίνεται το ίδιο αλλά για της κολώνες που χρησιμοποιούνται για το ταίριασμα του som και όπως και στο πρώτο βήμα η ενημέρωση γίνεται χρησιμοποιώντας όλο το διάνυσμα βάρους (11).

.3.9 Median SOM

Το median som έχει σαν κύριο στόχο να προσαρμόσει την κύρια ιδέα του αλγορίθμου για ενδεχόμενα δεδομένα για να ομαδοποιεί καλύτερα δεδομένα κοντινών αποστάσεων(13). Μπορεί να χρησιμοποιηθεί σε γραφήματα δεδομένων αλλά και σε αλληλουχίες DNA(14). Προτάθηκε πρώτα από τον Τούβο Κόχονεν και Πάνου Σομέρβουνο το 1998 σε μια προσπάθεια να επεκτείνουν τον αλγόριθμο som για σχεσιακά δεδομένα ομοιότητας και ανομοιότητας όπως kernels και πίνακες ανομοιότητας(5)(12). Τα δεδομένα υποτίθεται ότι μπορούν να περιγράφουν σε ένα συμμετρικό πίνακα ανομοιότητας σε διακριτή τοποθέτηση.(5)

Ένα παράδειγμα του median som αλγόριθμου στο πώς δουλεύει σε ένα σχήμα διακριτής βελτιστοποίησης σε λειτουργία ομαδοποίησης είναι

1. Ανάθεση όλων των δεδομένων στην καλύτερες μονάδες που αντιστοιχούνται
- 2.ενημέρωση των πρωτοτύπων μέσα στο σετ τιμών (5)

Με αυτόν τον τρόπο όσο ο αλγόριθμος εξερευνά ένα μη άπειρο σετ δεδομένων , επικεντρώνετε σε ένα τοπικό ελάχιστο της συνάρτησης ενέργειας. Ένα από τα κακά του αλγορίθμου είναι ότι όλα τα πρωτότυπα πρέπει να ανήκουν στο σετ δεδομένων μας και είναι επίσης πολύ ακριβό από την μεριά της υπολογιστικής δύναμης που χρειάζεται για να τρέξει. (5)

3.10 Relational SOM

Σε σχέση με τον median SOM Ο relational SOM μας επιτρέπει να δείξουμε τα πρωτότυπα μας με ένα εικονικό γραμμικό συνδυασμό όλων των στοιχείων μας στο σετ δεδομένων.(16) Δηλαδή αν τα δεδομένα μας μπορούν να περιγράφουν από ένα πίνακα ανομοιότητας $D=(\delta(x_i, x_j))_{i,j=1,\dots,n}$ τότε μπορούν να ενσωματωθούν σε ένα ψευδό ευκλιδικό χώρο.(5)

Το πρώτο βήμα του αλγορίθμου είναι να βρει τον νικητή νευρώνα μιας παρατήρησης όπως και σε ένα κλασσικό SOM το οποίο μπορεί να γενικευθεί απευθείας στις ανομοιοτήτες της on-line ρύθμισης αλλά και του batch.

(5)

Για την on-line ρύθμιση η πράξεις είναι ίδιες με τον κανονικό som ενώ για την batch ρύθμιση μόνο η ανανέωση των πρωτοτύπων είναι η ίδια με τον κανονικό batch αλγόριθμο.

Παρόλο που είναι δυνατή η δημιουργία ενός batch relational som η δημιουργία ενός on-line relation som προτιμήθηκε από (15) καθώς η batch version παρόλο είναι ντετερμινιστική και συνήθως συγκεντρώνεται σε λίγα βήματα έχει αρκετά προβλήματα στην οργάνωση του τελικού χάρτη, στο να παράγει ανισόρροπες κλάσεις ενώ ταυτόχρονα είναι πάρα πολύ εξαρτημένη από την αρχικοποίηση της. Παρακάτω εξηγείτε η βασική ιδέα του on-line RSOM σε παράδειγμα που θεωρούμε ότι έχουμε μόνο μια παρατήρηση.

Ξεκινάμε με το να κάνουμε αρχικοποίηση της τιμές των δεδομένων καθώς όμως και τις επαναλήψεις που θα εκτελέσει ο αλγόριθμος και το μέγεθος του χάρτη. Έπειτα επιλέγεται ένα τυχαίο δεδομένο και μπενούμε στον βροχο του αλγοριθμου.

Εξετάζουμε της αποστάσεις μεταξύ του δεδομένου που επιλέξαμε και των νευρώνων για να βρεθεί ποιος νευρώνας είναι κοντινότερος στο δεδομένο. Κατόπιν όπως και στον κανονικό som μεταβάλλουμε τα βάρη των νευρώνων και επαναλαμβάνουμε για όσες επαναλήψεις έχουμε επιλέξει. Για μια πιο αναλυτική εξήγηση του αλγορίθμου μπορεί κάποιος να αναφερθεί στο(15).

3.10.1 Fuzzy relational som

Ο fuzzy relational som είναι μια παραλλαγή του relational som αλγοριθμου(17). Η αρχική ιδέα προτάθηκε από τους R.D. Pascual-Marqui, A.D. Pascual-Montano, K. Kochi, J.M. Carazo, για έναν fuzzy SOM αλγορίθμου όπου είναι μια κανονικοποίηση του fuzzy c-means αλγοριθμου (18). Με σκοπό να χρησιμοποιηθεί με ασαφή δεδομένα και βάρη όπως για παράδειγμα σε οικονομικό καταμερισμό δεδομένων πιστωτικών καρτών που χρησιμοποιήθηκε στο(19). Ο fuzzy som στο (18) είναι βασισμένος στην εξίσωση κόστους που παράγεται αφότου εισαχθούν δυο τροποποίησης στον αλγόριθμο του fuzzy c-means αρχικά τα διανύσματα κώδικα κατανέμονται σε ένα κανονικοποιημένο λίγων διαστάσεων πλέγμα όπως και στο som και μια τιμή ποινής προθέεται για να εγγυηθεί την ομαλή κατανομή των διανυσμάτων κώδικα στο πλέγμα και για να βοηθήσουν να διατηρηθεί η τοπολογική δομή των δεδομένων(17)(20)(21). Ο fuzzy relational som (frsom) προτάθηκε από τους (17) και είναι ένας συνδυασμός των αλγορίθμων relational som και Non-Euclidean relational fuzzy clustering(NERFCM) (17) και ο τρόπος με τον οποίο λειτουργεί είναι παρόμοιος με αυτού του απλού batch som (BSOM) αλλά με κάποιες διαφορές μεταξύ τους.(17)

Μια από τις διαφορές μεταξύ του FRSOM και BSOM είναι στην εισαγωγή των δεδομένων, φυσικά όπου στο FRSOM τα δεδομένα που εισέρχονται είναι σχετικά(17) ενώ στο BSOM είναι αντικειμενικά. Μια άλλη διαφορά του είναι ο τρόπος με τον οποίο χειρίζονται της απόστασης. Για παράδειγμα στον BSOM η απόστάσεις μεταξύ των αντικειμένων υπολογίζονται και το αντικείμενο με την μικρότερη απόσταση θεωρείται ο νικητής νευρώνας. Ενώ στο FRSOM λόγω του ότι τα δεδομένα μας δεν είναι αντικειμενικά αλλά σχετικά η απόσταση για τον νικητή νευρώνα υπολογίζεται με βάση την ανομοιότητα των δεδομένων. Παρακάτω εξηγείτε περιληπτικά το πως λειτουργεί ο αλγόριθμος.(17)

Αρχικά κάνουμε μια αρχικοποίηση των τιμών των νευρώνων του χαρτί όπως και στον κανονικό som αυτό μπορεί να αποτελεί τυχαίες τιμές ή τιμές ανάλογες με τα δεδομένα μας. Επίσης αρχικοποιούμε και την τιμή m η οποία είναι το fuzzifier η αλλιώς η μεταβλητή θόλωσης η οποία

ορίζεται σε μια τιμή μεγαλύτερη του 1. Τέλος ορίζουμε της επανάληψης και την ακτίνα γειτνωσίας για τους νευρώνες.

Έπειτα προχωράμε στην εκπαίδευση του αλγορίθμου η οποία αποτελείται από δύο βρόχους.

Στον πρώτο βρόχο υπολογίζουμε της αποστάσεις μεταξύ νευρώνων και των δεδομένων για να βρούμε τον νικητή νευρώνα και υπολογίζουμε την γειτονιά που θα ανήκει η οποία επηρεάζεται από την μεταβλητή θόλωσης.

Τέλος στον επόμενο βρόχο ανανεώνουμε της τιμές των νευρώνων η οποίες επίσης επηρεάζονται από την μεταβλητή θόλωσης και την ακτίνα γειτνωσίας.

Για μια πιο αναλυτική περιγραφή του αλγορίθμου μπορεί κάποιος να αναφερθεί στο (17)

Η μεγάλη διαφορά μεταξύ στους δύο αλγορίθμους βρίσκεται στην χρήση της μεταβλητής θόλωσης όπου μπορεί να βρει υποδομές σε ομάδες των δεδομένων που άλλη αλγόριθμοι με πιο ξεκάθαρους και ευδιάκριτους τρόπους ομαδοποίησης δεν θα καταφέρανε να βρουν .

Ο συγκεκριμένος αλγόριθμος μετά από τεστς βρέθηκε να λειτουργεί καλύτερα σε σχετικά δεδομένα τα οποία δεν είχαν μεγάλες απόστασης μεταξύ τους η και που πολλές φορές επικάλυπταν το ένα με το άλλο πράγμα που άλλη αλγόριθμοι θα είχαν περισσότερες δυσκολίες ενώ αντιθέτως σε δεδομένα τα οποία είχαν μεγαλύτερες αποστάσεις μεταξύ τους με λιγότερες περίπτωσης όπου μια ομάδα δεδομένων επικάλυπτε μια άλλη βρέθηκε να μην λειτουργεί όσο καλά όσο άλλοι.(17)

3.11 Kernel SOM

Το Kernel SOM είναι ένας τύπος του SOM αλγορίθμου ο οποίος είναι kernelized δηλαδή τα δεδομένα έχουν τοπογραφηθεί σε χάρτη μεγάλων διαστάσεων μέσω ενός μη γραμμικού χάρτη .Με σκοπό να μετατρέψουμε δύσκολα προβλήματα τα οποία είναι μη γραμμικά σε γραμμικά τοπογραφώντας τα σε χάρτες μεγάλων διαστάσεων.(23) Φυσικά δεν υπάρχει μόνο ένας τρόπος να γίνει αυτό και έχουν γραφτεί αρκετά kernelized SOM αλγόριθμοι . Γενικότερα μέσω πειραμάτων που έχουν γίνει χρησιμοποιώντας και on-line αλλά και batch version του αλγορίθμου. Έχει παρατηρηθεί ότι ο kernel som είναι αρκετά καλός στο να ομαδοποιεί και να εμφανίζει κορυφές σε δεδομένα τα οποία δεν έχουν “τέλειες” γειτονίες μεταξύ τους. (22)Μεταξύ των δύο αλγορίθμων είναι επίσης αξιότιμο να αναφερθεί ότι ο batch kernel som βρέθηκε να είναι γρηγορότερος και φθηνότερος από τον on-line kernel som.(23) Παρακάτω αναφέρονται και η δύο εκδόσεις του αλγορίθμου ξεκινώντας με ένα παράδειγμα της Nathalie villa και Fabrice rossi πάνω στην on-line ρύθμιση του kernel som αλγορίθμου.

3.11.1 On-line kernel som

Παρακατώ επεξηγείται ένα παράδειγμα ενός on-line kernel som . Ο συγκεκριμένος αλγόριθμος δέχεται σαν δεδομένα εικόνες όπου όπως και στον κανονικό som αρχικά χαρτογραφούνται σε ένα χάρτη νευρώνων μικρών διαστάσεων η οποίοι συσχετίζονται μεταξύ τους μέσω γειτονικών σχέσεων .(23) Παρόλο που το on-line kernel είναι κάτι που έχει προταθεί και φτιαχτεί παρατηρείτε πιο αργό από ότι το batch kernel som έως και 10 φορές και άνω και για αυτό δεν προτιμάται.(23)

Η βασική ιδέα του αλγορίθμου στην αρχικοποίηση του είναι η ίδια με τον κανονικό αλγόριθμο ,η κυρία διαφορά του ξεκινάει με τον τρόπο στον οποίο αναπαριστούμε τους νευρώνες μας . Η οποίοι δεν είναι απλώς ένα διάνυσμα τιμών αλλά αναπαριστώνται από ένα πρότυπο στον χώρο hilbert. Τα δεδομένα ομαδοποιούνται με βάση το πόσο κοντά είναι σε αυτά τα πρότυπα και εννοείται ότι έπειτα σε κάθε βήμα τα πρότυπα ανανεώνονται έτσι ώστε ανάλογα όπως και με τον κανονικό som. Για μια πιο περιγραφική εξήγηση μπορεί κάποιος να αναφερθεί στο (23)

Batch kernel som

Ένα άλλο παράδειγμα του αλγορίθμου το batch kernel som προτάθηκε από τους Romain Boulet a, Bertrand Jouve a, Fabrice Rossi b, Nathalie Villa (22)με σκοπό ανάλυσης social networks δηλαδή να βρεθούν ενώσεις μεταξύ διαφορετικών ομάδων ανθρώπων και τι τους ενώνει. Ο τρόπος που το κάνει αυτό ξεκινάει τοπογραφώντας τα αρχικά δεδομένα σε ένα μεγάλο διαστάσεων χώρο hilbert (22) μέσω ενός χάρτη χαρακτηριστικών και έπειτα ομαδοποιούνται χρησιμοποιώντας ένα κλασσικό batch som αλγόριθμο.(22) Είναι επίσης αξιόλογο να αναφερθεί ότι τα αποτελέσματα του αλγορίθμου βασίζονται αρκετά πολύ στην αρχικοποίηση του αλγορίθμου και φυσικά στην εκτέλεση του. Παρακάτω είναι ένα παράδειγμα πολύ βασικό παράδειγμα του αλγορίθμου και της βασικής του ιδέας.(22)

Παρόμοια με το προηγούμενο παράδειγμα ξεκινάμε με την αρχικοποίηση των τιμών ,δεδομένων και τον προτύπων των νευρώνων σε χώρο hilbert. Έπειτα για κάθε νευρώνα και δεδομένο ξεκινάμε να κάνουμε την ομαδοποίηση του κάθε δεδομένου προς το πιο κοντινό πρότυπο hilbert του κάθε νευρώνα. Τέλος όπως και στον κανονικό batch som ανανεώνουμε τα πρότυπα hilbert των νευρώνων και επαναλαμβάνουμε αναλόγως με το πόσα βήματα ,νευρώνες και δεδομένα έχουμε. Για μια πιο αναλυτική περιγραφή του αλγορίθμου μπορεί κάποιος να αναφερθεί στο(22).

3.11.2 Bagged kernel som

Παρόλο που τα kernel SOM είναι αρκετά καλά στο να διαχειρίζονται γενικοποιημένα δεδομένα έχουν και πάλι μια μεγάλη αδυναμία στο να διαχειρίζονται μεγάλα σετ δεδομένων, αυτό γίνεται κυρίως λόγω του ότι τα δεδομένα μας περιγράφονται σε χάρτες μεγάλων διαστάσεων λόγω της φύσης τους κάνοντας τα χρονοβόρα να παραχθούν σε μεγάλα σετ δεδομένων και μεγάλα σε όγκο και επίσης δεν είναι εύκολο κάποιος να καταλάβει ένα μεγάλο σετ δεδομένων όταν αυτά είναι χαρτογραφημένα σε ένα μεγάλο διαστάσεων χάρτη.(25) Με σκοπό να λυθούν αυτά τα κύρια προβλήματα προτάθηκε από Jerome Mariette, Madalina Olteanu, Julien Boelaert, και Nathalie Villa-Vialaneix ένας διαφορετικός τύπου kernel som είναι ο bagged kernel. Η βασική ιδέα του bagged kernel som είναι περάσουμε τα δεδομένα μας μέσα από πολλαπλά som με τυχαία μικρά υπό δείγματα των δεδομένων με σκοπό να επιτύχουμε ένα αποτέλεσμα όμοιο των kernel som που αναφέρθηκαν προηγούμενος αλλά με μικρότερες διάστασης και λιγότερα λάθη. (25)

Τα βήματα που ακολούθη ο αλγόριθμος είναι τα εξής: αρχικά χρησιμοποιώντας την ρύθμιση on-line του αλγόριθμου που αναφέρθηκε παραπάνω εκπαιδεύουμε μικρά τυχαία υπό δείγματα των δεδομένων μας. Έπειτα επιλέγονται τα πρωτότυπα που παρατηρούνται ότι εκπροσωπούν περισσότερο τα δεδομένα μας σαν τα πρώτα βάρη για να βοηθήσουν να επιλέξουν τα επόμενα πρωτότυπα. Έπειτα κάθε φορά που μια παρατήρηση ένα καινούργιο υπό δείγμα υπολογίζετε το οποίο χρησιμοποιείτε για σαν κριτήριο ποιότητας για να επιλεχθούν η επόμενες πιο σημαντικές μεταβλητές. Τέλος δημιουργείται ένας τελευταίος χάρτης με της επιλεγμένες παρατήρησης που έγιναν και γίνεται μια τελευταία ομαδοποίηση των δεδομένων. Για μια πιο αναλυτική περιγραφή του αλγορίθμου μπορεί κάποιος να αναφερθεί στο (25).

Ο συγκεκριμένος αλγόριθμος έχει γενικότερα μικρότερο υπολογιστικό κόστος σε σχέση με τον εξ ολοκλήρου kernel. Αυτό φυσικά δεν έρχεται με τα αρνητικά του μιας και έχει μεγαλύτερο τοπογραφικό λάθος σε μεγάλα σετ δεδομένων ενώ είχε μικρότερο σε μικρότερα ταυτόχρονα το ακριβώς αντίθετο συμβαίνει στο λάθος στην ποσοτικοποίηση. Είναι επίσης άξιο να αναφερθεί ότι σε σχέση με έναν απλό batch som αλγόριθμο το τοπογραφικό λάθος ήταν πάλι αντίθετα μεγαλύτερο σε σετ δεδομένων ενώ το αντίθετο γινότανε σε λάθη στην ποσοτικοποίηση σε σχέση με τον bagged kernel som. (25) Ο συγκεκριμένος αλγόριθμος χρησιμοποιήθηκε στην ομαδοποίηση ego-facebook network τα οποία είναι κύκλοι λιστών φίλων από το facebook. (24) Σε αυτό το παράδειγμα δόθηκε μια γεμάτη λίστα σε ένα κανονικό kernel on-line som, ένα δείγμα χίλιων δεδομένων σε ένα bagged kernel som και το ίδιο τυχαία επιλεγμένο δείγμα σε έναν κανονικό kernel som. Στο παράδειγμα παρατηρήθηκε μικρότερο επίπεδο λάθους στους δυο αλγορίθμους που δόθηκε ένα κομμάτι από το σετ δεδομένων. Ενώ ταυτόχρονα το bagged kernel είχε το μικρότερο βαθμό λάθους μεταξύ τους. (25)

4 Ανάλυση των δεδομένων και εξήγηση τους

Τα δεδομένα που θα χρησιμοποιηθούν για το πρακτικό κομμάτι παρακάτω έχουν την μορφή συμμετρικών πινάκων ανομοιότητας. Δηλαδή είναι τετραγωνική πίνακες $X \times X$ η οποίοι έχουν μέσα τιμές ευκλιδιανών αποστάσεων η οποίες έχουν ως σκοπό να δείξουν το πόσο διαφορετικά είναι τα δεδομένα μεταξύ τους. Σκοπός των δεδομένων αυτών είναι να δείξουμε της διαφορές, δυσκολίες και αποτελέσματα μεταξύ των δύο αλγορίθμων που θα χρησιμοποιηθούν ο κλασσικός SOM και ο RSOM. Κύρια αποτελέσματα που θα έχουμε είναι η ομαδοποίηση των σετ δεδομένων σε ομάδες από της οποίες δεν γνωρίζουμε αν κάποιο δεδομένο ανήκει σε κάποια ομάδα με σκοπό να δούμε πως διαχειρίζονται τα δεδομένα η δυο αλγόριθμοι. Λόγω της φύσης αυτών των δεδομένων πρόκεινται για μεγάλων διαστάσεων δεδομένα τα οποία δεν γίνεται να δειχτούν στον χώρο εύκολα παρόλα αυτά μέσω απλοποίησης των δεδομένων με την χρήση κλασσικής πολυδιάστατης κλιμάκωσης θα γίνει μια επίδειξη των δεδομένων στο χώρο για την βοήθεια κατανόησης της μορφής των δεδομένων. Παρακάτω δίνεται μια μικρή εξήγηση των δεδομένων που θα χρησιμοποιηθούν καθώς και οποιαδήποτε επεξεργασία που μπορεί να έγινε πάνω τους για να είναι συμβατά.

4.1 iris dataset

Το συγκεκριμένο σετ δεδομένων πρόκειται για το αρκετά γνωστό iris dataset το οποίο χρησιμοποιείτε συχνά. Πρόκειται για ένα σετ δεδομένων το οποίο έχει 150 διαφορετικά δείγματα λουλουδιών iris τα οποία περιγράφονται από 4 διαφορετικά χαρακτηριστικά τα οποία είναι

Το μήκος του σέπαλου: Το ποσό μακρύ είναι το σέπαλο του λουλουδιού iris.

Το πλάτος του σέπαλου: Το πόσο πλάτη είναι το σέπαλο του λουλουδιού iris

Το μήκος του πετάλου: Το πόσο μακρύ είναι το πέταλο του λουλουδιού iris

Το πλάτος του πετάλου : Το πόσο πλατύ είναι το πέταλο του λουλουδιού iris

Τα συγκεκριμένα χαρακτηριστικά περιγράφονται σε εκατοστά με σκοπό να περιγράψουν το λουλούδι το οποίο μπορεί να ανήκει σε μία από της τρεις ομάδες η οποίες είναι.

Iris setosa: Μικρότερο σε μέγεθος από ότι τα υπόλοιπα και χαρακτηριστικές διαστάσεις στα πέταλα και σέπαλα.

Iris versicolor: Μεσαίο σε μέγεθος με χαρακτηριστικά σε διάστασης πετάλων και σέπαλα να πέφτουν αναμεσά στον setosa και virginica.

Iris virginica: Τα μεγαλύτερα σε μέγεθος από τα τρία με της μεγαλύτερες επίσης διαστάσεις σε πέταλα και σέπαλα μεταξύ των τριών ομάδων .

(40)

Παρόλα αυτά το συγκεκριμένο dataset έχει ένα πρόβλημα με τα δεδομένα μας το οποίο είναι η διαστάσεις του χάρτη. Όπως προαναφέρθηκε οι διαστάσεις του πίνακα των δεδομένων μας πρέπει να είναι ίσες μεταξύ τους και επίσης πρέπει να δείχνουν της διαφορές μεταξύ των δεδομένων μας. Για την λύση αυτού του προβλήματος θα χρειαστεί να μετατρέψουμε τα δεδομένα σε ευκλιδικές αποστάσεις. Η μετατροπή είναι μια κλασική μετατροπή σε ευκλείδεια απόσταση απλώς γίνεται για όλον τον πίνακα και για όλα τα δεδομένα. Για παράδειγμα έστω ότι έχουμε ένα πίνακα με δεδομένα δύο διαστάσεων. Όπως για παράδειγμα με την (εικόνα 11) παρακάτω. Η μετατροπή σε ένα πίνακα διαστάσεων θα ακολουθούσε την λογική η κάθε καταχώρηση πίνακα αποστάσεων θα ήταν ουσιαστικά η απόσταση δύο εισόδων του αρχικού πίνακα για παράδειγμα όπως παρακάτω. Η απόσταση μεταξύ του A με του A είναι 0 μιας και μιλάμε για τον εαυτό του αλλά η απόσταση μεταξύ του A και B θα είναι 16.

	X	Y		A	B	C	D	E	F
A	9	49	A	0	16	47	72	77	79
B	24	54	B	16	0	37	57	65	66
C	51	28	C	47	37	0	40	30	35
D	81	54	D	72	57	40	0	31	23
E	81	23	E	77	65	30	31	0	10
F	86	32	F	79	66	35	23	10	0

(Εικόνα 11) Πίνακας εισόδων δεδομένων στα αριστερά και πίνακας αποστάσεων στα δεξιά μετά την μετατροπή τους σε ευκλείδειες αποστάσεις από(41)

Ο υπολογισμός των αποστάσεων μεταξύ δυο τιμών του πίνακα είναι πάρα μια μαθηματική πράξη της ευκλιδικής απόστασης χ . Η πράξη αυτή είναι:

$$D = \sqrt{(\chi_2 - \chi_1)^2 + (y_2 - y_1)^2}$$

Παρόλο που τα δεδομένα του iris είναι τετραδιάστατα ακολουθούν την ίδια λογική απλά η πράξη μετατρέπεται με παρόμοιο τρόπο για να ικανοποιεί της παραπάνω διαστάσεις δηλαδή για

παράδειγμα ας πάρουμε της πρώτες δύο τιμές του iris (Εικόνα 12)

5.1 3.5 1.4 0.2

4.9 3.0 1.4 0.2

(Εικόνα 12) πρώτα 2 δεδομένα στο σετ δεδομένων iris.

Η πράξη που θα κάνουμε για να υπολογίσουμε την απόσταση μεταξύ τους θα είναι :

$$D = \sqrt{(5,1 - 4,9)^2 + (3,5 - 3,0)^2 + (1,4 - 1,4)^2 + (0,2 - 0,2)^2}$$

με αποτέλεσμα περίπου:

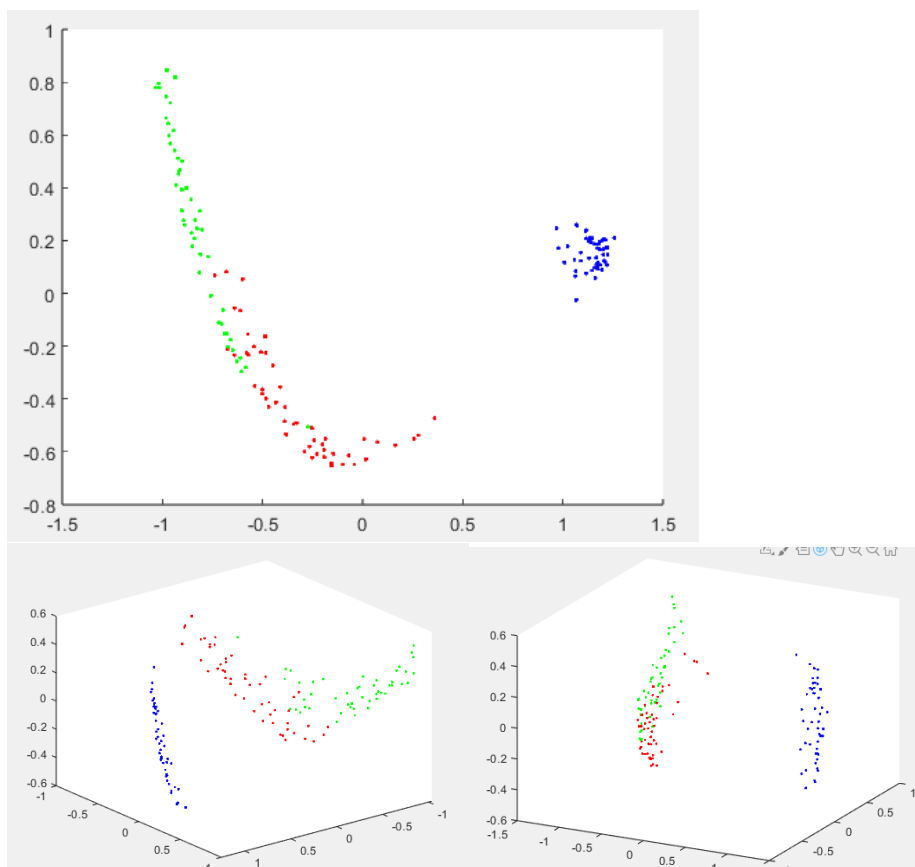
$$D = 0.538516$$

Με αυτήν την λογική μεταφράζουμε ουσιαστικά όλα τα δεδομένα μας σε ένα πίνακα αποστάσεων 150x150 τα οποία δείχνουν της διαφορές μεταξύ τους για να μπορέσουμε να τα χρησιμοποιήσουμε στον αλγόριθμο RSOM .

Για να επιτευχθεί αυτό θα γίνει χρήση μιας βιβλιοθήκης python με όνομα SciPy (appendix 1) . Συγκεκριμένα θα καταλήξουμε με ένα χάρτη διαστάσεων 150*150 οποίος θα δείχνει της διαφορές μεταξύ καθενός δείγματος ενός λουλουδιού μεταξύ ενός άλλου.

Για να μπορέσουμε να πάρουμε μια καλύτερη εικόνα των δεδομένων και της μορφής τους στον χώρο θα κάνουμε χρήση μιας συνάρτησης της matlab με όνομα cmdscale. Είναι μια συνάρτηση η οποία κάνει κλιμάκωση των δεδομένων ενός πίνακα με δεδομένα όχι μόνο ανομοιότητας αλλά και δεδομένα σε ευκλιδικές αποστάσεις η και σε δεδομένα ομοιότητας(45). Γνωστή και ως ανάλυση κυρίων συντεταγμένων η στα αγγλικά ,”principal coordinate analysis” μας βοηθάει να μειώσουμε της διαστάσεις σε ένα πίνακα με σκοπό να μπορέσουμε να της απεικονίσουμε σε διαστάσεις που μπορούμε να διαχειριστούμε. Δηλαδή παίρνουμε πίσω ένα πίνακα συντεταγμένων των οποίο μπορούμε να δείξουμε ευκολότερα σε δισδιάστατο η τρισδιάστατο χώρο. Για περισσότερες πληροφορίες για το πως γίνεται αυτό μπορεί κάποιος να διαβάσει παραπάνω εδώ(44).

Παρακάτω (εικόνα 13) έχουμε τα αποτελέσματα των δεδομένων μετά από το script απλοποιημένα στον χώρο σε τρεις διαστάσεις .Επίσης έχουν δοθεί χρώματα στα σημεία των δεδομένων όπου τα πρώτα 50 χρωματίζονται ως μπλε, τα επόμενα 50 ως κόκκινα και τα τελευταία 50 ως πράσινα .Αξιοσημείωτο εδώ περα είναι να αναφερθεί ότι στο αρχικό σετ δεδομένων του iris τα πρώτα 50 παραδείγματα είναι δείγματα του λουλουδιού setosa , έπειτα τα επόμενα 50 του versicolor και τέλος του virginica. Μπορούμε να δούμε ότι εμφανίζουν δύο ξεκάθαρες ομάδες μεταξύ τους . Κάτι που είναι αρκετά ενδιαφέρον εάν κάποιος σκεφτεί ότι πρόκειται για τρεις αρχικές ομάδες . Όμως αξιότιμο επίσης να αναφερθεί ότι μια από τις ομάδες στο σετ δεδομένων iris. Συγκεκριμένα το versicolour πρόκειται για ένα φυτό το οποίο έχει τιμές ενδιάμεσα των άλλων δύο συγκεκριμένα η τιμές του είναι αρκετά κοντά με το virginica ενώ το setosa είναι το μικρότερο και πιο απομακρυσμένο από τα τρία σε τιμές. Πράγμα που βλέπουμε και στα δεδομένα παρακάτω με το παράδειγμα του setosa να είναι σε μια δικιά του απομονωμένη ομάδα με μπλε ενώ το versicolor σε κόκκινο και virginica σε πράσινο βλέπουμε ότι σχηματίζουν μια μεγαλύτερη δικιά τους ενωμένη και ελάχιστα ανακατεμένη ομάδα.



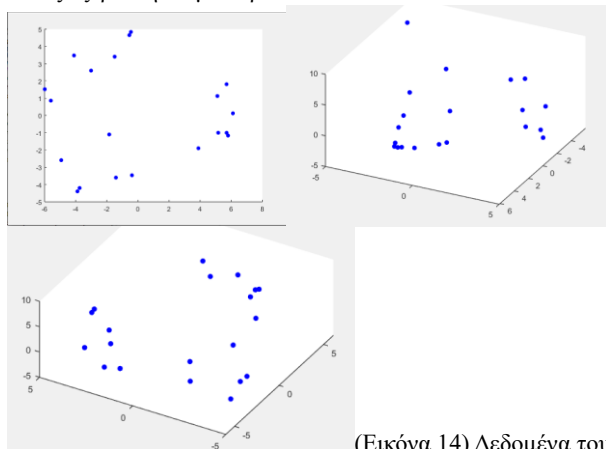
(Εικόνα 13) Δεδομένα του σετ δεδομένων iris σε ευκλιδιανές αποστάσεις απεικονισμένα σε τρεις διαστάσεις σε διαφορετικές γωνίες, χρωματισμένες είναι επίσης και η ομάδες με το setosa χρωματισμένο με μπλε, το versicolor με κόκκινο και με πράσινο το virginica.

Παρακάτω στα πειράματα θα δούμε μερικές πολύ αξιότιμες παρατηρήσεις στα δεδομένα αλλά προς το παρόν τα σχεδιαγράμματα μας δίνουν μια ιδέα του τι είδους δεδομένα έχουμε να κάνουμε.

4.2 Σετ υπολειμμάτων αμινοξέων

Το συγκεκριμένο σετ δεδομένων πρόκειται για τα χαρακτηριστικά υπολειμάτων αμινοξέων τα οποία αντικαταστάθηκαν με διαφορετικά αμινοξέα(43). Πρόκειται για ένα μικρότερο σε διαστάσεις πίνακα συγκεκριμένα 20x20 και παρόλο που δεν είναι ευκλείδεια απόσταση όπως συνηθίζεται η απόσταση grantham θυμίζει αρκετά τα χαρακτηριστικά της. Λόγω του ότι είναι αρκετά όμοιες μεταξύ τους δεν υπάρχει κάποιο πρόβλημα στην χρήση της. (42)

Η απόσταση grantham η αλλιώς grantham score είναι ένας τρόπος για να μέτρησε κάποιος της διαφορές μεταξύ δυο αμινοξέων τα οποία μεταλλάσσονται μέσα σε μια πρωτεΐνη. Το συγκεκριμένο σετ δεδομένων παρόλο που είναι μικρότερο πιστεύεται ότι μπορεί να δείξει πως διαχειρίζονται η δύο αλγόριθμοι μικρότερα σε διαστάσεις δεδομένα ανομοιότητας. Αλλά ταυτόχρονα και το πως διαχειρίζονται δεδομένα που δεν είναι σε ευκλιδικές αποστάσεις. Γενικότερα σε σχέση με το προηγούμενο σετ δεδομένων μπορούμε να δούμε τα δεδομένα απλοποιημένα στο χώρο (εικόνες) ότι χωρίζονται σε 2 με 3 ομάδες και είναι αρκετά ενδιαφέρον να τις εξερευνήσουμε παραπάνω.

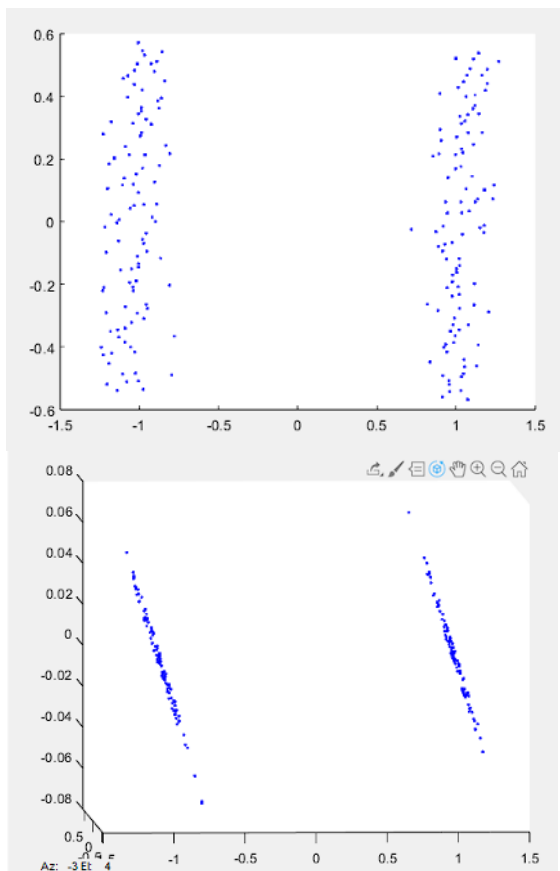


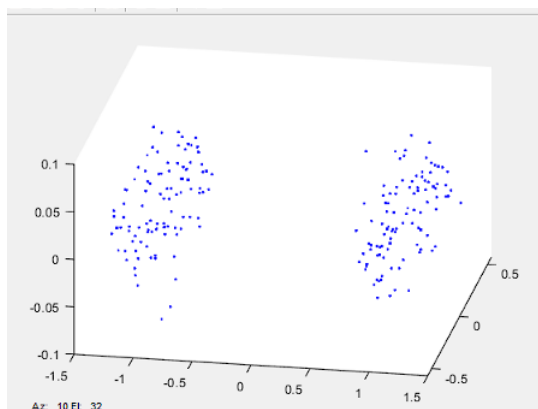
(Εικόνα 14) Δεδομένα του σετ δεδομένων υπολειμμάτων αμινοξέων σε τρεις διαστατό χώρο από διαφορετικές γωνίες.

4.3 Παράδειγμα συνθετικού Δεδομένων ανομοιότητας δύο ομάδων

Το συγκεκριμένο σετ δεδομένων πρόκειται για ένα πίνακα με δεδομένα ανομοιότητας διαστάσεων 200*200 . Πρόκειται για ένα μεγάλο σε διάστασης σετ δεδομένων το οποίο βρίσκεται στο πρόγραμμα που θα χρησιμοποιηθεί το som toolbox. Σκοπός του είναι να δείξει σαν παράδειγμα πως λειτουργεί ο RSOM χωρίς να υπάρχει κάποιο παράρτημα της προέλευσής του από που μπορεί να είναι οπότε είναι αρκετά πιθανό το σετ δεδομένων να είναι συνθετικό. Επρόκειτο για ένα σετ δεδομένων το οποίο στον χώρο έχει δύο αρκετά ξεκάθαρες ομάδες (εικόνα 15) .Παρόλο της απλότητας του είναι ένα εξαιρετικό παράδειγμα για να δει κάποιος της διάφορες μεταξύ των δύο αλγόριθμους αλλά καθώς και της δυσκολίες μεταξύ που μπορεί να

συναντήσουν.





(Εικόνα 15) Δεδομένα του συνθετικού δοκιμαστικού σετ δεδομένων ανομοιότητας σε τρεις διαστατό χώρο από διαφορετικές γωνίες

5 Αποτελέσματα

Στο παρακάτω κομμάτι θα περιγραφή το πρακτικό κομμάτι της εργασία όπου θα δείξουμε της διαφορές μεταξύ του κανονικού SOM και RSOM μέσω διεξαγωγής παραδειγμάτων μεταξύ των σετ δεδομένων. Θα δειχθούν πως η δύο αλγόριθμοι διαχειρίζονται τα δεδομένα που τους δίνονται , θα εξετασθεί η ποιότητα των δεδομένων για να δειχθούν η διαφορές τους στην διαχείριση των δεδομένων. Όπως προαναφέρθηκε τα δεδομένα μας πρόκειται για δεδομένα σε πίνακες ανομοιότητας όπου δείχνουν της διαφορές μεταξύ τους χωρίς να υπάρχει κάποιος τίτλος η αλλιώς ταμπέλα για το κάθε δεδομένο . Ο κύριος λόγος που χρησιμοποιούμε πίνακες ανομοιότητας εννοείται πως είναι λόγω του RSOM μιας και με αυτά τα δεδομένα λειτουργεί και ο κυρίως λόγος που φτιάχτηκε είναι για να διαχειρίζεται τέτοιου είδους δεδομένα. Κυρίως σκοπός αυτού του κομματιού της εργασίας είναι να δούμε πόσο καλύτερη είναι η διαχείριση του σε σχέση με τον κανονικό αλγόριθμο αλλά και υπάρχει περίπτωση χρήσης όπου ο κανονικός som θα προτιμώταν. Εννοείται ότι τα προηγούμενα κομμάτια για τους αλγόριθμους αυτά μας είναι απαραίτητα για την κατανόηση των αποτελεσμάτων των αλγόριθμων. Παρακάτω θα εμβαθύνουμε στο λογισμικό που θα χρησιμοποιηθεί καθώς και θα δειχθούν τα παραδείγματα των αποτελεσμάτων.

5.1 Λογισμικό που θα χρησιμοποιηθεί

Το λογισμικό που θα χρησιμοποιηθεί πρόκειται για το SOM toolbox πάνω στην matlab. Πρόκειται για μια βιβλιοθήκη της matlab η οποία γράφτηκε από το πανεπιστήμιο τεχνολογίας του helsinki το 1997 . Είναι μια βιβλιοθήκη η οποία παρέχεται δωρεάν και γράφτηκε από τους Esa Alhoniemi, Johan Himberg, Juha Parhankangas and Juha Vesanto. Επίσης έχουν γίνει με τους καιρούς και συνεισφορές κομματιών κώδικα εκ των οποίων ένα από αυτά είναι και το rsom. Να τονιστεί ότι η βιβλιοθήκη δεν παρέχεται μαζί με την matlab και πρέπει να εγκατασταθεί . Ο λόγος της δημιουργίας του SOM toolbox ξεκίνησε για να καλύψει ένα κενό για τα λογισμικά τα οποία υπήρχαν εκείνη την περίοδο. Για παράδειγμα μια από της επιλογές που είχε κάποιος ήταν ένα δωρεάν λογισμικό ονόματος SOM_PAK.(34) Το λογισμικό SOM PAK γράφτηκε το 1992 σε ANSI C από το πανεπιστήμιο τεχνολογίας του helsinki (44). Το πρόβλημα του ήταν ότι δεν ήταν όσο ευέλικτο και με ποικιλία σε τροποποίησης όσο η matlab. Αλλά ταυτόχρονα το neural networks toolbox με το οποίο ερχόταν η matlab είδη εγκατεστημένο δεν ήταν όσο εξελιγμένο όσο θα ήθελαν. Με αυτήν την λογική δημιουργήθηκε η βιβλιοθήκη SOM toolbox μιας και κάποιος θα μπορούσε να τροποποίηση τον κώδικα στην matlab για να καλύψει της ανάγκες του. Ταυτόχρονα επίσης μπορούσαν να χρησιμοποιήσουν τα εργαλεία της matlab για την εικονοποίηση των δεδομένων και όχι μόνο. Το συγκεκριμένο λογισμικό είναι διαθέσιμο στην σελίδα του πανεπιστημίου στο (46) στο οποίο περιέχεται η οδηγίες του προγράμματος καθώς και παραπάνω πληροφορίες για αυτό. Στην συγκεκριμένη εργασία θα χρησιμοποιηθεί η εκδοχή 2.1 του προγράμματος το οποίο βρίσκεται κοινοποιημένο για κατέβασμα στο github λόγω του ότι είναι η πιο πρόσφατη εκδοχή η οποία περιέχει κομμάτια τα οποία έχουν συνεισφέρει χρήστες και ένα από αυτά τυχαίνει να είναι ο κώδικας για τον RSOM(34). Ο κώδικας για το RSOM

συνεισφέρθηκε στο λογισμικό το 2012 από τον Alexander Schulz από το πανεπιστήμιο biefeld της Γερμανίας. Η συνεισφορά περιέχει τα εργαλεία για να τρέξει κάποιος την παραλλαγή RSOM και να την δείξει και θα εξηγηθεί παρακάτω περισσότερο.(45)

5.2 Αποτελέσματα στο σετ δεδομένων iris

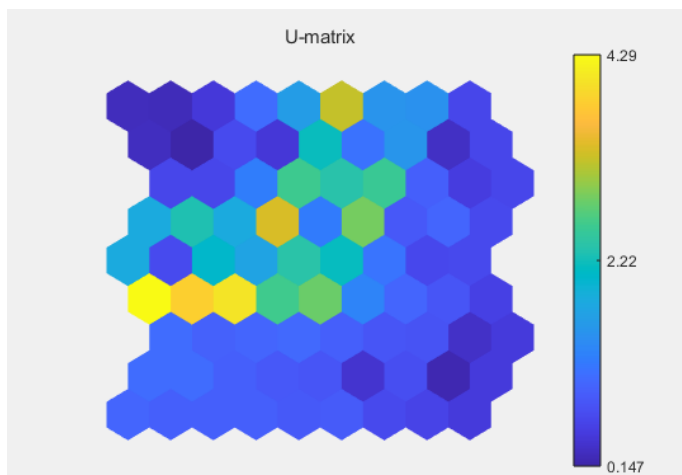
Ξεκινώντας με το παράδειγμα για τα δεδομένα του γνωστού σετ δεδομένων iris το οποίο έχουμε μεταφράσει σε πίνακα αποστάσεων . Αρχικά θα ξεκινήσουμε με ένα παράδειγμα το οποίο θα έχει μέγεθος χάρτη 5x5 με εξαγωνικό πλέγμα αλλά και με τετραγωνικό πλέγμα. Σκοπός είναι αρχικά να βρούμε πιο από τα δυο πλέγματα θα μας δείξει με καλύτερο τρόπο την πληροφορία αλλά και να βρούμε πιο μέγεθος χάρτη είναι κατάλληλο . Με σκοπό εννοείται παρόμοιες ρύθμισης μεταξύ των αλγορίθμων να εξετασθούν και στους δυο αλγόριθμους. Ξεκινώντας με τον κανονικό SOM αρχικά να θυμίσουμε το σχήμα των δεδομένων . Όπως γνωρίζουμε το αρχικό σετ δεδομένων του iris πρόκειται για 3 ομάδες από τις οποίες η δύο είναι αρκετά όμοιες με μια να ξεχωρίζει . Παρόλα αυτά μετά την μετατροπή των δεδομένων μπορούσαμε να δούμε καθαρά μόνο δύο ομάδες με μια από τις ομάδες να αποτελείται από δύο ομάδες ενωμένες. Ενώ η τρίτη ομάδα ξεχωρίζει πιο ξεκάθαρα. Έχοντας λοιπόν αυτά στο μυαλό μας προχωράμε στην πρώτη ρύθμιση του παραδείγματος .

5.2.1 Μέγεθος χάρτη 5x5 normal som

Αρχικά ας ξεκινήσουμε με ένα μικρό σε διαστάσεις παράδειγμα ενός 5x5 σε διαστάσεις χάρτη νευρώνων σε σχήμα πλέγματος εξαγώνο. Ο αλγόριθμος έχει ρυθμιστεί να τρέχει για το 60% σε τραχύ προσαρμογή δηλαδή να κάνει μεγάλες αλλαγές και στο υπόλοιπο κομμάτι σε ακριβή προσαρμογή έτσι ώστε να προσαρμοστεί πιο σωστά. Τέλος ο χάρτης εκπαιδεύεται για 500 επαναλήψεις που δηλαδή αυτό μας κάνει μία 300-200 τραχύ προς ακριβή εκπαίδευση . Η επιλογή 500 επαναλήψεων έγινε μετά από αρκετά τεστ και βρέθηκε ότι μετά από αυτά η διαφορά σε αποτελέσματα ήταν από μηδαμινή έως μη ανιχνεύσιμη και τα αποτελέσματα που περνάμε ήταν αρκετά όμοια μεταξύ τους.

Commented [GU1]: ελληνικά

Commented [GU2]: ελληνικά

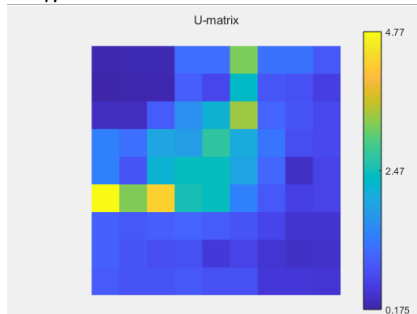


(Εικόνα 16) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 5x5, σε κανονικό som, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

Στην παραπάνω εικόνα(εικόνα 16) μπορούμε να δούμε το σχεδιάγραμμα του umatrix το οποίο πρόκειται για της αποστάσεις των νευρώνων μεταξύ τους. Όπως εξηγήθηκε και προηγουμένως σκοπός του είναι να μας δήσει τα αποτελέσματα του εκπαιδευμένου χάρτη σε πληροφορία που μπορούμε να καταλάβουμε πιο ευκολά. Συγκεκριμένα στα σχεδιαγράμματα ο διαχωρισμός των ομάδων γίνεται με της απόστασης των νευρώνων με μεγάλες αποστάσεις να θεωρούνται ως όρια ομάδων ενώ μικρότερες αποστάσεις μεταξύ των νευρώνων να θεωρούνται το σώμα μιας ομάδας. Στα σχεδιαγράμματα όπως και η μπάρα στα δεξιά τους συνιστά με κίτρινο σηματοδοτούμαι μεγάλες αποστάσεις μεταξύ των νευρώνων ενώ με μπλε μικρότερες αποστάσεις μεταξύ τους. Σε αυτό το παράδειγμα μπορούμε να δούμε δύο ξεκάθαρες ομάδες κάτι που είχαμε δει και στην ανάλυση των δεδομένων αλλά δεν έχει καταφέρει να γίνει διαχωρισμός της τρίτης ομάδας. Πράγμα που μας δείχνει ότι σε αυτό το παράδειγμα ο αλγόριθμος απέτυχε να διαχωρίσει και της τρεις ομάδες βρίσκοντας μόνο της δυο από αυτές.

Παρακάτω (εικόνα 17) είναι ένας ίσος σε μήκος χάρτης με της ίδιες ρυθμίσεις αλλά σε τετραγωνικό αντί για εξαγωνικό πλέγμα με σκοπό να δούμε της διαφορές μεταξύ των

πλεγμάτων.

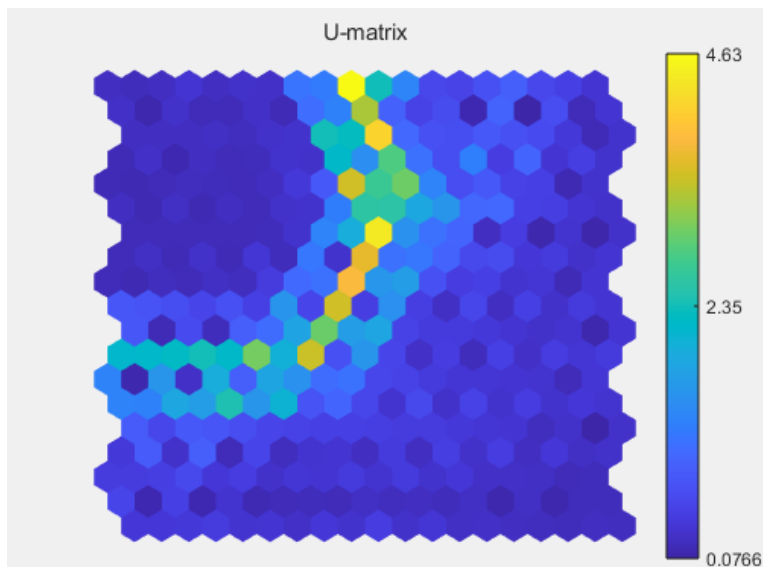


(Εικόνα 17) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 5x5, σε κανονικό som, σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα.

Γενικότερα βλέπουμε ότι τα αποτελέσματα μεταξύ των δυο παραδειγμάτων είναι όμοια μεταξύ τους χωρίς κάποια ενδιαφέρων διαφορά μεταξύ τους.

5.2.2 Μέγεθος χάρτη 10x10 normal som

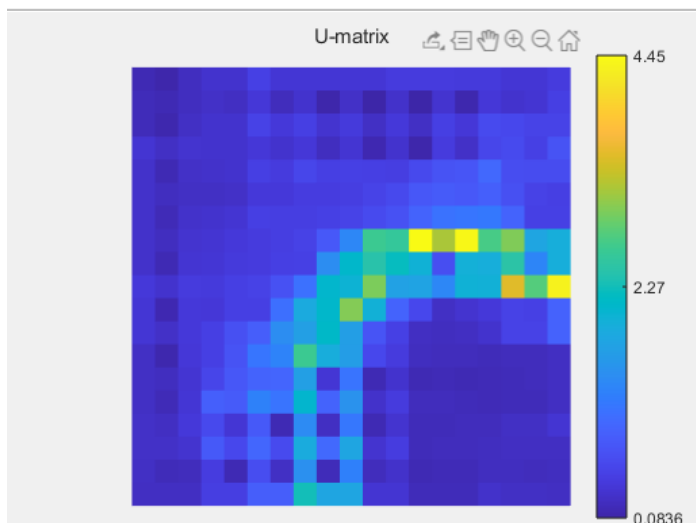
Προχωρώντας με το παράδειγμα του iris dataset θα συνεχίσουμε με ένα μεγαλύτερο σε διαστάσεις χάρτη. Ένα από τα προβλήματα που γενικότερα θα παρατηρηθούν είναι ότι τα αποτελέσματα και η ποιότητά τους μπορεί όχι μόνο να επηρεαστεί από αν ο χάρτης μας είναι πολύ μικρός αλλά και αν είναι πολύ μεγάλος. Αυτό ονομάζεται ‘underfitting’ και ‘overfitting’ αντίστοιχος και είναι κάτι το οποίο μπορεί να εμφανιστεί όχι μόνο στον αριθμό των νευρώνων αλλά και στο ποσό πολύ εκπαιδεύονται τα δεδομένα. Παρακάτω (εικόνα 18) είναι τα αποτελέσματα του αλγορίθμου σε σχεδιάγραμμα του umatrix.



(Εικόνα 18) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 10x10, σε κανονικό som, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

Παρόλο που αυξήθηκε στο παρών παράδειγμα το μέγεθος του χάρτη βλέπουμε ότι δεν υπάρχει κάποια μεγάλη διαφορά στα αποτελέσματα παίρνουμε. Η μεγαλύτερη διαφορά σε σχέση με το προηγούμενο παράδειγμα είναι η ευκρίνεια του χάρτη στον οποίο μπορούμε να δούμε τα όρια μεταξύ των δύο ομάδων πιο ξεκάθαρα όμως βλέπουμε ότι συνεχίζεται να μην βλέπουμε την τρίτη ομάδα με την δεύτερη μεγαλύτερη ομάδα στον χάρτη στην οποία θα περιμέναμε να δούμε την τρίτη ομάδα σύμφωνα με την ανάλυση των δεδομένων, να φαίνεται να είναι αρκετά όμοια διασπαρμένη στις αποστάσεις μεταξύ της.

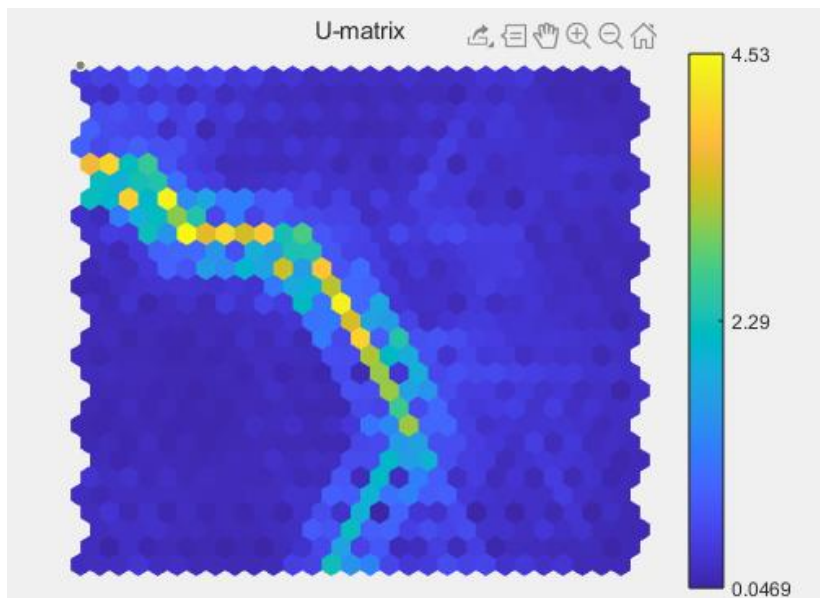
Στο τετραγωνικό πλέγμα παρακάτω (εικόνα 19) επίσης δεν βλέπουμε κάποια μεγάλη διαφορά στα αποτελέσματα που παίρνουμε από το umatrix. Με μόνο δύο ομάδες να φαίνονται ξεκάθαρα.



(Εικόνα 19) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 10x10, σε κανονικό som, σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα.

5.2.3 μέγεθος χάρτη 15x15 normal som

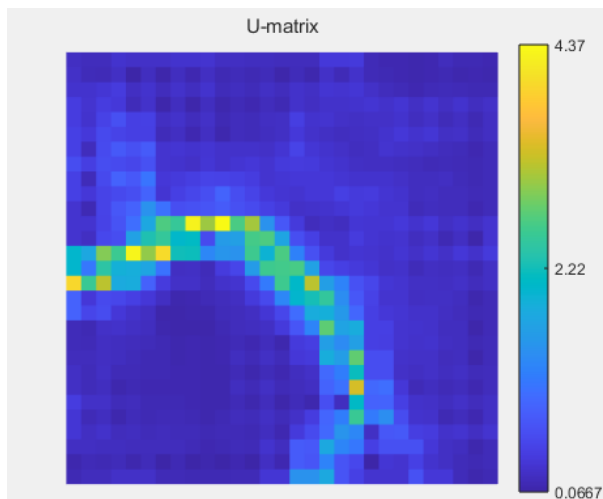
Τέλος για το παράδειγμα στον κανονικό som βλέπουμε ένα παράδειγμα σε διαστάσεις χάρτη 15x15 (εικόνα 20). Λόγω της διαστάσεις του μπορούμε να δούμε τα δεδομένα σε μεγαλύτερη ευκρίνεια και με περισσότερη λεπτομέρεια . Αν και από όσο φαίνεται δεν έχει αλλάξει κάτι ουσιαστικό στα δεδομένα σε σχέση με τον προηγούμενο χάρτη. Οπότε μπορούμε να πούμε ότι πέρα από το προηγούμενο μέγεθος μπορεί μεν να βελτιώνει την εικόνα αλλά δεν μας δείχνει κάποια ουσιαστική αλλαγή.



(Εικόνα 20) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 15x15, σε κανονικό som, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

μέγεθος χάρτη 15x15 σε τετραγωνικό πλέγμα

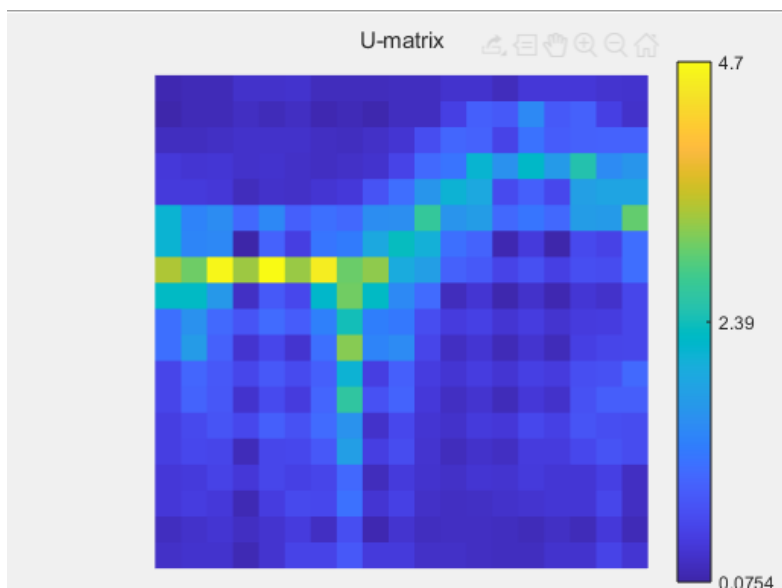
Στο τετραγωνικό πλέγμα(εικόνα 21) δεν βλέπουμε πάλι μεγάλες διαφορές στα δεδομένα μας το κανονικό umatrix με της δύο πιο κύριες ομάδες να φαίνονται ξεκάθαρα . Κάτι όμως που αρχίζουμε να βλέπουμε είναι της δύο ομάδες στην μια μεγαλύτερη ομάδα να αρχίζουν να αλλοιώνονται και να μην φαίνονται τόσο ξεκάθαρα στο τετραγωνικό πλέγμα. Κάτι που όπως προαναφέρθηκε μπορεί και να οφείλεται στο ότι αρχίζουμε να κάνουμε των χάρτη overfit . Με αποτέλεσμα να μην βρίσκει πια μοτίβα αλλά να μαθαίνει τα δεδομένα μας .



(Εικόνα 21) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 15x15, σε κανονικό som, σε επίπεδο σχήμα, σε τετραγωνικό πλέγμα.

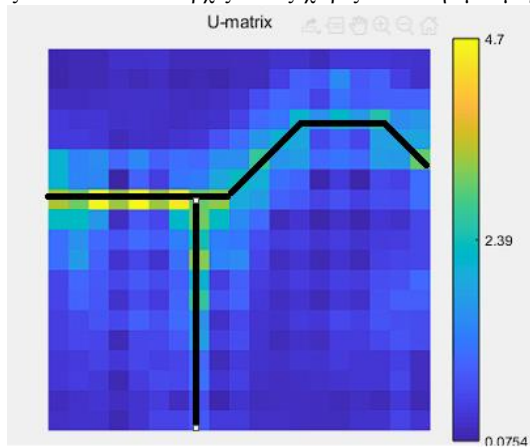
5.2.4 Μέγεθος χάρτη 10x10 normal som σε κυλινδρικό σχήμα

Παρακάτω (εικόνα 22) βλέπουμε μια διαφορετική προσέγγιση στο πρόβλημα. Ως τώρα βλέπουμε ότι ο αλγόριθμος δυσκολεύεται να ξεχωρίσει την τρίτη ομάδα στο σετ δεδομένων μας παρόλο της αλλαγής του πλέγματος και μεγέθους του χάρτη. Οπότε μπορεί το πρόβλημα να βρίσκεται στο σχήμα του χάρτη αντί στους νευρώνες ή στον αλγόριθμο. Παρακάτω φαίνονται τα αποτελέσματα του αλγορίθμου αλλά αντί το επίπεδο των νευρώνων να είναι σε επίπεδο σχήμα είναι σε σχήμα κυλίνδρου.



(Εικόνα 22) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 10x10, σε κανονικό som, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

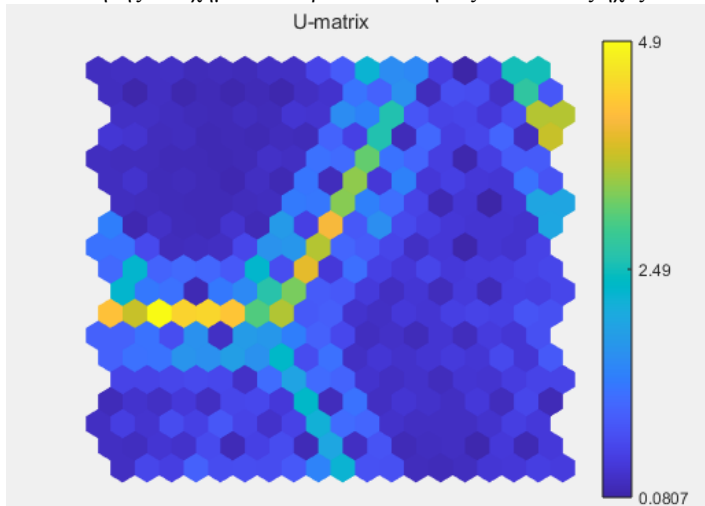
Τα αποτελέσματα όπως βλέπουμε είναι πολύ πιο ξεκάθαρα αφού βλέπουμε η τρίτη ομάδα να ξεκολλάει και να αρχίζει να ξεχωρίζει από την μια μεγάλη που είχαμε στα παραδείγματα .



(Εικόνα 23) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 10x10, σε κανονικό som, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα. Με μαύρες γραμμές για να τονιστούν καλύτερα τα όρια μεταξύ των ομάδων.

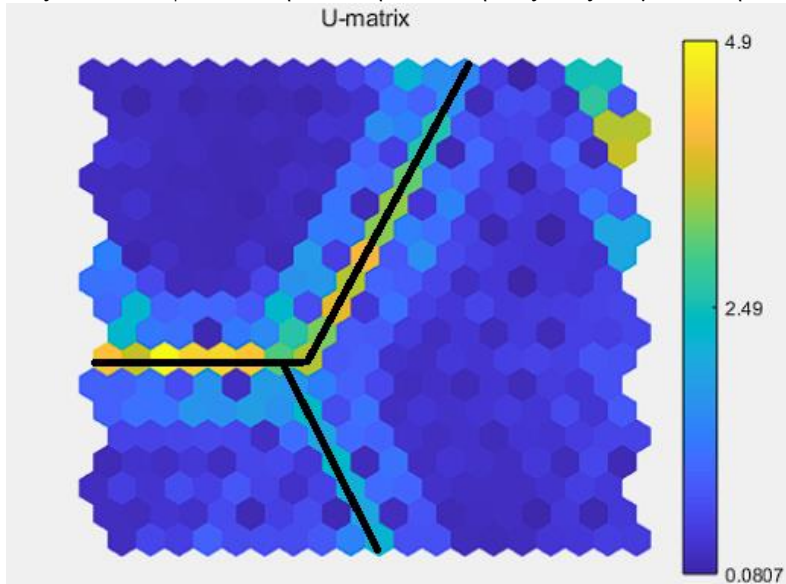
Συγκεκριμένα στην παραπάνω εικόνα (εικόνα 23) μπορούμε να δούμε με μαύρη χρωματισμένη γραμμή τα όρια στα οποία βλέπουμε ότι δεν μας δείχνουν μόνο δύο ομάδες. Αλλά βλέπουμε για πρώτη φορά και των διαχωρισμό μιας τρίτης ομάδας στα δεδομένα η οποία όπως και στην ανάλυση των δεδομένων βλέπουμε ότι βρίσκεται μέσα στην μεγαλύτερη ομάδα.

Τα δεδομένα βλέπουμε επίσης (εικόνα 24) να συμφωνούνε μεταξύ τους και για το εξαγωνικό πλέγμα με κυλινδρικό σχήμα χάρτη μάλιστα φαίνεται πιο ξεκάθαρο. Η απομονωμένη ομάδα είναι συνήθης σε σχήμα κυλίνδρου και συνήθως είναι απλός ήχος



(Εικόνα 24) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 10x10, σε κανονικό som, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

Όπως και στο τετραγωνικό πλέγμα μπορούμε παρόμοιος παρακάτω (εικόνα 25) να δούμε τα όρια των ομάδων χρωματισμένα με μαύρη γραμμή. Τα οποία όρια δίνουν ξεκάθαρες τρεις ομάδες στα δεδομένα με τα όρια της τρίτης ομάδας να είναι φυσικά πιο απαλά και κοντά μεταξύ τους. Κάτι που οφείλεται στην κοντινή απόσταση τους όπως είδαμε και στην ανάλυση τους.



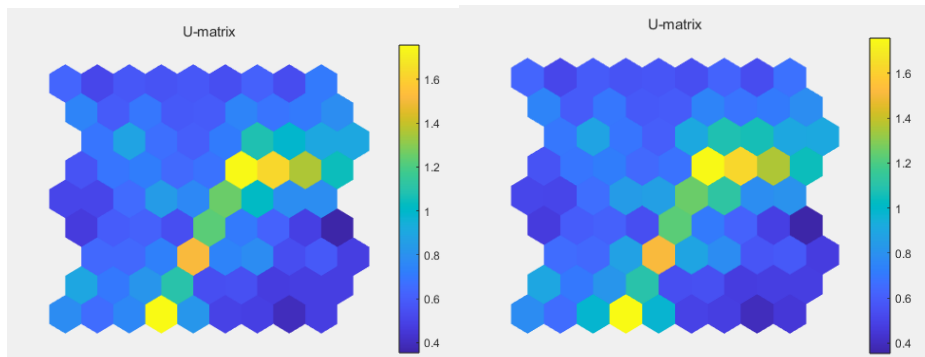
(Εικόνα 25) Σχεδιάγραμμα umatrix σετ δεδομένων iris με μέγεθος χάρτη 10x10, σε κανονικό som, σε κυλινδρικό σχήμα, σε εξαγωνικό πλέγμα. Με μαύρες γραμμές για να τονιστούν καλύτερα τα όρια μεταξύ των ομάδων.

Μπορούμε να δούμε μια αρκετή καθαρή εικόνα των τριών ομάδων οπότε με διαφορετικό σχήμα στον χάρτη μπορούμε να πούμε ότι ο κανονικός som κατάφερε να τα ομαδοποίηση με επιτυχία.

5.2.5 μέγεθος χάρτη 5x5 rsom σε επίπεδο σχήμα

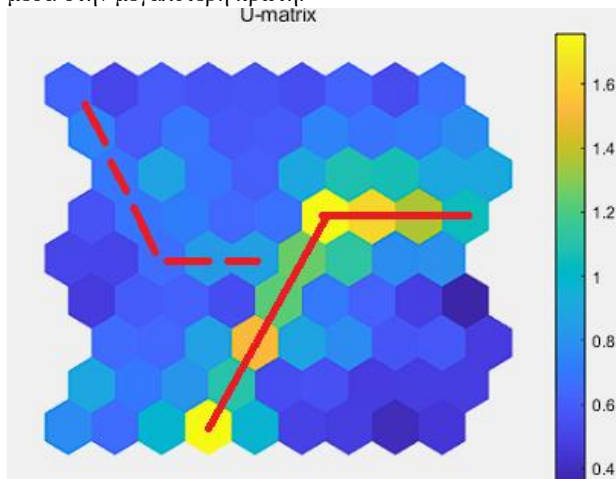
Προχωρώντας στο rsom ξεκινάμε τα παραδείγματα με αρκετά παρόμοιο τρόπο με τον κανονικό som. Η εκπαίδευση που κάνουμε είναι επίσης 500 επαναλήψεις όπως και με τον κανονικό som σε batch setting σε διαστάσεις χάρτη 5x5. Το umatrix το οποίο θα παρουσιάζεται θα είναι σε δύο μορφές η οποίες διαφέρουν στο πως δείχνονται η αποστάσεις μεταξύ των νευρώνων. Η μορφή αυτές είναι σε median και mean δηλαδή σε διάμεση απόσταση και μέση απόσταση. Η διαφορά τους είναι κυρίως ότι η μέση απόσταση πολλές φορές κάνει πιο έντονες της αποστάσεις μεταξύ των νευρώνων και μας βοηθάει να βρούμε ομάδες καλύτερα.

Commented [GU3]: στην λεζάντα



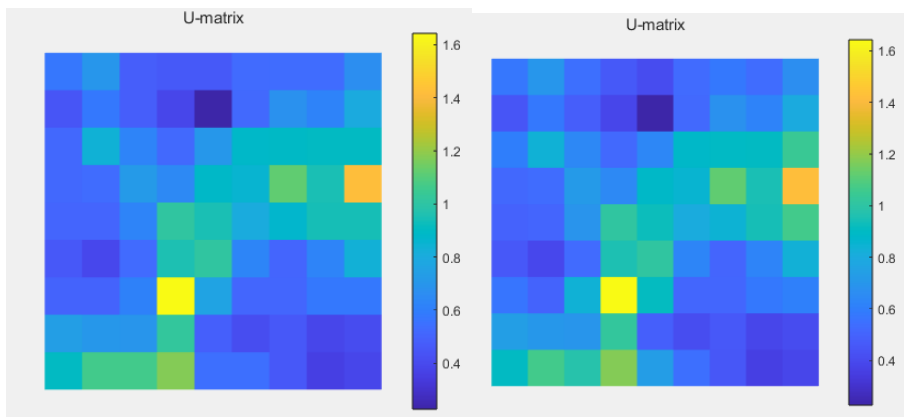
(Εικόνα 26) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

Στα πρώτα αποτελέσματα (Εικόνα 26) βλέπουμε το umatrix σε διάμεση και μέση απόσταση αντιστοίχως και όπως και με τον κανονικό som εύκολα μπορούμε να παρατηρήσουμε την μια από της δύο ομάδες αρκετά καθαρά. Όμως μπορούμε επίσης να δούμε και μια δεύτερη ομάδα μέσα στην μεγαλύτερη πρώτη.



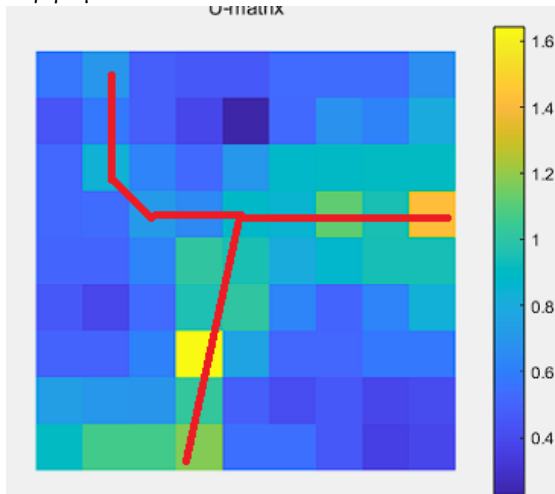
(Εικόνα 27) Σχεδιάγραμμα umatrix σετ δεδομένων iris σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα. Με κόκκινες γραμμές για να τονιστούν καλύτερα τα όρια μεταξύ των ομάδων και με διακεκομμένες γραμμές για τα πιο απαλά όρια.

Με κόκκινο χρώμα έχει σημειωθεί στην παραπάνω εικόνα (εικόνα 27) η κύρια διαχωριστική γραμμή και με διακεκομμένες μπορούμε να δούμε μια λίγο πιο έντονη απόσταση μεταξύ των νευρώνων αυτό μπορούμε να το πάρουμε σαν ένα σημάδι για μια ομάδα ανάμεσα στην ομάδα κάτω που είδαμε ότι ισχύει και στον κανονικό som αλλά εδώ πέρα ακόμα και στο βασικό επίπεδο σχήμα και σε πολύ μικρότερες διαστάσεις από ότι ο κανονικός som ξεκίνησε να μας δίνει αποτελέσματα.



(Εικόνα 28) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα.

Αλλάζοντας το πλέγμα σε τετραγωνικό φαίνεται να συμφωνεί με τα δεδομένα παραπάνω(εικόνα 28) αλλά κάνει και την δεύτερη ομάδα που παρατηρείτε πάνω να φαίνεται ακόμα πιο έντονα. Με αυτά τα αποτελέσματα μπορούμε είδη να δούμε της διαφορές στις δυνατότητες των δύο αλγορίθμων.

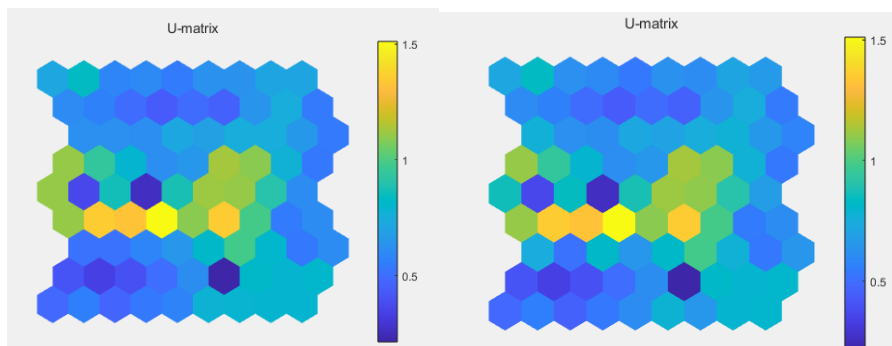


(Εικόνα 29) Σχεδιάγραμμα umatrix σετ δεδομένων iris σε διάμεσες αποστάσεις , με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα . Με κόκκινες γραμμές για να τονιστούν καλύτερα τα όρια αναμεσά στις ομάδες τους.

Όπως παραπάνω (εικόνα 29) με κόκκινη γραμμή πάλι σηματοδοτούμε τα όρια των ομάδων στον χάρτη.

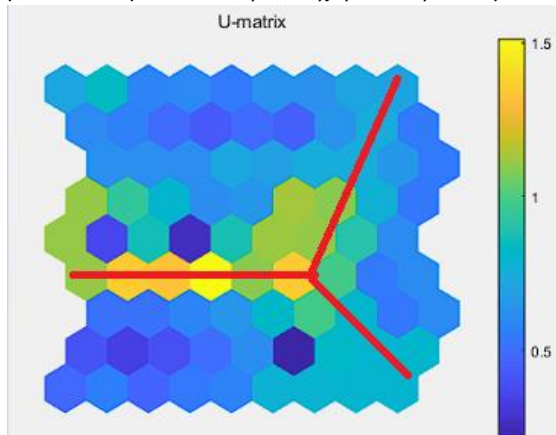
5.2.6 μέγεθος 5x5 rsom σε κυλινδρικό σχήμα

Προχωρώντας σε παρόμοιο παράδειγμα αλλά με σχήμα χάρτη σε κύλινδρο το οποίο στον κανονικό som έκανε μια μεγάλη διαφορά συνεχίζουμε να εξερευνούμε για διαφορές και αν βοηθάει πραγματικά στα δεδομένα.



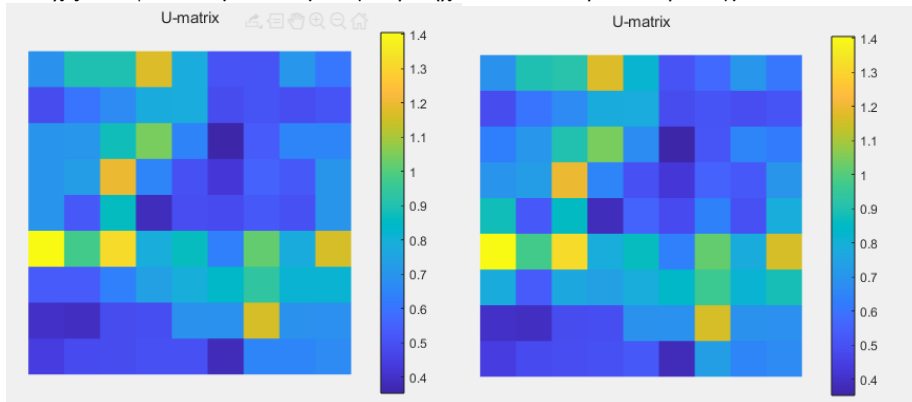
(Εικόνα 30) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

Ξεκινώντας λοιπόν με τα πρώτα δεδομένα βλέπουμε (εικόνα 30) ότι στο εξαγωνικό πλέγμα εντοπίζονται η δυο καθαρές ομάδες αλλά ταυτόχρονα η τρίτη ομάδα εντοπίζεται πιο καθαρά μάλιστα ακόμα και το κύριο διαχωριστικό βλέπουμε ότι ξεκινάει να τινί προς αυτό.



(Εικόνα 31) Σχεδιάγραμμα umatrix σετ δεδομένων iris σε διάμεσες αποστάσεις , με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα ,σε κυλινδρικό πλέγμα . Με κόκκινες γραμμές για να τονιστούν καλύτερα τα όρια αναμεσά στις ομάδες τους.

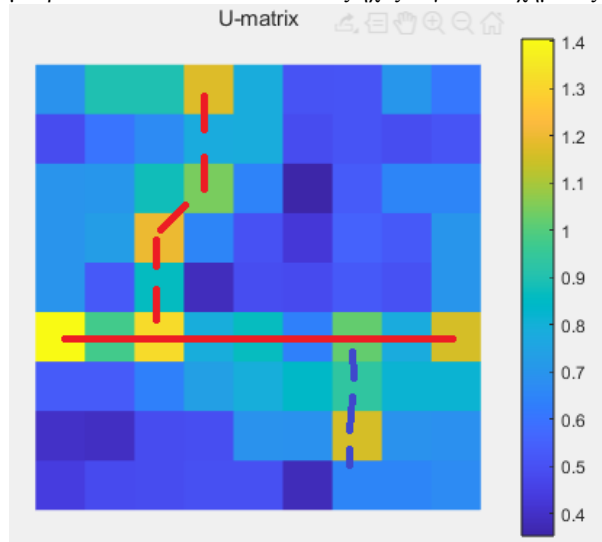
Κλασσικά βλέπουμε στην παραπάνω εικόνα (εικόνα 31) με κύρια κόκκινη γραμμή τον διαχωρισμό αλλά ταυτόχρονα ο ίδιος ο διαχωρισμός των κύριων ομάδων τείνει προς τον διαχωρισμό της τρίτης ομάδας αναμεσα στην κύρια. Οπότε το μοτίβο του κυλινδρικού χάρτη συνεχίζει να φαίνεται μάλιστα με λιγότερο ήχο από ότι το πρώτο παράδειγμα .



(Εικόνα 32) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε tsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

Τα αποτελέσματα στην παραπάνω εικόνα (εικόνα 32) στον τετραγωνικό χάρτη ξεκινάνε να δείχνουν διαφορετικά αποτελέσματα από ότι στον εξαγωνικό. Γενικότερά βλέπουμε και στο προηγούμενο παράδειγμα αλλά και σε αυτό με το εξαγωνικό πλέγμα ότι η ομάδες παρόλο ξεκάθαρες ήταν κυρίως ίσες σε αποστάσεις (αλλιώς χρώμα) μέσα τους όμως ξεκινώντας από τον τετραγωνικό μπορούμε να δούμε ότι η δύο κυρίες ομάδες είναι έντονες ενώ η τρίτη ομάδα ξεκινάει όχι μόνο να μην φαίνεται ξεκάθαρα αλλά ακόμα και να αρχίσει να φαίνεται και μια τέταρτη πολύ μικρότερη ομάδα κάτω δεξιά χρωματισμένη με μπλε στην παρακάτω εικόνα(εικόνα 33).Βεβαία είναι σημαντικό να αναφερθεί όπως και στον κανονικό som ότι θα

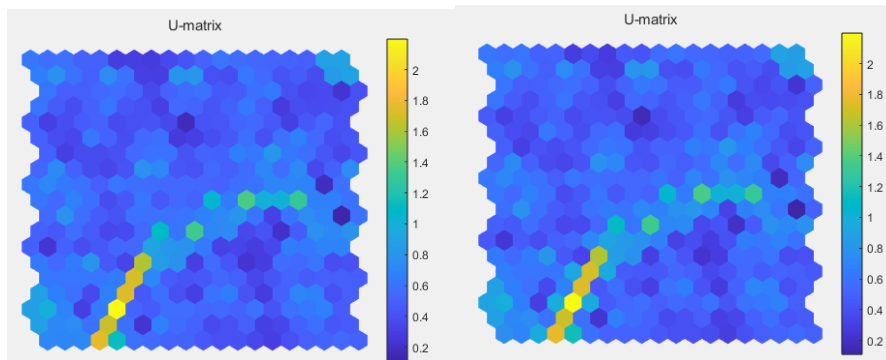
μπορούσε κάλλιστα να είναι απλός ήχος λόγω του σχήματος.



(Εικόνα 33) Σχεδιάγραμμα umatrix σετ δεδομένων iris σε διάμεσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα. Με κόκκινες γραμμές για να τονιστούν καλύτερα τα όρια μεταξύ των ομάδων και με διεκεκομμένες γραμμές για της δύο δευτερέων ομάδες που εμφανίζονται.

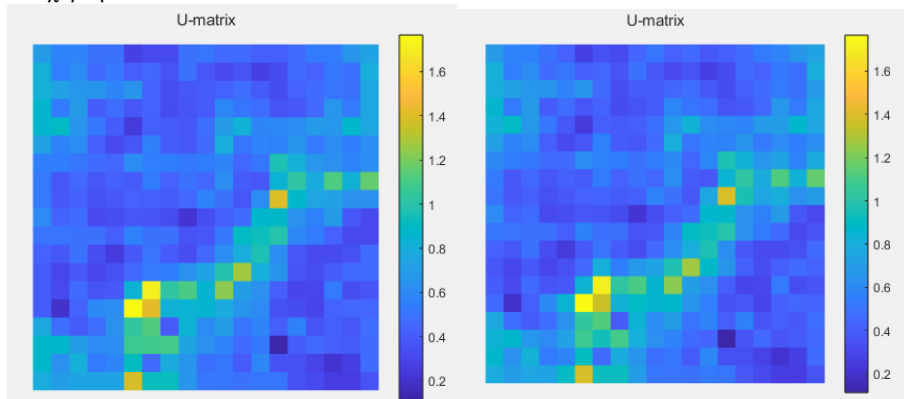
5.2.7 Μέγεθος χάρτη 10x10 rsom

Προχωρώντας σε μεγαλύτερες διαστάσεις ξεκινάμε για πρώτη φορά να βλέπουμε (εικόνα 34) διαφορές στα αποτελέσματα που παίρνουμε από το rsom. Στα umatrix παρακάτω βλέπουμε ότι ακόμα και η δύο καθαρές ομάδες που είχαμε αρχικά έχουν ξεκινήσει να ενώνονται μεταξύ τους με την τρίτη ομάδα να εξαφανίζεται ανάμεσα στην πρώτη στο πλήθος.



(Εικόνα 34) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

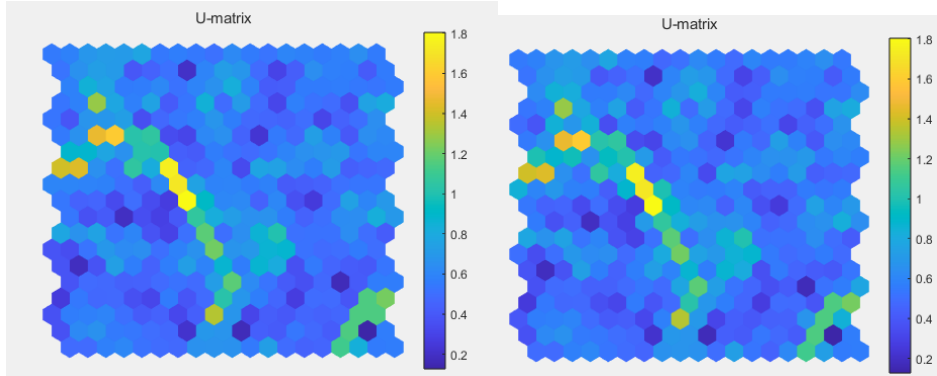
Τα αποτελέσματα δεν είναι πολύ καλύτερα σε τετραγωνικό πλέγμα . Αρχικά βλέπουμε (εικόνα 35) ότι η δύο ομάδες έχουν διαχωριστεί καλύτερα αλλά ταυτόχρονα η τρίτη ομάδα είναι και πάλι μπερδεμένη μέσα στην πρώτη μεγάλη. Ξεκινάμε να βλέπουμε ότι σε μεγαλύτερους χάρτες ο rsom έχει μεγαλύτερη δυσκολία να δείξει αποτελέσματα κάτι που είναι το αντίθετο με τον κανονικό som όπου η μεγαλύτερη δυσκολία βρισκόταν στα μικρά σε διάστασης αποτελέσματα του χάρτη.



(Εικόνα 35) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

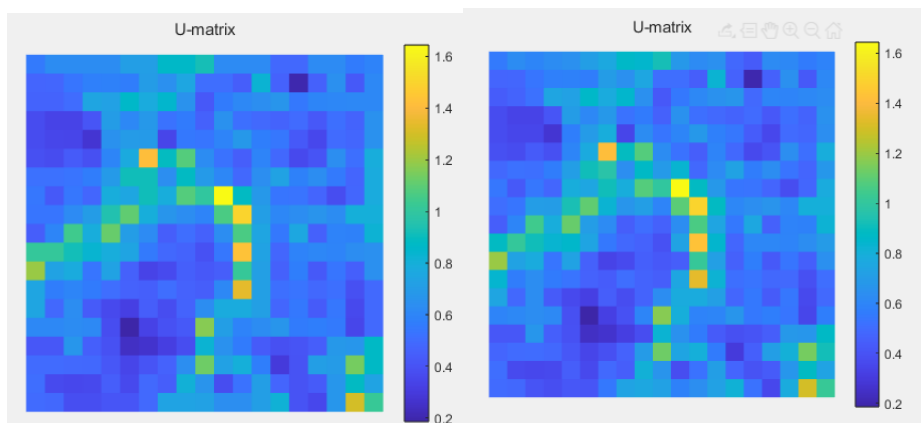
5.2.8 Μέγεθος 10x10 rsom σε κυλινδρικό σχήμα

Προχωρώντας στο κυκλικό σχήμα του χάρτη νευρώνων συνεχίζουμε να βλέπουμε πιο μπερδεμένα δεδομένα από ότι με το μικρότερο παράδειγμα στο εξαγωνικό πλέγμα. Παρόλο που βλέπουμε στην παρακάτω εικόνα (εικόνα 36) και πάλι πιο καθαρά της δύο ομάδες από ότι ήταν στο εξαγωνικό πλέγμα σε παράλληλο σχήμα συνεχίζουμε να βλέπουμε μια ένωση μεταξύ των δύο ομάδων και την έλλειψη της τρίτης καθαρής ομάδας. Επίσης ξεκινάμε να βλέπουμε artifacts του κυκλικού χάρτη το οποίο φαίνεται συχνά στην κάτω δεξιά γωνία. Οπου βλέπουμε μικρές γωνιακές ομάδες.



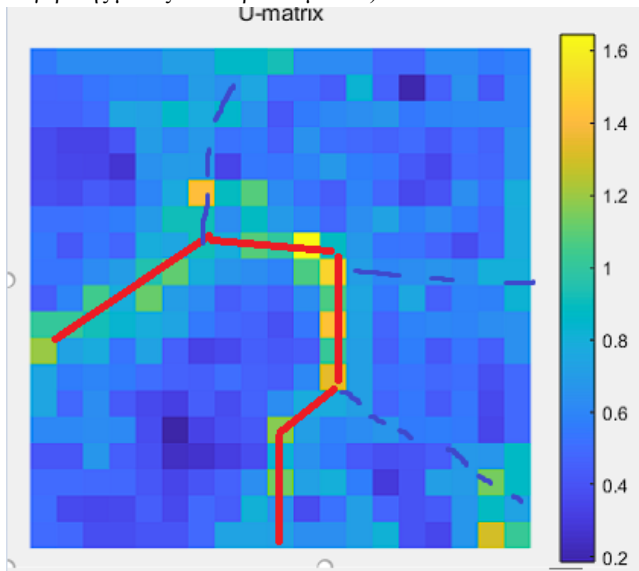
(Εικόνα 36) Σχεδιαγράμματα umatrix σε δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε κυλινδρικό σχήμα, σε εξαγωνικό πλέγμα.

Σε τετραγωνικό πλέγμα βλέπουμε στην παρακάτω εικόνα (εικόνα 37) λίγο πιο καθαρά των διαχωρισμό των ομάδων, ταυτόχρονα βλέπουμε και ξεκινάμε να βλέπουμε και περισσότερο δευτερεύον ομάδες στην κυρία μεγάλη ομάδα όμως συνεχίζουμε να μην έχουμε όσο καθαρή εικόνα είχαμε σε μικρότερους σε διάστασης χάρτες.



(Εικόνα 37) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

Παρακάτω(εικόνα 38) βλέπουμε ξανά με κόκκινο των διαχωρισμό των κύριων ομάδων αλλά ταυτόχρονα με μπλε βλέπουμε και διάφορες μικρό ομάδες μέσα στην μεγάλη ομάδα, η πιο κυριά από αυτές είναι στα αριστερά αλλά ταυτόχρονα υπάρχουν και σημάδια από τα δεξιά της πρώτης ομάδας πράγμα που μας δείχνει ότι σιγουρά υπάρχουν και άλλες ομάδες (μιας και τεχνικά τα δεδομένα μας δεν αποτελούνται από 3 ομάδες καθαρά αλλά από 6 μιας και πρόκειται για σύγκρισης μεταξύ των τριών ομάδων)

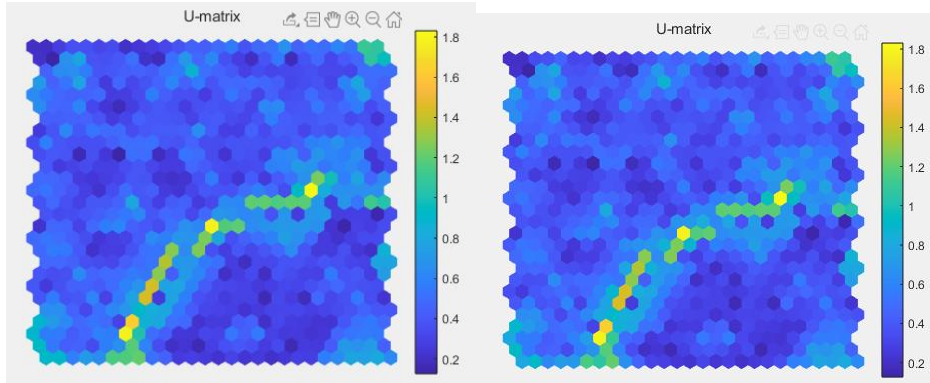


(Εικόνα 38) Σχεδιάγραμμα umatrix σετ δεδομένων iris σε διάμεσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα. Με κόκκινες γραμμές για να τονιστούν καλύτερα τα όρια μεταξύ των ομάδων και με διακεκομμένες γραμμές για της δύο δευτερεύων ομάδες που εμφανίζονται.

5.2.9 Μέγεθος 14x14 rsom σε επίπεδο σχήμα

Συνεχίζοντας με τα παραδείγματα στους μεγαλύτερους σε διαστάσεις χάρτες αρχικά να αναφερθεί ότι η εξαγωνική χάρτες θα είναι σε 14x14 διαστάσεις ενώ η τετραγωνική σε 15x15 λόγω του ότι η εξαγωνική δεν μπορούσαν να δειχθούν στην μεγαλύτερη διάσταση. Ξεκινώντας πάλι σε επίπεδο σχήμα βλέπουμε στην παρακάτω εικόνα (εικόνα 39) ότι τα δεδομένα μας έχουν συνεχίσει να κρατάνε της δυο καθαρές ομάδες με την τρίτη να χάνεται ακόμα πιο πολύ στην μεγαλύτερη ομάδα όποτε είναι πιο ασφαλές να ξεκινάμε να θεωρούμε ότι τα δεδομένα μας αρχίζουν να γίνονται overfit και ξεκινάνε απλώς να απομνημονεύουν τα δεδομένα αντί να βρίσκουν μοτίβο και να κάνουν σωστή ομαδοποίηση. Τα κενά πάνω στις δύο ομάδες αρχίζουν

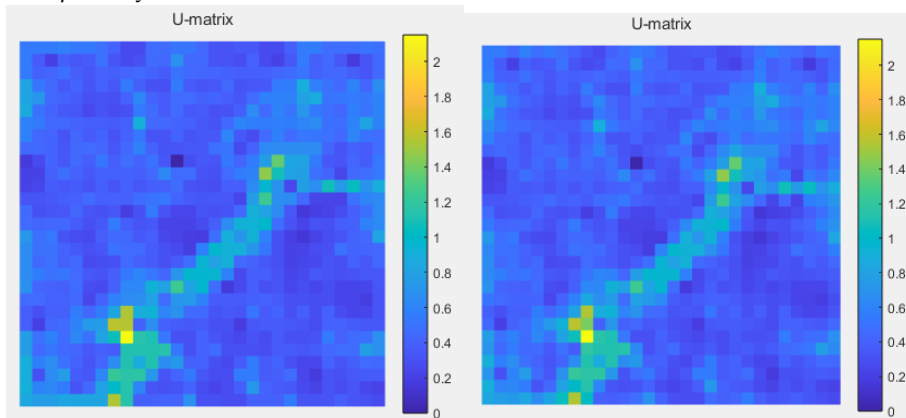
επίσης να γίνονται περισσότερα αλλά συνεχίζουμε ξεκάθαρα να της ξεχωρίζουμε.



(Εικόνα 39) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 14x14, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

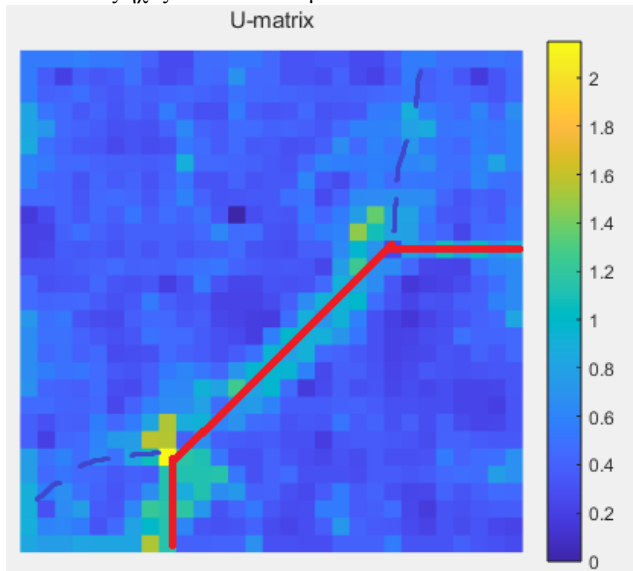
Μέγεθος χάρτη 15x15rsom

Σε τετραγωνικό πλέγμα τα αποτελέσματα παρακάτω (εικόνα 40) δεν βελτιώνονται και πολύ βλέπουμε ότι τα όρια και πάλι φαίνονται καθαρά στις δύο κύριες ομάδες με λιγότερα κενά πάνω στα όριά τους.



(Εικόνα 40) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 15x15, σε rsom, σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα.

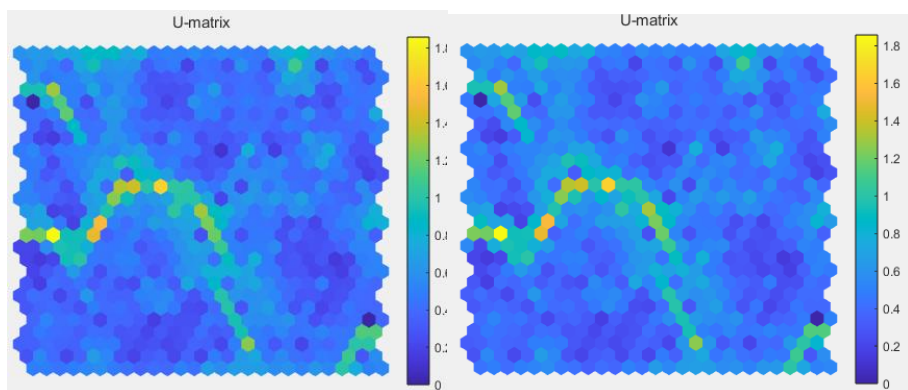
Ταυτόχρονα μπορούμε να παρατηρήσουμε στην παρακάτω εικόνα (εικόνα 41) δύο μικρότερες ομάδες ανάμεσα στην μεγάλη ομάδα όμως τα όρια είναι πάρα πολύ αχνά και θα μπορούσε να είναι απλός ήχος στα αποτελέσματα.



(Εικόνα 41) Σχεδιάγραμμα umatrix σετ δεδομένων iris σε διάμεσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 15x15, σε rsom, σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα. Με κόκκινες γραμμές για να τονιστούν καλύτερα τα όρια μεταξύ των ομάδων και με διακεκομμένες γραμμές για της δύο δευτερευόν ομάδες που εμφανίζονται.

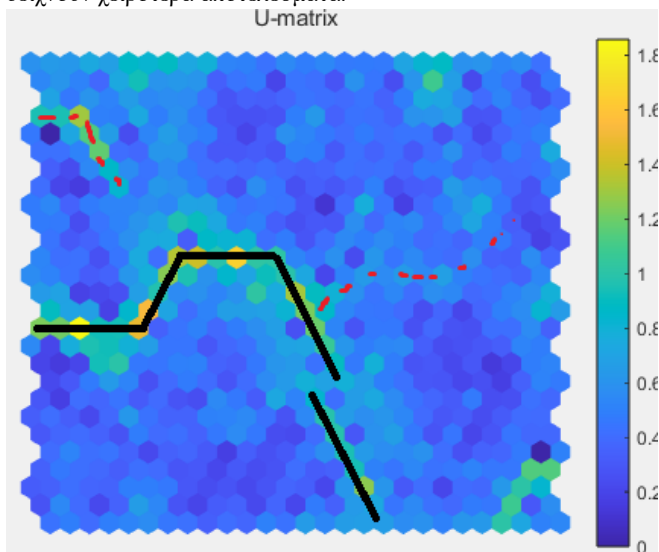
5.2.10 Μέγεθος χάρτη 14x14 rsom σε κυλινδρικό σχήμα

Προχωρώντας στο κυλινδρικό σχήμα οι δύο ομάδες φαίνονται πιο ξεκάθαρα (εικόνα 42) αλλά ταυτόχρονα ξεκινάμε να βλέπουμε μερικά μισοτελειωμένα όρια σε ομάδες τα , αλλά όπως και με τα προηγούμενα αποτελέσματα δεν φαίνεται κάποια πραγματική ομάδα με έναν εμφανή διαχωρισμό στα δεδομένα.



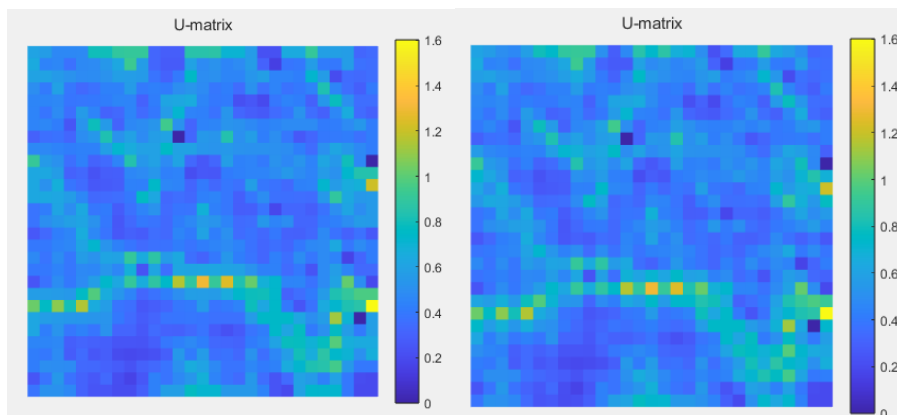
(Εικόνα 42) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 14x14, σε rsom, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

Γενικότερα βλέπουμε (εικόνα 43) ότι υπάρχουν ακόμα κάποιοι διαχωρισμοί στην μεγαλύτερη ομάδα αλλά συνεχίζεται να ακολουθείτε το μοτίβο των μεγαλύτερων διαστάσεων να μας δείχνουν χειρότερα αποτελέσματα.



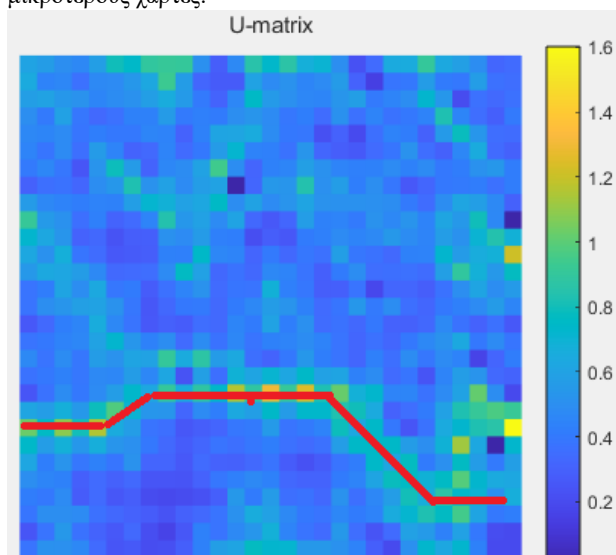
(Εικόνα 43) Σχεδιάγραμμα umatrix σετ δεδομένων iris σε διάμεσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 14x14, σε t-sne, σε κυλινδρικό σχήμα, σε εξαγωνικό πλέγμα. Με μαύρες γραμμές για να τονιστούν καλύτερα τα όρια μεταξύ των ομάδων και με διακεκομμένες γραμμές για της δύο δευτερεύων ομάδες που εμφανίζονται.

Τα αποτελέσματα (εικόνα 44) από το τετραγωνικό πλέγμα σε διάστασης 15x15 δεν είναι τόσο διαφορετικά επίσης με την κυρία διαφορά να είναι ότι η μικρότερη κυρία ομάδα τώρα καταλαμβάνει ολόκληρο το κάτω μέρος του umatrix.



(Εικόνα 44) Σχεδιαγράμματα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 15x15, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

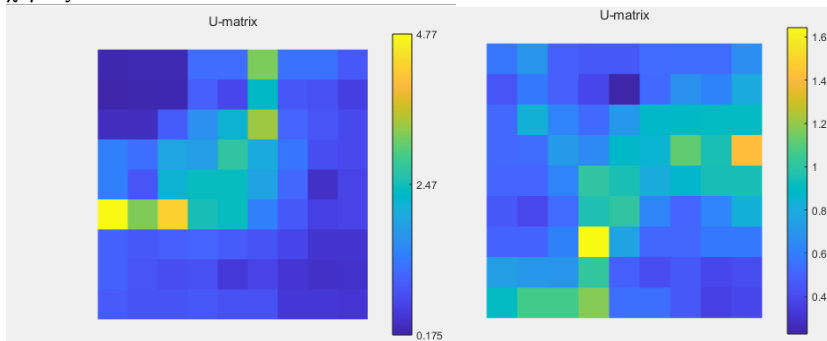
Το μόνο ενδιαφέρον κομμάτι στις ομάδες βρίσκεται σε μια μικρή καθαρή υποομάδα στην μεγαλύτερη ομάδα αλλά συνεχίζει να μην μας δείχνει καλά αποτελέσματα σε σχέση με τους μικρότερους χάρτες.



(Εικόνα 45) Σχεδιάγραμμα umatrix σετ δεδομένων iris στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 14x14, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

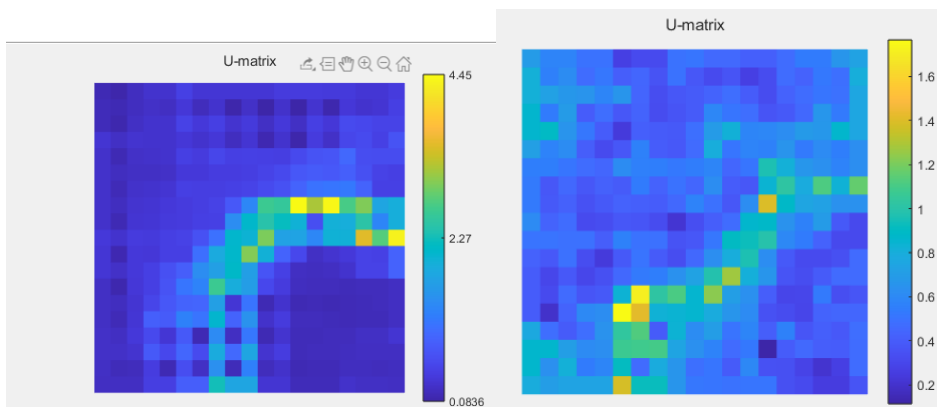
5.2.11 σύγκριση των καλύτερων αποτελεσμάτων.

Γενικότερα και οι δύο χάρτες κατάφεραν να βρουν τις τρεις ομάδες που είχαμε αλλά με αρκετά αντίθετους τρόπους. Αρχικά ο κανονικός som είδαμε ότι δυσκολευόταν αρκετά να βρει την τρίτη ομάδα σε σχέση με τον rsom όπου μας εμφάνισε αποτελέσματα από μικρούς σε διαστάσεις χάρτες.



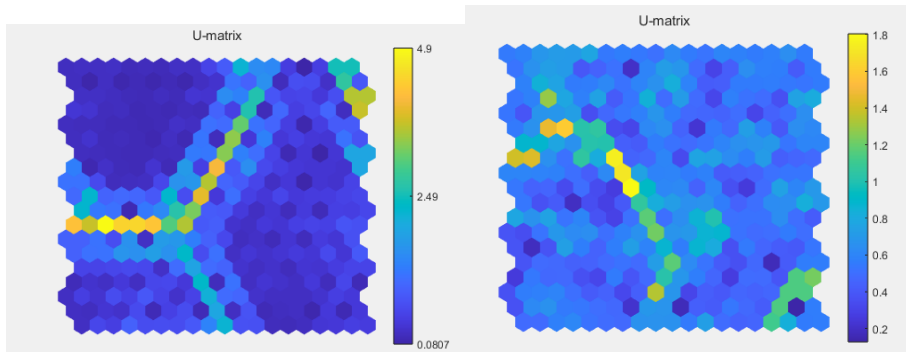
(Εικόνα 46) Σχεδιάγραμμα umatrix μεγέθους χάρτη 5x5 κανονικού som από τα αριστερά και rsom από τα δεξιά σε επίπεδο σχήμα.

Όπως βλέπουμε (εικόνα 46) ο κανονικός som στα αριστερά είχε μια ισότιμη απόσταση και μια καθαρή μια ομάδα στο umatrix ενώ ο rsom έδειξε μια πιο καθαρή τρίτη ομάδα. Όμως όσο διεξήγαμε πειράματα και μεγαλώναμε της διαστάσεις είδαμε ότι ο rsom ξεκινούσε να μας δείχνει χειρότερα αποτελέσματα ενώ το κανονικό som ξεκίνησε να δείχνει καλύτερα αποτελέσματα .



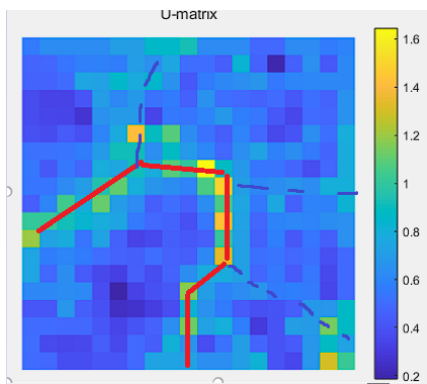
(Εικόνα 47) Σχεδιάγραμμα umatrix μεγέθους χάρτη 10x10 κανονικού som από τα αριστερά και rsom από τα δεξιά σε επίπεδο σχήμα.

Ξεκινώντας από διαστάσεις 10x10 ξεκινάμε είδη να βλέπουμε (εικόνα 47) τον rsom να ξεκινάει να μας δίνει χειρότερα αποτελέσματα από τα προηγούμενα. Συγκεκριμένα στις διαστάσεις 10x10 και η δυο αλγόριθμοι απέτυχαν να βρουν και να δείξουν την τρίτη ομάδα.



(Εικόνα 48) Σχεδιάγραμμα umatrix μεγέθους χάρτη 10x10 κανονικού som από τα αριστερά και rsom από τα δεξιά σε κυλινδρικό σχήμα.

Στο παράδειγμα διαστάσεων 10x10 αλλά σε κυλινδρικό σχήμα χάρτη είδαμε (εικόνα 48) την πρώτη μεγάλη διάφορα στον κανονικό som. Τα αποτελέσματα είναι αρκετά ξεκάθαρα στον κανονικό som με τις τρεις ομάδες να φαίνονται ξεκάθαρα ενώ στον rsom τα αποτελέσματα συνεχίζουν να μας μπερδεύουν αρκετά χωρίς να παίρνουμε κάποιο καθαρό αποτέλεσμα πέρα από της δύο ξεκάθαρες ομάδες που περνάμε συνήθως, τα καλύτερα αποτελέσματα από τον rsom ήρθαν από το τετραγωνικό πλέγμα.



(Εικόνα 49) Σχεδιάγραμμα umatrix μεγέθους χάρτη 5x5 rsom σε επίπεδο κυλινδρικό σχήμα με κόκκινες γραμμές για να τονίσουμε τα ορία και διακεκομμένες γραμμές για της δευτερεύων ομάδες.

Όπου είδαμε 2 πιο καθαρές ομάδες στην μεγαλύτερη κύρια ομάδα αλλά ταυτόχρονα και άλλες μικρότερες ομάδες μέσα της η οποίες δεν εμφανίστηκαν σε άλλα παραδείγματα.

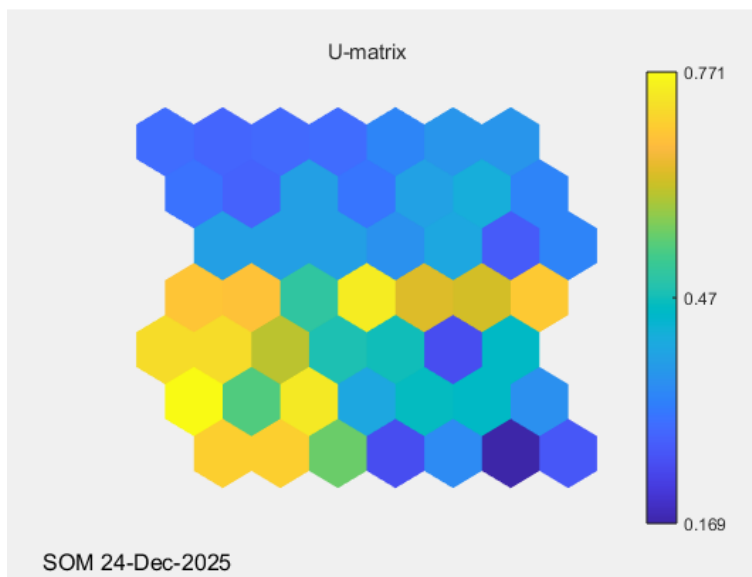
Γενικότερα από τα πειράματα βλέπουμε ότι το σχήμα από επίπεδο σε κυλινδρικό βοήθησε στην εικονοποίηση των δεδομένων και επίσης στο τωρινό παράδειγμα είδαμε ότι ο κανονικός som διαχειριζόταν δεδομένα καλύτερα σε μεγαλύτερους σε διαστάσεις χάρτες τα δεδομένα ενώ αντιθέτως ο rsom διαχειριζόταν τα δεδομένα καλύτερα σε μικρότερους σε διαστάσεις χάρτες. Κρατώντας αυτά τα αποτελέσματα στο μυαλό μας προχωράμε στο επόμενο παράδειγμα όπου μιλάμε για δεδομένα σε υπολείμματα αμινοξέων σε απόσταση grantham όπου πρόκειται για ένα μικρότερο σε διαστάσεις σετ δεδομένων.

5.3 παράδειγμα σετ δεδομένων υπολειμμάτων αμινοξέων

Προχωρώντας στο επόμενο παράδειγμα το οποίο πρόκειται για τα δεδομένα ανομοιότητας αποστάσεων grantham σε αμινοξέα τα οποία έχουν πρόσφατα μεταλλαχθεί σε μια πρωτεΐνη. Βλέπουμε από την ανάλυση των δεδομένων ότι πρόκειται φυσικά για ένα πολύ μικρότερο σε διαστάσεις πίνακα δεδομένων. Ταυτόχρονα όμως μπορούσαμε να δούμε και 3 πιο ξεκάθαρες ομάδες από την κλιμάκωση που κάναμε στα δεδομένα. Το παράδειγμα θα γίνει με ίδιες ρυθμίσεις με το iris αυτό σημαίνει φυσικά 500 επαναλήψεις χρόνο εκπαίδευσης και batch training. Όμως λόγω του ότι τα δεδομένα είναι μικρότερα θα χρειαστεί να χειριστούμε της διάστασης των χαρτών πιο διακριτικά αφού το να κάνουμε overfit τα δεδομένα είναι εννοείται πιο εύκολο. Για αυτόν τον λόγο αντί για 5x5-15x15 που ήταν η προηγούμενες διαστάσεις θα γίνουν σε πιο μικρές διαστάσεις, 4x4-6x6. Ένας παράγοντας που επιλέχτηκε το 6x6 είναι διότι στο rsom συγκεκριμένα υπάρχει μεγάλο πρόβλημα overfit μετά της διαστάσεις 6x6 αλλά ταυτόχρονα σε διαστάσεις 3x3 δυσκολευόμαστε να δούμε κάτι ουσιαστικό από τα αποτελέσματα που παίρνουμε.

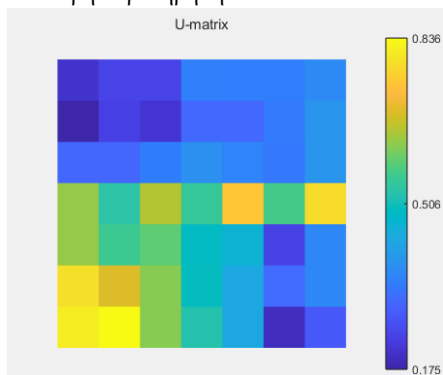
5.3.1 Διαστάσεις χάρτη 4x4 επίπεδο σχήμα SOM

Ξεκινώντας με το πρώτο παράδειγμα διαστάσεων 4x4 σε επίπεδο σχήμα με εξαγωνικό πλέγμα. Βλέπουμε παρακάτω (εικόνα 50) δύο ξεκάθαρες ομάδες. Λόγω του μεγέθους του σχήματος φυσικά δεν έχουμε αρκετά δεδομένα για να χρησιμοποιήσουμε αλλά βλέπουμε μια πολύ βασική μορφή του χάρτη.



(Εικόνα 50) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 4x4, σε κανονικό som , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα .

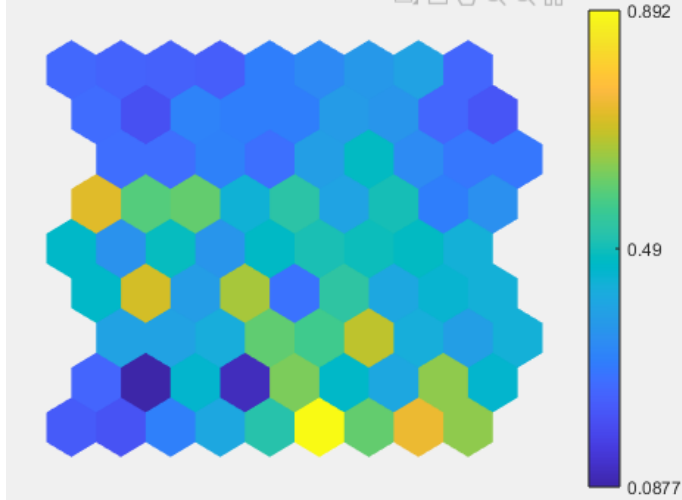
Στο τετραγωνικό πλέγμα επίσης δεν βλέπουμε (εικόνα 51) κάποια ιδιαίτερη διαφορά και τα δεδομένα μεταξύ των δύο πλεγμάτων συμφωνούν μεταξύ τους οπότε δεν υπάρχει κάποια ιδιαίτερη παρατήρηση πάνω σε αυτό που δεν έγινε ήδη στο εξαγωνικό πλέγμα.



(Εικόνα 51) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 4x4, σε κανονικό som , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

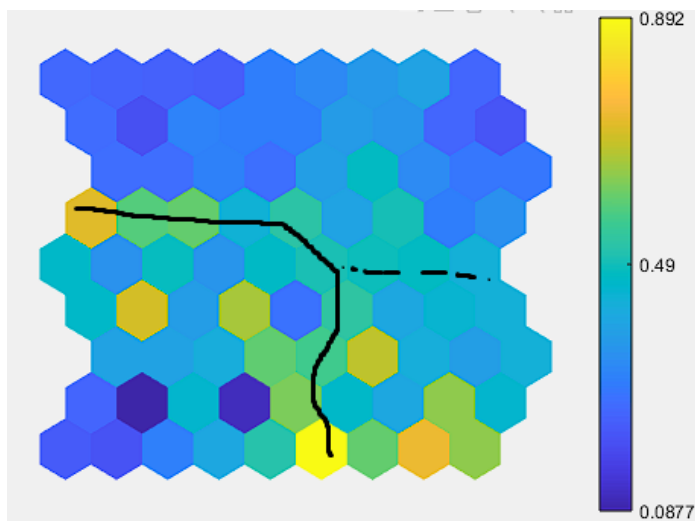
5.3.2 διαστάσεις χάρτη 5x5 σε επίπεδο σχήμα στον κανονικό som

Προχωρώντας στο επόμενο παράδειγμα στο οποίο η διαστάσεις του χάρτη νευρώνων αυξάνονται σε 5x5 ξεκινάμε να βλέπουμε (εικόνα 52) περισσότερες πληροφορίες. Αρχικά όπως και στο προηγούμενο παράδειγμα βλέπουμε της δύο ξεκάθαρες ομάδες αλλά ταυτόχρονα ξεκινάμε να βλέπουμε και μια τρίτη ομάδα να αρχίζει να φαίνεται αρκετά ξεκάθαρα.



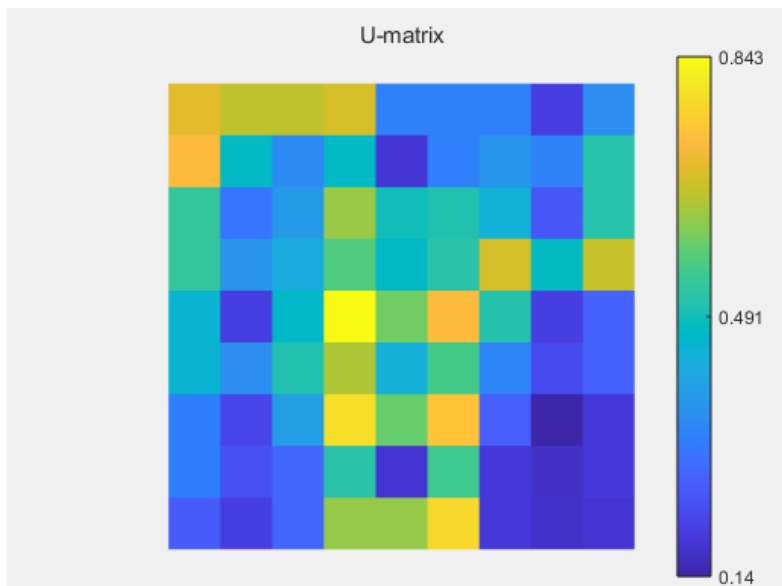
(Εικόνα 52) Σχεδιάγραμμα umatrix σε δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα .

Παρακάτω(εικόνα 53) μπορούμε έχουν χρωματιστή τα πιο κύρια όρια με μαύρη γραμμή ενώ με διακεκομμένες δείχνεται ένα τρίτο πιο απαλό όριο στις ομάδες που βλέπουμε στα δεδομένα. Παρόλο που δεν είναι όσο ξεκάθαρη όσο η δύο κύριες ομάδες μπορούμε να δούμε ότι ο χάρτης ξεκινάει να εξελίσσεται παραπάνω και να μας δείχνει περισσότερη πληροφορία.



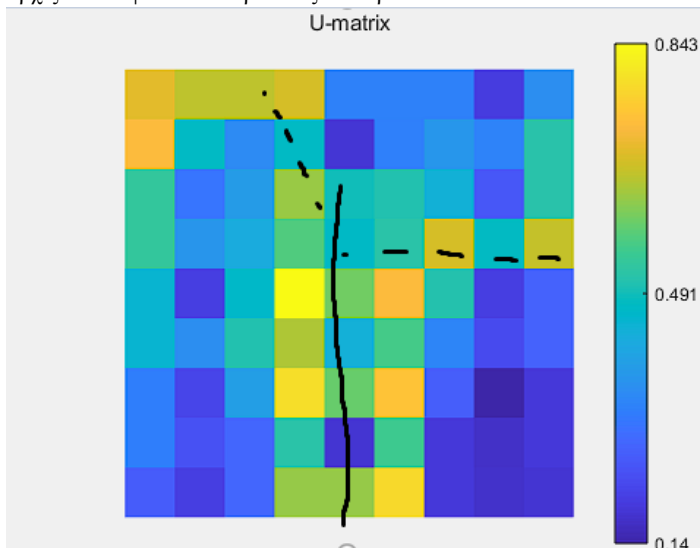
(Εικόνα 53) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα . Με μαύρες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερεύων ομάδες

Τώρα όμως στο τετραγωνικό πλέγμα βλέπουμε (εικόνα 54) τις τρεις ομάδες που ξεκίνησαν να φαίνονται πολύ πιο ξεκάθαρα. Όχι μόνο και πάλι βλέπουμε της δύο κύριες ομάδες αλλά βλέπουμε να ακολουθεί το σχήμα από την προηγούμενη εικόνα απλώς όμως γυρισμένη σε 90 μοίρες. Αυτό είναι ένα φαινόμενο που σίγουρα παρατηρήθηκε και στο παράδειγμα του iris αφού η νευρώνες αρχικοποιούνται σε τυχαίες θέσης οπότε είναι συχνό φαινόμενο παραδείγματα μεταξύ του αλγορίθμου να είναι περιστραμένα αρκετές φορές.



(Εικόνα 54) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

Όπως και στο προηγούμενο παράδειγμα βλέπουμε (εικόνα 55) με κύρια γραμμή το μεγαλύτερο όριο μεταξύ των ομάδων και με διακεκομμένες γραμμές βλέπουμε τα όρια της τρίτης ομάδας να αρχίζουν να φαίνονται αρκετά ξεκάθαρα.



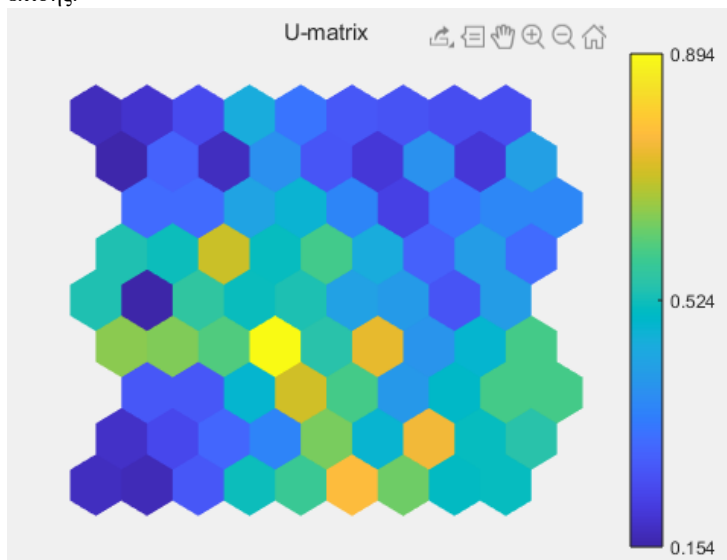
(Εικόνα 55) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα . Με μαύρες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερευόν ομάδες.

5.3.3 Διαστάσεις 5x5 σε κυλινδρικό σχήμα στον κανονικό som

Όπως είδαμε στο προηγούμενο παράδειγμα το σχήμα που επίσης επιλέγουμε για τον χάρτη έπαιξε μεγάλο ρόλο στα δεδομένα που πήραμε. Για αυτόν τον λόγο ξεκινώντας από της διαστάσεις 5x5 στις οποίες παίρνουμε καλά αποτελέσματα είναι μια στιγμή να τρέξουμε τον αλγόριθμο και σε κυλινδρικό σχήμα για να δούμε αν συμφωνούνε τα δεδομένα η αν από την άλλη τα επηρεάζει αρνητικά.

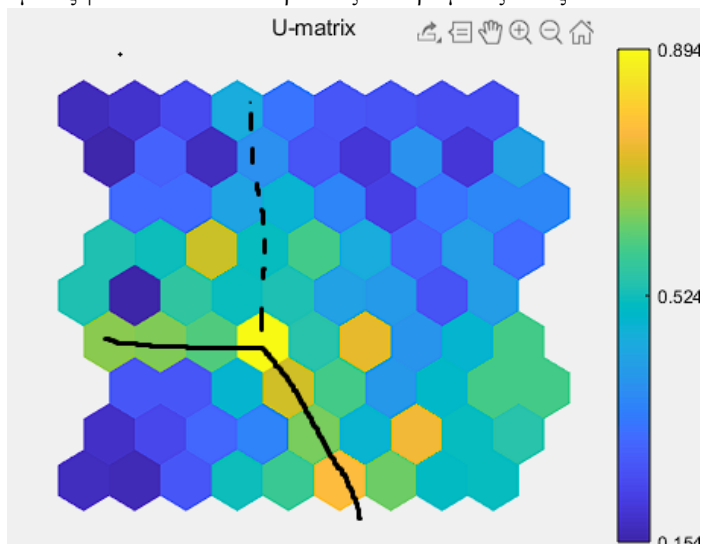
Ξεκινώντας με το εξαγωνικό σχήμα μπορούμε να δούμε (εικόνα 56) ότι παίρνουμε παρόμοια αποτελέσματα με το επίπεδο πλαίσιο όπου η δύο ξεκάθαρες ομάδες φαίνονται αλλά ταυτόχρονα μπορούμε παρομοίως με τον τετραγωνικό χάρτη να δούμε πιο ξεκάθαρα την τρίτη ομάδα

επίσης.



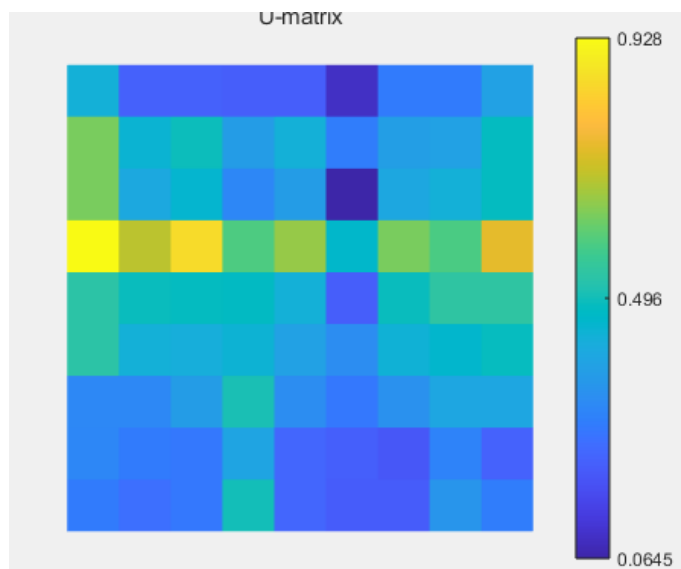
(Εικόνα 56) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα .

Παρόμοια με το προηγούμενο παράδειγμα μπορούμε να διακρίνουμε (εικόνα 57) στις διακεκομμένες γραμμές ένα όριο το οποίο ξεκινάει μεταξύ των δεδομένων . Ταυτόχρονα η δυο ομάδες φαίνονται και πάλι αρκετά ξεκάθαρα μεταξύ τους.



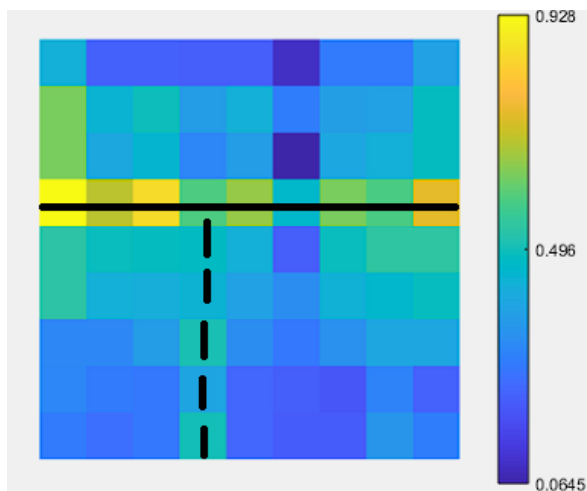
(Εικόνα 57) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα . Με μαύρες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερεύων ομάδες.

Στο τετραγωνικό πλέγμα βλέπουμε (εικόνα 58) τρεις ομάδες αρκετά ξεκάθαρα και μάλιστα τα όρια της τρίτης ομάδας πιο ξεκάθαρα από ότι το εξαγωνικό πλέγμα.



(Εικόνα 58) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα .

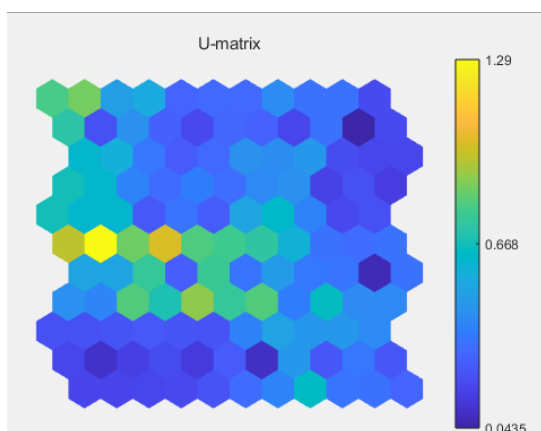
Όπως και στα προηγούμενα παραδείγματα στην (εικόνα 59) με μαύρο χρώμα έχει χρωματιστή το κύριο όριο ενώ με διακεκομμένες το πιο απαλό όριο αναμεσά στην δεύτερη ομάδα που την χωρίζει.



(Εικόνα 59) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε κανονικό som , σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα . Με μαύρες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερεύων ομάδες.

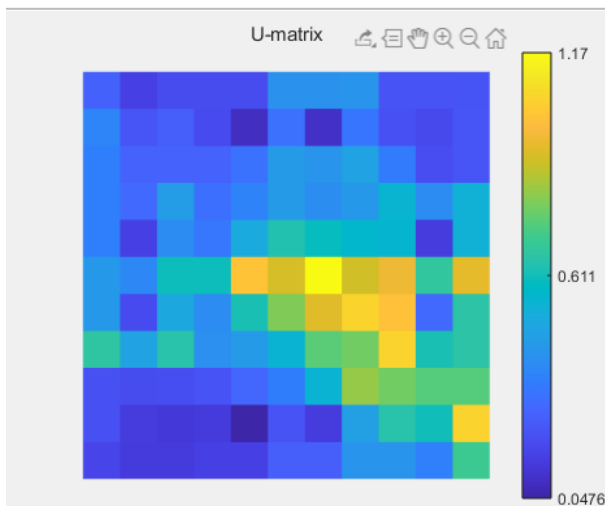
5.3.4 Διαστάσεις χάρτη 6x6 σε επίπεδο σχήμα στον κανονικό som

Προχωρώντας στον μεγαλύτερο σε διαστάσεις χάρτη με διαστάσεις 6x6 και ξεκινώντας με το επίπεδο σχήμα . Βλέπουμε (εικόνα 60) ότι δεν γίνεται κάποια ιδιαίτερη αλλαγή στο βασικό σχήμα αλλά ταυτόχρονα βλέπουμε ότι η τρίτη ομάδα στο εξαγωνικό πλέγμα έχει αρχίσει να εξαφανίζεται και να ενώνεται ξανά με την μεγαλύτερη ομάδα κάτι που θα μπορούσε να μας δείχνει ότι ξεκινάμε να κάνουμε overfit τα δεδομένα.



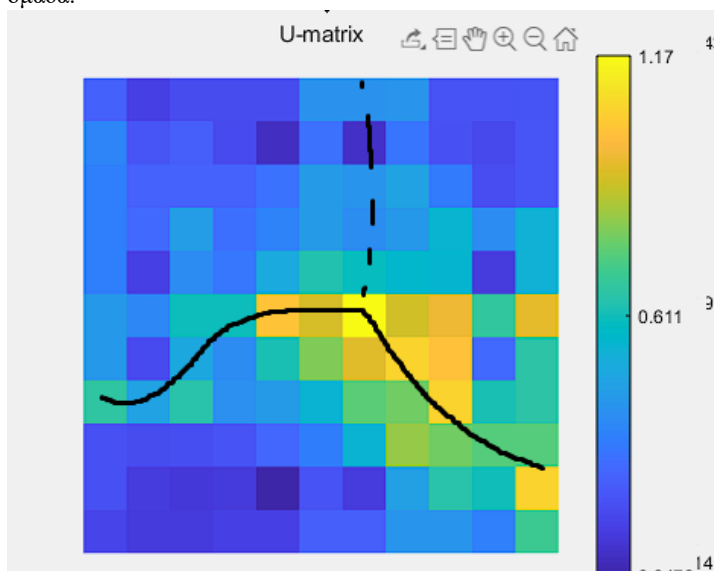
(Εικόνα 60) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 6x6, σε κανονικό som , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα .

Αντιθέτως όμως στο τετραγωνικό πλέγμα μπορούμε να δούμε(εικόνα 61) ακόμα την τρίτη ομάδα αρκετά καθαρά κάτι που ξεκινάει να μας δείχνει διαφορές στα αποτελέσματα που παίρνουμε μεταξύ των δυο πλεγμάτων.



(Εικόνα 61) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 6x6, σε κανονικό som , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

Ταυτόχρονα όμως είναι σημαντικό να αναφερθεί ότι παρόλο που βλέπουμε το όριο ξεκινάμε επίσης να βλέπουμε ότι τα όρια στην τρίτη ομάδα έχουν αρχίσει να γίνονται πιο αχνά στο σημείο με της διακεκομμένες γραμμές όπου αναδεικνύετε το όριο μεταξύ των ομάδων μπορούμε να δούμε ότι υπάρχει μια μικρή γειτονιά που ήδη ξεκινάει να ενώνει την ομάδα κάτι που πιθανότατα είναι και ο λόγος που στο εξαγωνικό πλέγμα δεν μπορούμε να δούμε την τρίτη ομάδα.

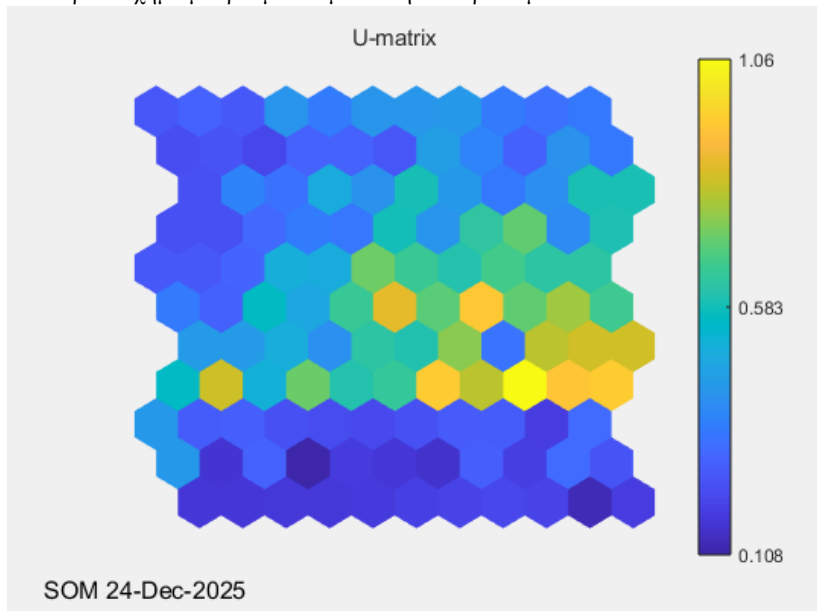


(Εικόνα 62) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 6x6, σε κανονικό som , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα . Με μαύρες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερεύων ομάδες.

5.3.5 Διαστάσεις χάρτη 6x6 σε κυλινδρικό σχήμα σε κανονικό som

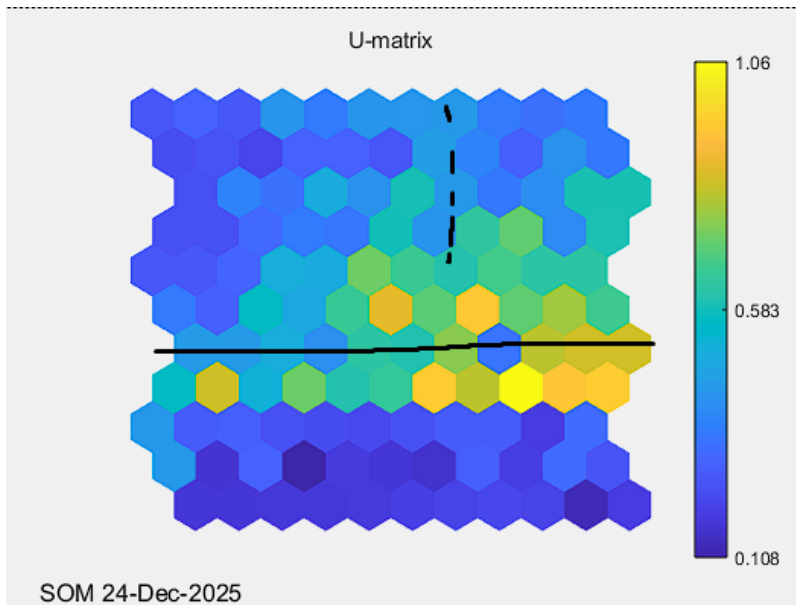
Τέλος των παραδειγμάτων για τον κανονικό som έχουμε το παράδειγμα σε διαστάσεις 6x6 αλλά σε κυλινδρικό σχήμα. Γενικότερα βλέπουμε (εικόνα 63) ότι τα δεδομένα παραμένουν με δύο πιο ξεκάθαρες ομάδες με την τρίτη ομάδα να φαίνεται αρκετά ελάχιστα αλλά σε σχέση με τον μη

κυλινδρικό σχήμα μπορούμε ακόμα να την διακρίνουμε.



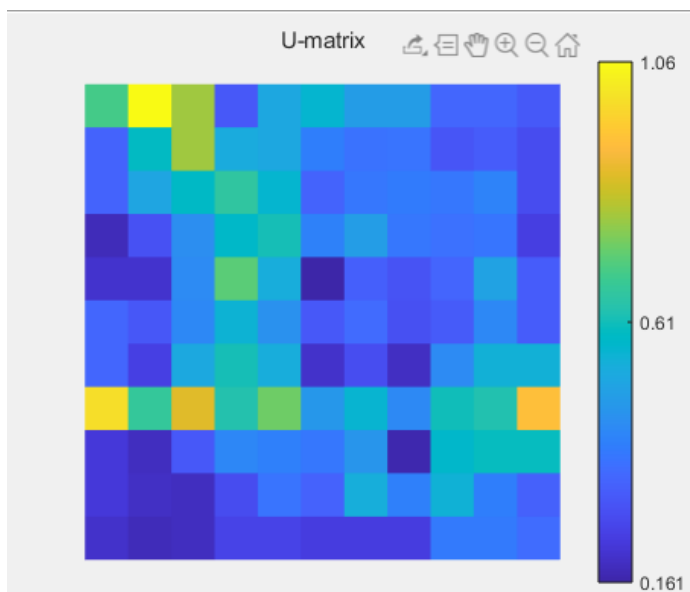
(Εικόνα 63) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 6x6, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα .

Το όριο είναι όπως βλέπουμε (εικόνα 64) αρκετά αχνό και είναι στα όρια να εξαφανιστεί επίσης αλλά βλέπουμε ότι μπορεί ακόμα να διακριθεί βλέπουμε ταυτόχρονα μια μικρή ένωση των δύο κύριων ομάδων και που μπορεί και να γίνεται λόγω του σχήματος του χάρτη αλλά δεν είναι κάτι το οποίο θα μπορούσε να μπερδέψει κάποιον προς το παρών.



(Εικόνα 64) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 6x6, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα . Με μαύρες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερεύων ομάδες.

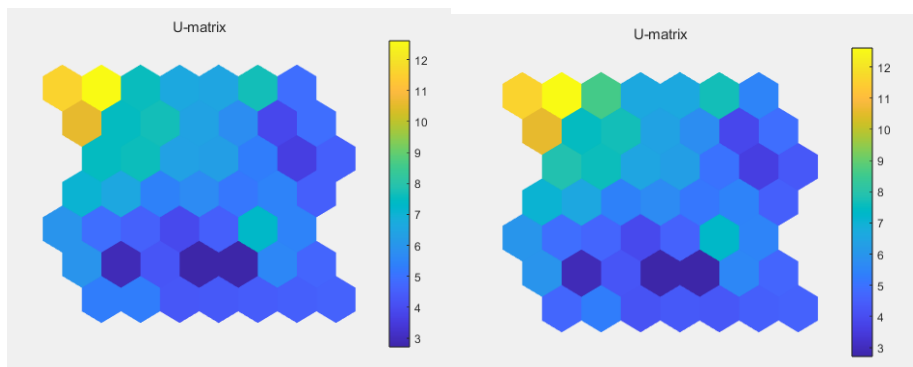
Προχωρώντας όμως στο τετραγωνικό πλέγμα μπορούμε να παρατηρήσουμε(εικόνα 65) κάτι αρκετά ενδιαφέρον. Αρχικά 1 βλέπουμε τρεις ομάδες αλλά ταυτόχρονα δεν βλέπουμε δύο κεντρικές και κυρίες ομάδες. Βλέπουμε 3 ομάδες η οποίες είναι πιο διάσπαρτες στον χώρο αλλά εύκολα δια κρίσιμες.



(Εικόνα 65) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 6x6, σε κανονικό som , σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα .

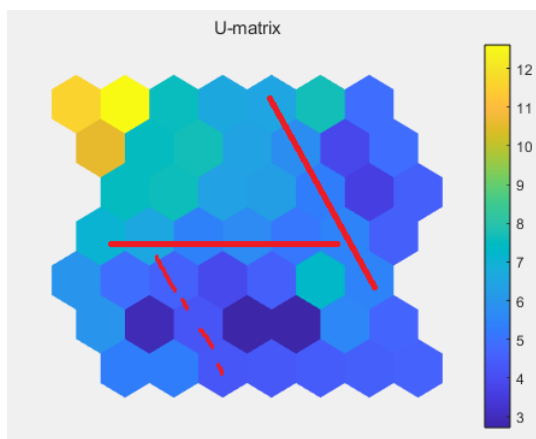
5.3.6 Διαστάσεις χάρτη 4x4 σε επίπεδο σχήμα σε rsom

Προχωρώντας στο rsom όπου θα χρησιμοποιήσουμε παρόμοιες ρυθμίσεις με τον κανονικό som ξεκινάμε με χάρτη σε διαστάσεις 4x4. Τα umatrix όπως και στο παράδειγμα του iris θα είναι σε δυο ρυθμίσεις αποστάσεων η οποίες είναι σε διάμεση και μέση απόσταση. Παρόμοια με τον κανονικό som βλέπουμε(εικόνα 67) ότι γενικά ο χάρτης είναι αρκετά μικρός για να μας δείξει τα αποτελέσματα σε μεγάλη ευκρίνεια όμως μπορούμε να διακρίνουμε ήδη μερικά χαρακτηριστικά πάνω στον χάρτη, για παράδειγμα έχουμε μια “νεκρή ζώνη” πάνω αριστερά με μια ομάδα στα αριστερά της και όπως βλέπαμε και στον κανονικό μπορούμε να ξεκινήσουμε να βλέπουμε μια μεγάλη ομάδα η οποία αποτελείται από δυο μικρότερες, κάτι που είδαμε να γίνεται και από τον κανονικό.



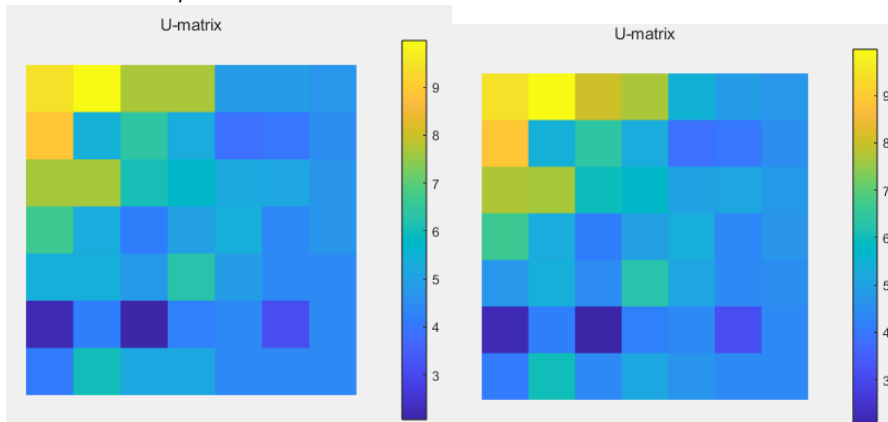
(Εικόνα 67) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 4x4, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

Γενικότερα βλέπουμε(εικόνα 68) και πάλι μερικά όρια που ορίζονται στο umatrix από τον χάρτη αλλά ταυτόχρονα ένα χαρακτηριστικό οροί μέσα στα όρια της ομάδας το οποίο δείχνεται από της διακεκομμένες γραμμές παρακάτω.



(Εικόνα 68) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 4x4, σε rsom , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα . Με κόκκινες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερεύων ομάδες.

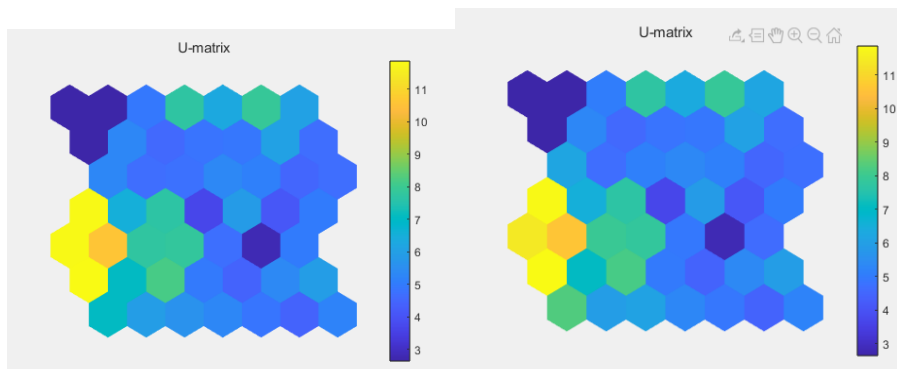
Στο τετραγωνικό πλέγμα του umatrix παίρνουμε παρόμοια αλλά ταυτόχρονα πολύ πιο μπερδεμένα δεδομένα κάτι που δεν διευκολύνει συγκεκριμένες παρατηρήσεις για τον συγκεκριμένο χάρτη παρόλα αυτά το βασικό σχήμα των ομάδων βλέπουμε (εικόνα 69) ότι μένει πάνω κάτω σταθερό.



(Εικόνα 69) Σχεδιαγράμματα umatrix σε δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 4x4, σε rsom, σε επίπεδο σχήμα, σε τετραγωνικό πλέγμα.

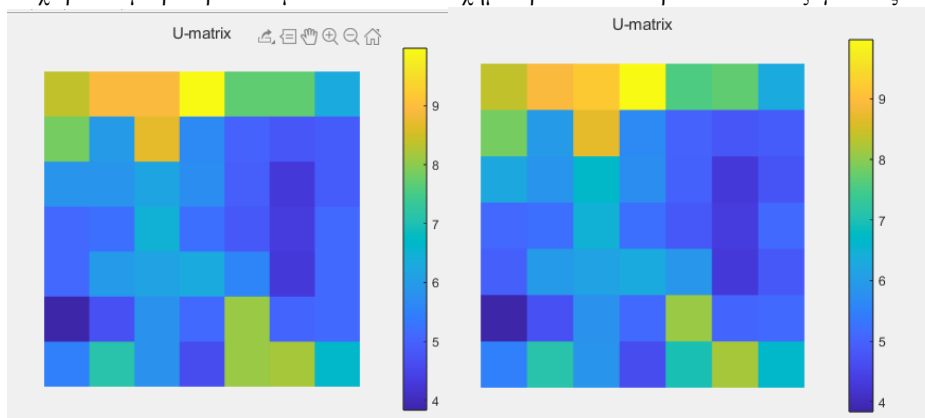
5.3.7 Διαστάσεις χάρτη 4X4 σε κυλινδρικό σχήμα σε RSOM

Προχωρώντας θα δείξουμε της ίδιες διαστάσεις αλλά σε κυλινδρικό σχήμα λόγω των αποτελεσμάτων που παίρνονται. Αρχικά βλέπουμε (εικόνα 70) το εξαγωνικό πλέγμα το οποίο βλέπουμε ότι συσσωρεύει τους περισσότερους νευρώνες στο πάνω αριστερά κομμάτι του umatrix ενώ ταυτόχρονα βλέπουμε ότι ο χάρτης δυσκολεύεται να δήσει κάποια ουσιαστικά δεδομένα πάνω στο umatrix



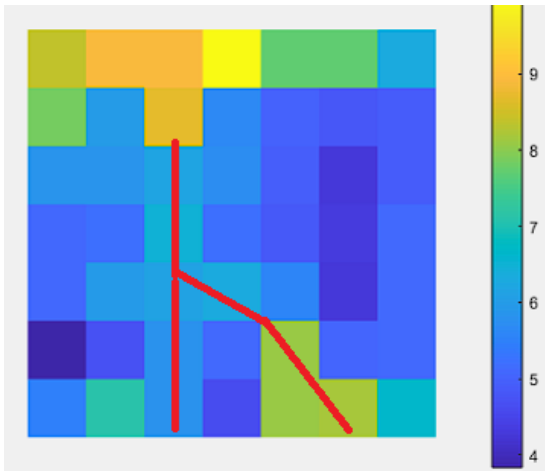
(Εικόνα 70) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 4x4, σε rsom, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

Αντιθέτως όμως στο τετραγωνικό πλέγμα παίρνουμε (εικόνα 71) πάρα πολύ καλά αποτελέσματα και ξεκάθαρα στο κυκλικό σχήμα συγκεκριμένα μπορούμε να δούμε 2 πολύ ξεκάθαρες ομάδες ενώ ταυτόχρονα και τον διαχωρισμό μιας τρίτης ομάδας. Κάτι που γενικότερα βλέπουμε πρώτη φορά σε παραδείγματα και μπορεί να είναι σημάδι του τετραγωνικού πλέγματος να μπορεί να διαχειριστεί μικρότερα δεδομένα σε κυκλικό σχήμα αρκετά καλύτερα από ότι ο εξαγωνικός.



(Εικόνα 71) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 4x4, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

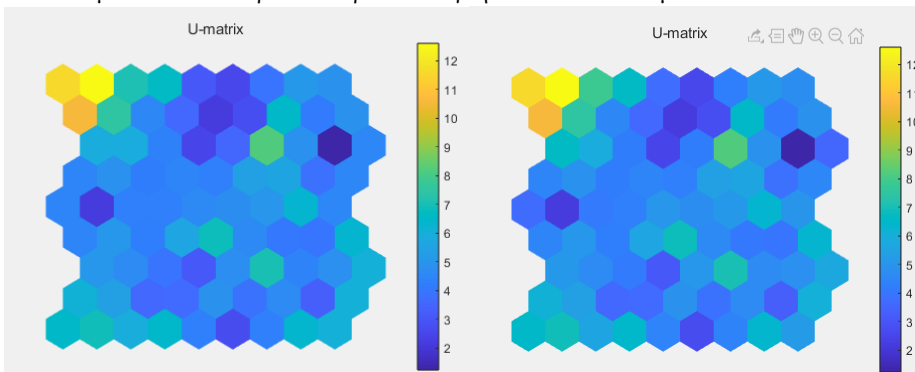
Παρακάτω βλέπουμε ότι τα όρια φαίνονται αρκετά ξεκάθαρα μεταξύ των τριών ομάδων σε σχέση με το εξαγωνικό πλέγμα στο οποίο η ομάδες δεν φαίνονται ξεκάθαρα και που βλέπουμε τον χάρτη να ξεκάθαρα αποτυγχάνει να διαμορφωθεί ξεκάθαρα.



(Εικόνα 72) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 4x4, σε rsom , σε κυκλικό σχήμα ,σε τετραγωνικό πλέγμα . Με κόκκινες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερεύων ομάδες.

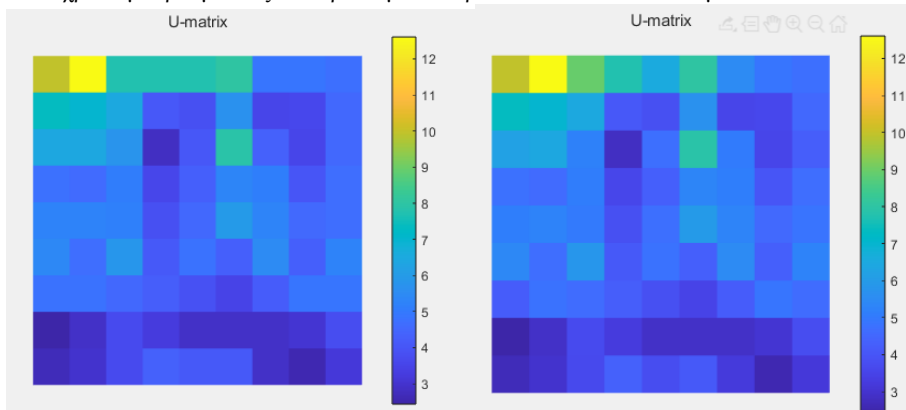
5.3.8 Διαστάσεις χάρτη 5x5 σε επίπεδο σχήμα σε rsom

Προχωρώντας στο επόμενο παράδειγμα πηγαίνουμε σε χάρτη μεγαλύτερων διαστάσεων 5x5. Ξεκινώντας με το εξαγωνικό πλέγμα βλέπουμε (εικόνα 73) και πάλι ένα αρκετά χαοτικό αποτέλεσμα από το umatrix ,μπορούμε να ξεκαθαρίσουμε μικρές ομάδες αλλά ταυτόχρονα δεν είναι αρκετά ξεκάθαρες σε σημείο που θα μπορούσαν να θεωρηθούν και απλώς σαν ήχος τον αποτελεσμάτων. Γενικότερα είναι αρκετά απογοητευτικά αποτελέσματα.



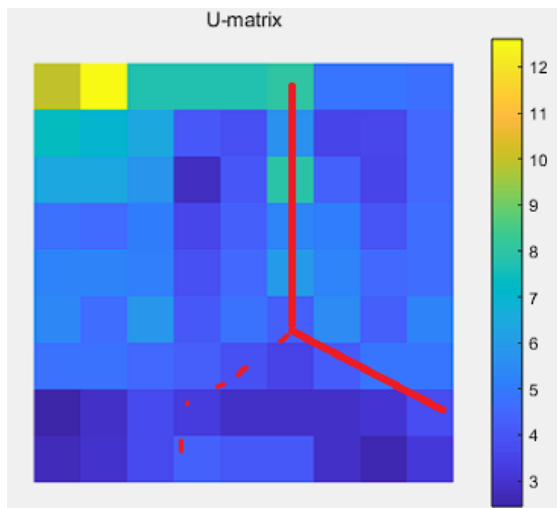
(Εικόνα 73) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα.

Αντιθέτως όμως με τα αποτελέσματα του εξαγωνικού πλέγματος στο τετραγωνικό πλέγμα βλέπουμε πιο ξεκάθαρα αποτελέσματα. Αρχικά μπορούμε να δούμε (εικόνα 74) πολύ μεγαλύτερες ομάδες κάτι που φυσικά επηρεάζεται από το αυξημένο μέγεθος του χάρτη αλλά ταυτόχρονα μπορούμε να ξεκαθαρίσουμε και όρια σε αυτά τα αποτελέσματα.



(Εικόνα 74) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα.

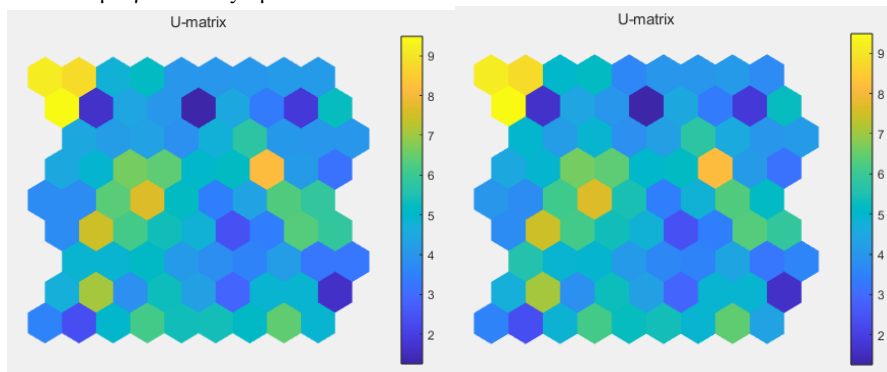
Γενικότερα τα δεδομένα που παίρνουμε και το σχήμα τους μοιάζει αρκετά με τα δεδομένα που πήραμε από το παράδειγμα με κυλινδρικό σχήμα 4x4 έχουμε μια ξεκάθαρη ομάδα στα δεξιά ενώ ταυτόχρονα βλέπουμε ένα μικρό όριο μέσα στην μεγαλύτερη ομάδα σηματοδοτημένη με μπλε διακεκομμένες γραμμές. Ταυτόχρονα όμως μπορούμε να διακρίνουμε (εικόνα 75) και άλλο ένα όριο με κόκκινες γραμμές το οποίο μας δείχνει μια μικρότερη ομάδα. Τώρα εννοείται πως αυτό θα μπορούσε να είναι μια τέταρτη ομάδα αλλά ταυτόχρονα θα μπορούσε να είναι και ήχος ο οποίος ανήκει στην μεσαία ομάδα . Εννοείται ότι θα χρειαστούν περισσότερα αποτελέσματα για να πάρουμε μια πιο συγκεκριμένη παρατήρηση.



(Εικόνα 75) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 4x4, σε rsom , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα . Με κόκκινες γραμμές για να τονιστούν τα όρια και διακεκομμένες γραμμές για δευτερευόν ομάδες.

5.3.9 Διαστάσεις χάρτη 5x5 σε κυλινδρικό σχήμα σε rsom

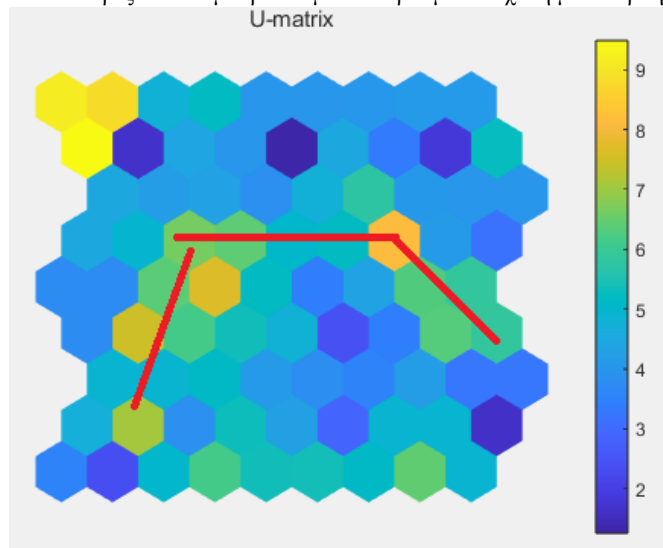
Προχωρώντας στο κυλινδρικό σχήμα του χάρτη για το παράδειγμα βλέπουμε (εικόνα 76) στο εξαγωνικό και πάλι μη επική αποτελέσματα. Η μπορούμε να ξεκαθαρίσουμε κάποιες ομάδες αλλά τα δεδομένα είναι πολύ απομακρυσμένα και τυχαία για να έχουμε κάποιο καλό αποτέλεσμα για να δείξουμε.



(Εικόνα 76) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος

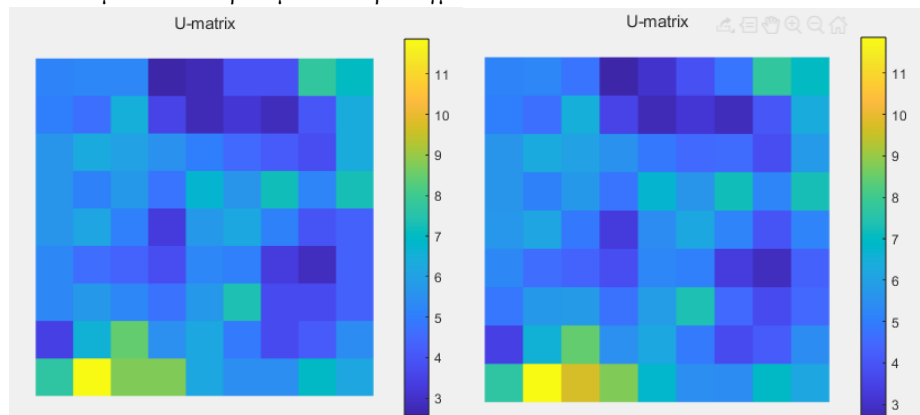
χάρτη 5χ5, σε isom, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

Παρόλα αυτά δύο ομάδες μπορούν να φανούν ξεκάθαρα (εικόνα 77) όμως και πάλι δεν είναι και οι καλύτερες που θα μπορούσαμε να πάρουμε σε σχέση με τα προηγούμενα αποτελέσματα.



(Εικόνα 77) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 5x5, σε rsom , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα . Με κόκκινες γραμμές για να τονιστούν τα όρια

Όμως όπως και στο επίπεδο σχήμα βλέπουμε (εικόνα 78) πάλι ότι το τετραγωνικό πλέγμα μας δίνει πολύ πιο ξεκάθαρα αποτελέσματα στο umatrix. Βλέπουμε δύο πολύ ξεκάθαρες ομάδες πάνω στον χάρτη οι οποίες βρίσκονται στην δεξιά και στην πάνω πλευρά του χάρτη ενώ ταυτόχρονα μια μικρότερη τρίτη ομάδα να χωρίζεται κάτι που ταιριάζει και με προηγούμενα αποτελέσματα που παίρνουμε από παραδείγματα.

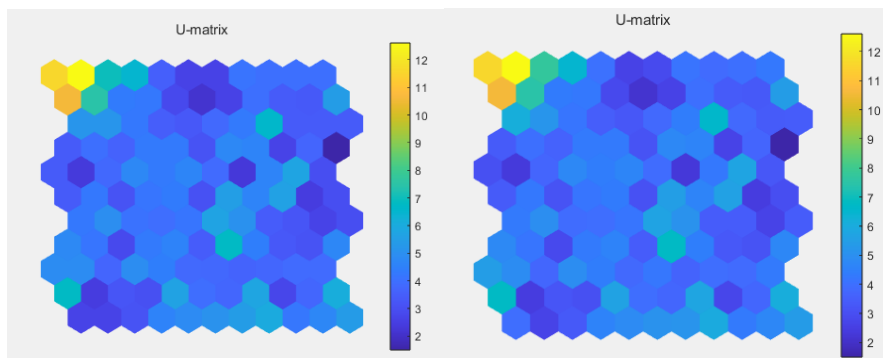


(Εικόνα 78) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε

διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

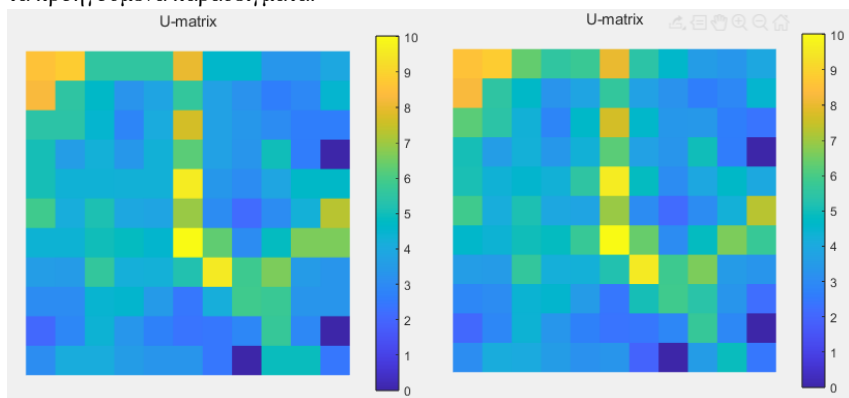
5.3.10 Διαστάσεις χάρτη 6x6 σε επίπεδο σχήμα σε rsom

Προχωρώντας στο μεγαλύτερο σε διαστάσεις παράδειγμα για το συγκεκριμένο set δεδομένων πάμε σε διαστάσεις 6x6. Ξεκινώντας με το εξαγωνικό πλέγμα βλέπουμε ότι τα αποτελέσματα(εικόνα 79) έχουν γίνει ακόμα πιο χαοτικά σε σημείο που δεν μπορούμε να πάρουμε κάποια καλή πληροφορία. Γενικότερα όσο μεγαλώνουν η διαστάσεις μπορούμε να παρατηρήσουμε ότι το umatrix του εξαγωνικού πλέγματος γίνεται όλο και πιο χαοτικό.



(Εικόνα 79) Σχεδιαγράμματα umatrix σε δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 6x6, σε rsom, σε επίπεδο σχήμα, σε εξαγωνικό πλέγμα.

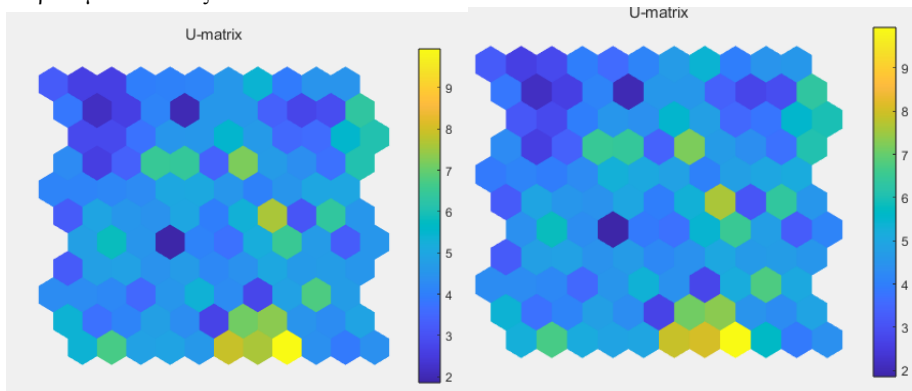
Από την άλλη συνεχίζοντας να ακολουθεί το μοτίβο ακριβώς αντιθέτως βλέπουμε (εικόνα 80) τα πιο καθαρά δεδομένα στο τετραγωνικό πλέγμα μιας και 3 ομάδες φαίνονται πάρα πολύ ξεκάθαρα με δύο ομάδες αριστερά και πάνω δεξιά μαζί με μια μικρότερη ομάδα κάτω αριστερά στον χάρτη του umatrix κάτι που συναινεί και με τα δεδομένα που παίρνουμε συνήθως και από τα προηγούμενα παραδείγματα.



(Εικόνα 80) Σχεδιαγράμματα umatrix σε δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 6x6, σε rsom, σε επίπεδο σχήμα, σε τετραγωνικό πλέγμα.

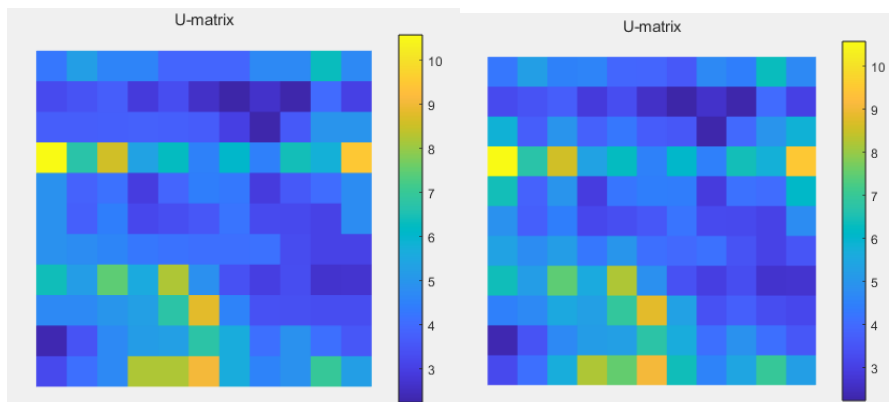
5.3.11 Διαστάσεις χάρτη 6x6 σε κυλινδρικό σχήμα σε rsom

Προχωρώντας στο τελευταίο παράδειγμα το οποίο πρόκειται για τον χάρτη νευρώνων σε διαστάσεις 6x6 σε κυλινδρικό σχήμα και ξεκινώντας με το εξαγωνικό πλέγμα. Βλέπουμε (εικόνα 81) ότι συνεχίζεται το χάος που έχουμε παρατηρήσει στα περισσότερα από τα παραδείγματα. Ομάδες πάνε να σχηματιστούν αλλά βλέπουμε ότι δεν υπάρχει κάτι ξεκάθαρο στο σχήμα που παίρνουμε στο τέλος.



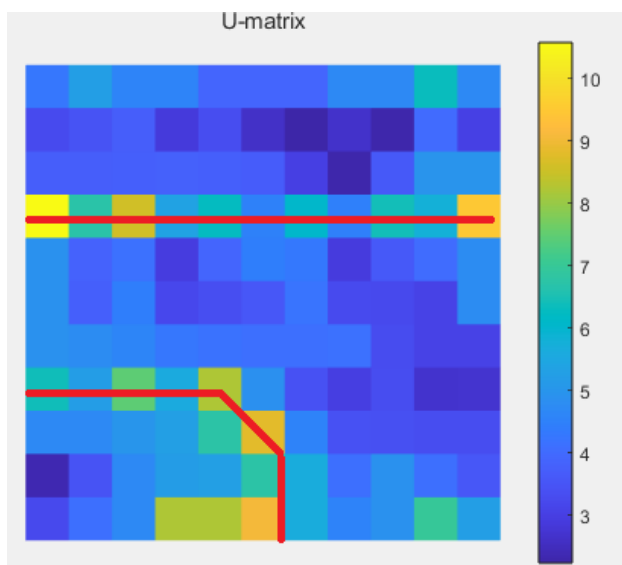
(Εικόνα 81) Σχεδιαγράμματα umatrix σε δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 6x6, σε rsom, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

Από την άλλη τα δεδομένα που παίρνουμε συνεχίζουν να είναι πιο ξεκάθαρα αλλά μπορούμε να δούμε ότι ξεκινάνε να αλλοιώνονται ακόμα και στο τετραγωνικό πλέγμα. Γενικά βλέπουμε (εικόνα 82) τρεις ομάδες και πάλι ξεκάθαρα διαχωρισμένες. Ταυτόχρονα μπορούμε να δούμε ότι ξεκινάνε να εμφανίζονται κενά στα όρια κάτι που ίσως είναι σημάδι ότι το ίδιο πράγμα που έχει συμβεί στο εξαγωνικό πλέγμα αρχίζει να συμβαίνει και στο τετραγωνικό πλέγμα.



(Εικόνα 82) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 6x6, σε t-som, σε κυλινδρικό σχήμα, σε τετραγωνικό πλέγμα.

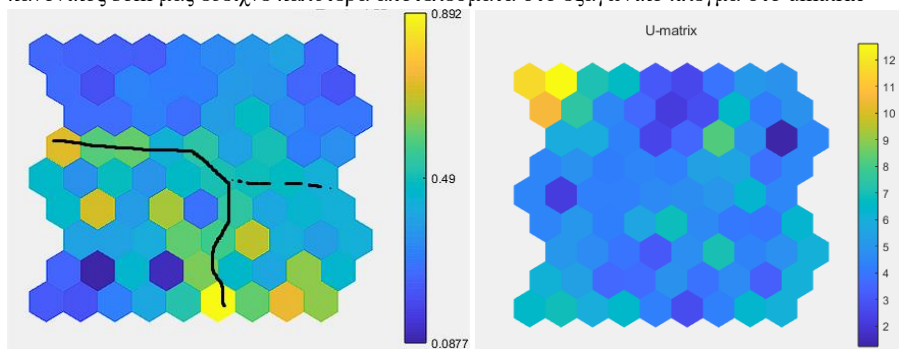
Παρακάτω βλέπουμε (εικόνα 83) τα όρια των ομάδων με μπλε γραμμές και βλέπουμε ότι είναι γενικά αρκετά ξεκάθαρα τα όρια μεταξύ των ομάδων χωρίς να αφήνεται χώρος για να δειχτεί το αντίθετο. Τρεις ομάδες φαίνονται ξεκάθαρα κάτι που συμφωνεί με τα προηγούμενα αποτελέσματα αλλά και με την ανάλυση των δεδομένων που κάναμε προηγουμένως.



(Εικόνα 83) Σχεδιάγραμμα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων , με μέγεθος χάρτη 6x6, σε rsom , σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα . Με κόκκινες γραμμές για να τονιστούν τα όρια καλύτερα.

5.3.12 Αποτελέσματα των παραδειγμάτων

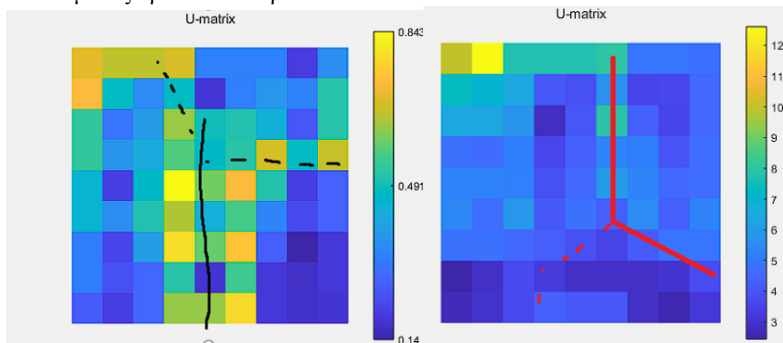
Μια από της μεγάλες διαφορές που είδαμε και στους δύο αλγορίθμους ήταν μεταξύ των πλεγμάτων τον αλγορίθμων. Παρόλο που στο προηγούμενο παράδειγμα με το iris είδαμε ότι η διαφορές πολλές φορές δεν ήταν αρκετά μεγάλες πέρα από την προοπτική που βλέπαμε τα δεδομένα . Στο συγκεκριμένο παράδειγμα (εικόνα 84) είδαμε και από τους δυο αλγορίθμους γιγάντια διαφορά στην ποιότητα των δεδομένων στο εξαγωνικό πλέγμα σε σχέση με το τετραγωνικό. Συγκεκριμένα μάλιστα στον rsom το εξαγωνικό πλέγμα μας έδωσε πάρα πολύ κακά αποτελέσματα τα οποία μειωνόντουσαν σε ποιότητα όσο μεγαλώναμε της διάστασης του αλγορίθμου. Παρόλα αυτά μεταξύ των δύο αλγορίθμων είδαμε ότι σε γενικές γραμμές ο κανονικός som μας έδειχνε καλύτερα αποτελέσματα στο εξαγωνικό πλέγμα στο umatrix



(Εικόνα 84) Σχεδιάγραμμα umatrix μεγέθους χάρτη 5x5 κανονικού som από τα αριστερά και rsom από τα δεξιά σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα . Με γραμμές για να τονιστούν τα όρια των ομάδων και διακεκομμένες γραμμές για δευτερεύων όρια.

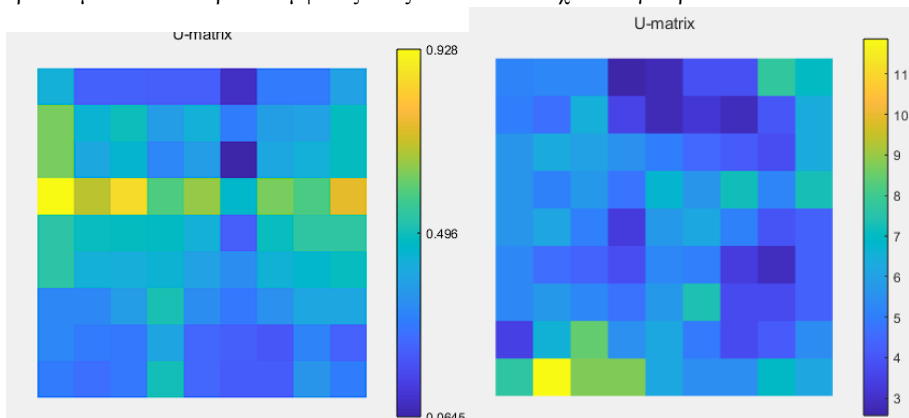
Παραπάνω μπορούμε να δούμε την διαφορά στο umatrix μεταξύ των δύο αλγόριθμών στις διαστάσεις 5x5 όπου ήδη βλέπουμε ότι μπορούμε να χαρακτηρίσουμε καλύτερα ομάδες στον κανονικό som στα αριστερά ενώ στον rsom δυσκολευόμαστε να βρούμε σαφή όρια .

Όσο αφορά όμως το τετραγωνικό πλέγμα (εικόνα 85) και η δύο αλγόριθμοι διαχειρίστηκαν τα δεδομένα αρκετά καλά. Η διαφοράς τους είναι κυρίως στο πόσο καθαρά δείχνανε τα δεδομένα και ξεκινώντας από της διαστάσεις μεγέθους 5x5 ξεκίνησαν και η δύο αλγόριθμοι να εντοπίζουν και 3 ομάδες αρκετά σταθερά.



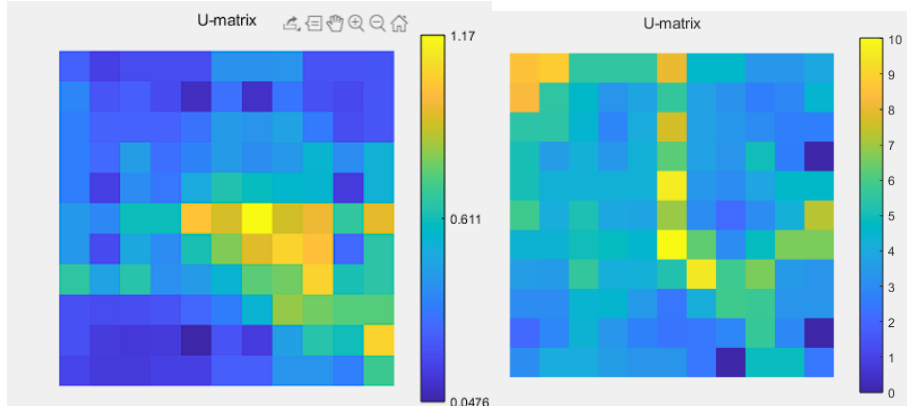
(Εικόνα 85) Σχεδιάγραμμα umatrix μεγέθους χάρτη 5x5 κανονικού som από τα αριστερά και rsom από τα δεξιά σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα . Με γραμμές για να τονιστούν τα όρια των ομάδων και διακεκομμένες γραμμές για δευτερεύων όρια.

Βλέπουμε ότι στο πρώτο παράδειγμα 5x5 σε επίπεδο σχήμα και η δύο αλγόριθμοι κατάφεραν να εντοπίσουν και της δύο ομάδες, συγκεκριμένα φάνηκε και μια μικρή τέταρτη ομάδα στον rsom κάτι όμως που θα μπορούσε να θεωρηθεί θόρυβος μεταξύ των δεδομένων μιας και δεν είδαμε κάτι παρόμοιο στα αποτελέσματα σε άλλα παραδείγματα. Γενικότερα όμως το καθαρό μοτίβο τριών ομάδων είναι αρκετά εμφανές ίσως στον rsom ελάχιστα λιγότερο.



(Εικόνα 86) Σχεδιάγραμμα umatrix μεγέθους χάρτη 5x5 κανονικού som από τα αριστερά και rsom από τα δεξιά σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα .

Παραπάνω (εικόνα 86) βλέπουμε το παράδειγμα σε κυλινδρικά σχήματα των ιδίων διαστάσεων και σε αντίθεση με τον κανονικό som μπορούμε να δούμε ότι ο rsom διαχώρισε πάρα πολύ καλά τις τρεις ομάδες ενώ ο κανονικός som παρόλο που και σε αυτόν μπορούν να παρατηρηθούν 3 ξεκάθαρες ομάδες υπάρχει μια μικρή θόλωση μεταξύ των ορίων.



(Εικόνα 87) Σχεδιάγραμμα umatrix μεγέθους χάρτη 6x6 κανονικού som από τα αριστερά και rsom από τα δεξιά σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

Τα καλύτερα αποτελέσματα όμως σίγουρα προήλθαν από τον rsom στις 6x6 διάστασης όπου μπορούμε να δούμε (εικόνα 87) με πάρα πολύ μεγάλη ευκρίνεια τρεις ξεκάθαρες ομάδες ενώ ταυτόχρονα ο som παρόλο που και εκεί μπορούμε να δούμε τρεις ομάδες η εικόνα τους είναι ξεκάθαρα πιο θολή και μικρότερης ευκρινείας.

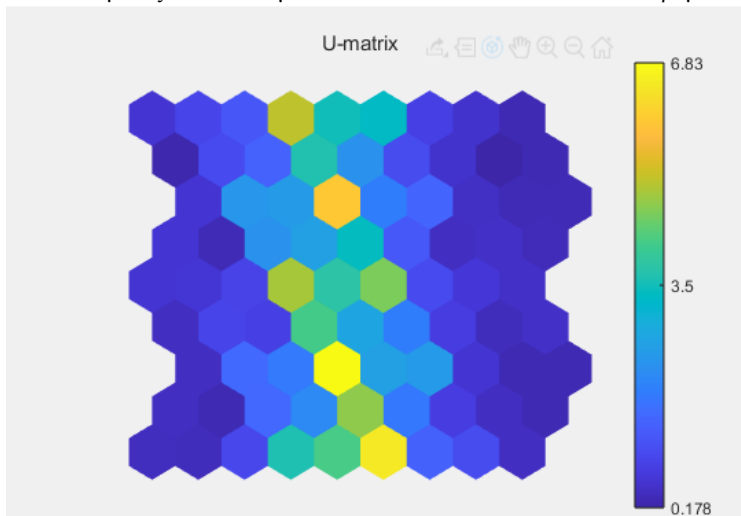
Τελικά συμπεράσματα είναι ότι μεταξύ των δύο αλγορίθμων ο som σιγουρά είχε καλύτερη ανάλυση στο εξαγωνικό πλέγμα με τα καλύτερα αποτελέσματα να εμφανίζονται στις διαστάσεις 5x5 σε κυλινδρικό πλέγμα κάτι το οποίο ο rsom δεν κατάφερε να μας δώσει επική αποτελέσματα μετά τις διαστάσεις 4x4 τα οποία λόγω των μικρών διαστάσεων δεν ήταν και τα καλύτερα. Σε γενικές γραμμές και η δύο αλγόριθμοι κατάφεραν να ομαδοποιήσουν τα δεδομένα αρκετά καλά σε διαστάσεις 5x5 και μετά αλλά σιγουρά ο rsom το έκανε πιο ξεκάθαρα από ότι ο κανονικός som. Βέβαια είναι άξιο επίσης να αναφερθεί ότι μεταξύ των δυο αλγορίθμων μόνο ο κανονικός som μπορούσε να συνεχίσει να ομαδοποιεί τα δεδομένα σε διαστάσεις μεγαλύτερες του 6x6 ενώ ο rsom αποτύχανε χωρίς να παίρνουμε αποτελέσματα. Οπότε ο κανονικός som είναι πιο ευέλικτος ενώ ο rsom μας έδωσε καλύτερα αποτελέσματα πιο γρήγορα μέχρι όμως ένα σημείο.

5.4 Παράδειγμα dissimilarity data

Προχωρώντας στο τελευταίο παράδειγμα των αλγόριθμων φτάνουμε στα Dissimilarity data. Όπως προαναφέρθηκε και στο κεφάλαιο με την ανάλυση των δεδομένων το συγκεκριμένο παράδειγμα βρέθηκε σαν ένα δοκιμαστικό σετ δεδομένων για να δώσει της δυνατότητες του rsom μέσα στο λογισμικό. Πρόκειται για ένα 200×200 σε διαστάσεις σετ δεδομένων και είναι ένα σετ δεδομένων το οποίο χωρίζεται σε δύο αρκετά ξεκάθαρες ομάδες. Με αυτό το παράδειγμα μπορούμε να δούμε όχι μόνο πως οι αλγόριθμοι διαχειρίζονται πιο απλά δεδομένα αλλά και πιθανών δυσκολίες που μπορεί να έχουν προσπαθώντας να τα διαχειριστούν. Τα παραδείγματα που θα δειχθούν θα είναι σε διαστάσεις 5×5 , 10×10 και 17×17 εκτός από ένα παράδειγμα στον rsom το οποίο θα αναφερθεί παρακάτω το οποίο είναι σε 16×16 σε διαστάσεις. Κάτι που έγινε λόγω limitation του ίδιου του rsom. Η ρυθμίσεις εκπαίδευσης που θα χρησιμοποιηθούν είναι ίδιες με τα προηγούμενα παραδείγματα δηλαδή 500 επαναλήψεις ρόνο εκπαίδευσης σε batch training με της ίδιες ρυθμίσεις του umatrix που έχουν αναδειχθεί ως τώρα.

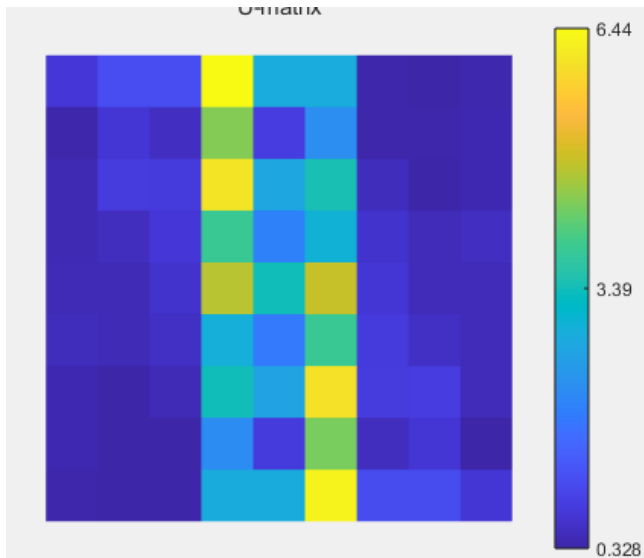
5.4.1 Διαστάσεις χάρτη 5×5 σε επίπεδο σχήμα σε κανονικό som

Ξεκινώντας με το εξαγωνικό πλέγμα σε διαστάσεις 5×5 μπορούμε να δούμε ότι ο αλγόριθμος έχει αρκετά ξεκάθαρα βρει και της δύο ομάδες αρκετά ξεκάθαρα. Μπορούμε να δούμε (εικόνα 88) ένα αρκετά ξεκάθαρο όριο μεταξύ των δύο ομάδων με μερικές αποστάσεις να είναι ελάχιστα πιο κοντά μεταξύ των δύο ομάδων αλλά τίποτα που να αλλοιώνει δραματικά το umatrix



(Εικόνα 88) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας, με μέγεθος χάρτη 5×5 , σε κανονικό som, σε επίπεδο σχήμα, σε εξαγωνικό πλέγμα.

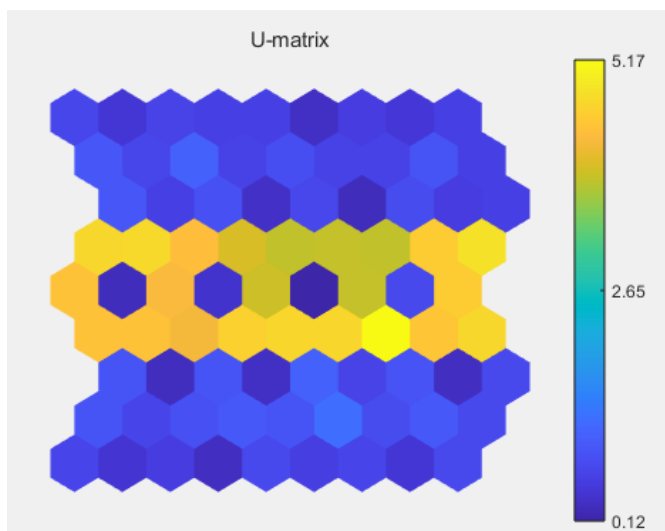
Ταυτόχρονα μπορούμε επίσης να δούμε (εικόνα 89) στο τετραγωνικό πλέγμα ότι επίσης η ομαδοποίηση έγινε με παρόμοια ποιότητα με το εξαγωνικό πλέγμα και ίσως ελάχιστα πιο καθαρά από την μεριά του ορίου.



(Εικόνα 89) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιοτητας , με μέγεθος χάρτη 5x5, σε κανονικό som , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

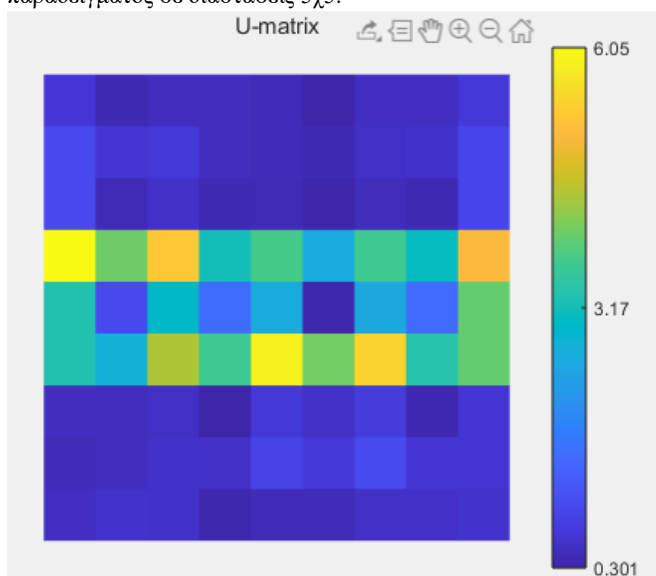
5.4.2 Διαστάσεις χάρτη 5x5 σε κυλινδρικό σχήμα σε κανονικό som

Στο κυλινδρικό σχήμα μπορούμε να δούμε(εικόνα 90) μια διαφορά στα όρια η οποία είναι το αρκετά πιο ξεκάθαρο όριο μεταξύ των ομάδων με τις μεγαλύτερες αποστάσεις. Σε σχέση με το επίπεδο σχήμα βλέπουμε ότι οι αποστάσεις μεταξύ των δύο ομάδων είναι οι μεγαλύτερες δυνατές.



(Εικόνα 90) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 5x5, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα .

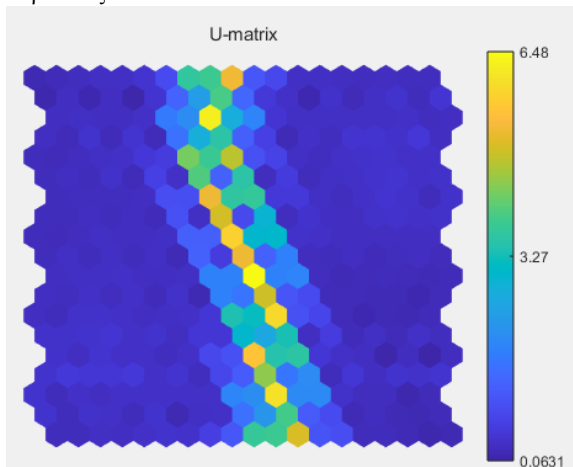
Ταυτόχρονα βλέπουμε ότι στο τετραγωνικό πλέγμα (εικόνα 91) έχουμε παρόμοια αποτελέσματα με το επίπεδο σχήμα χωρίς κάποια διαφορά επίσης με τα ίδια αποτελέσματα του επίπεδου παραδείγματος σε διαστάσεις 5x5.



(Εικόνα 91) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 5x5, σε κανονικό som , σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα πλέγμα .

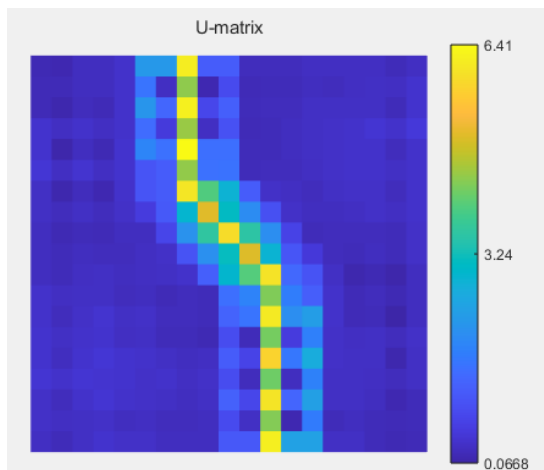
5.4.3 Διαστάσεις χάρτη 10x10 σε επίπεδο σχήμα σε κανονικό som

Στο εξαγωνικό πλέγμα διαστάσεων 10x10 ξεκινάμε να βλέπουμε (εικόνα 92) περισσότερες λεπτομέρειες στον χάρτη. Κάτι που φαινόταν στο μικρότερο παράδειγμα ήταν ότι η ομάδες ήταν ελάχιστα ανισόρροπες στο πάνω και κάτω άκρο τους με τους νευρώνες να είναι πιο κοντά ο ένας με τον άλλο . Βλέπουμε εδώ ότι η νευρώνες είναι ακόμα πιο διασκορπισμένη και μάλιστα ξεκινάνε να δείχνουνε αυτή την συσσώρευση επεκτείνοντας της ομάδες στο πάνω και κάτω άκρο τους.



(Εικόνα 92) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 10x10, σε κανονικό som , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

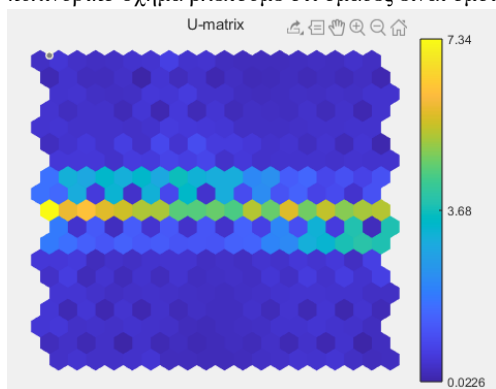
Στο τετραγωνικό πλέγμα βλέπουμε (εικόνα 93) επίσης ότι ο αλγόριθμος ξεκινάει να παίρνει ένα σχήμα μεταξύ των δύο ομάδων το οποίο είναι παρόμοιο με το σχήμα του umatrix του παραδείγματος 5x5 . Όπου βλέπουμε στις ομάδες τα πάνω και κάτω άκρα τους να έχουν πιο πυκνές και κοντινές αποστάσεις ενώ ταυτόχρονα το όριο φαινόταν λίγο πιο θολό. Εδώ πέρα μπορούμε να δούμε ότι ο χάρτης ξεκινάει να παίρνει το σχήμα τους σε μεγαλύτερες διαστάσεις και να της δείχνει καλύτερα. Όπως και στο εξαγωνικό πλέγμα αλλά ελάχιστα πιο ξεκάθαρα και με πολύ πιο καθαρά όρια στην ομάδα από ότι το εξαγωνικό πλέγμα.



(Εικόνα 93) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 10x10, σε κανονικό som , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα .

5.4.4 Διαστάσεις χάρτη 10x10 σε κυλινδρικό σχήμα σε κανονικό som

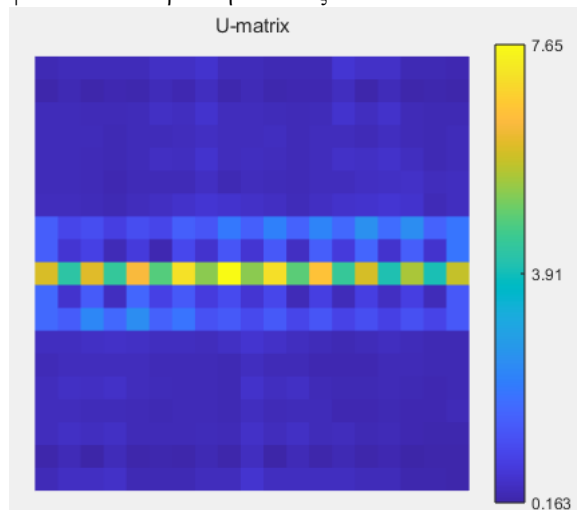
Προχωρώντας στο κυλινδρικό σχήμα μπορούμε να παρατηρήσουμε (εικόνα 94) πιο εύκολα ότι σε σχέση με το επίπεδο το umatrix δεν μας δίνει κάποιο σχήμα στα δεδομένα μας. Δηλαδή εκεί που στα προηγούμενα παραδείγματα βλέπαμε μια συσσώρευση των δεδομένων στο πάνω και κάτω μέρος τους κάτι που στις μεγαλύτερες διαστάσεις φάνηκε και στα όρια τους .Στο κυλινδρικό σχήμα βλέπουμε ότι ομάδες είναι ομοιόμορφες και ίσιες .



(Εικόνα 94) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 10x10, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα .

Στο τετραγωνικό πλέγμα βλέπουμε (εικόνα 95) παρόμοια αποτελέσματα με της δύο ομάδες να

χωρίζονται από το ίδιο όριο όμως με μια διαφορά σε σχέση με το εξαγωνικό πλέγμα που φαίνεται καλύτερα στην ενώσεις.

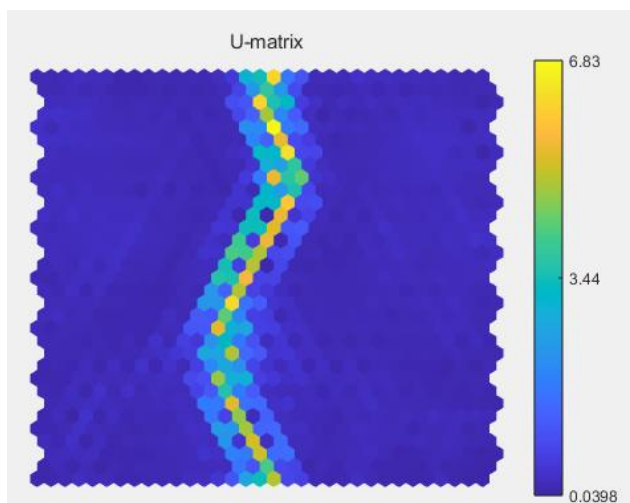


(Εικόνα 95) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιοότητας , με μέγεθος χάρτη 10x10, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα .

Σε σχέση με το εξαγωνικό πλέγμα βλέπουμε ότι η αποστάσεις των νευρώνων κοντά στα όρια δεν ακολουθούν το ίδιο παράδειγμα με την ένωση μεταξύ των ομάδων να της κάνει να είναι πιο κοντινές αλλά η διασπορά των νευρώνων βλέπουμε να είναι παρόμοια.

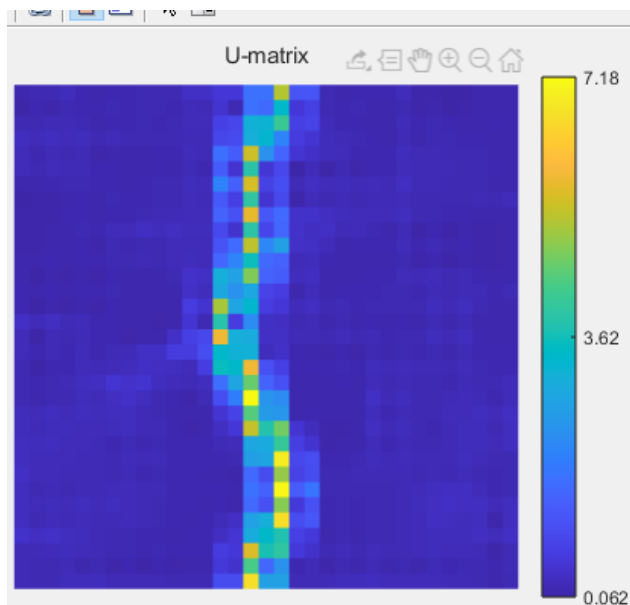
5.4.5. Διαστάσεις χάρτη 17x17 σε επίπεδο σχήμα σε κανονικό som

Προχωρώντας προς τον μεγαλύτερο σε διαστάσεις χάρτη νευρώνων στο εξαγωνικό πλέγμα βλέπουμε παρόμοια αποτελέσματα με τα προηγούμενα αλλά αυτή την φορά με περισσότερες γωνίες μεταξύ του ορίου των ομάδων. Γενικότερά βλέπουμε (εικόνα 96) ότι η ομάδα μας δεν μας δείχνουν κάποια μεγάλη διαφορά στα αποτελέσματα πέρα από το σχήμα του umatrix και δεν έχει γίνει κάποια ριζική αλλαγή πέρα της ελαφράς αλλοίωσης του ορίου σε μερικά σημεία του αλλά δεν έχει φτάσει σε σημείο που μπερδεύει τα δεδομένα.



(Εικόνα 96) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 17χ17, σε κανονικό som , σε επίπεδο σχήμα ,σε εξαγωνικό πλέγμα .

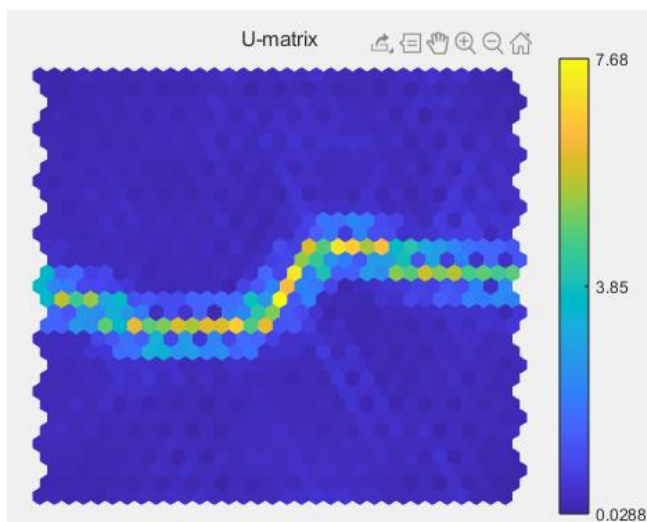
Επίσης για το τετραγωνικό πλέγμα (εικόνα 97) δεν υπάρχει κάτι ιδιαίτερο να αναφερθεί αφού ομοίως με το εξαγωνικό βλέπουμε παρόμοιο σχήμα στο umatrix , δύο ομάδες καθαρές και ελάχιστη αλλοίωση στο όριο μεταξύ των ομάδων.



(Εικόνα 97) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 17x17, σε κανονικό som , σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

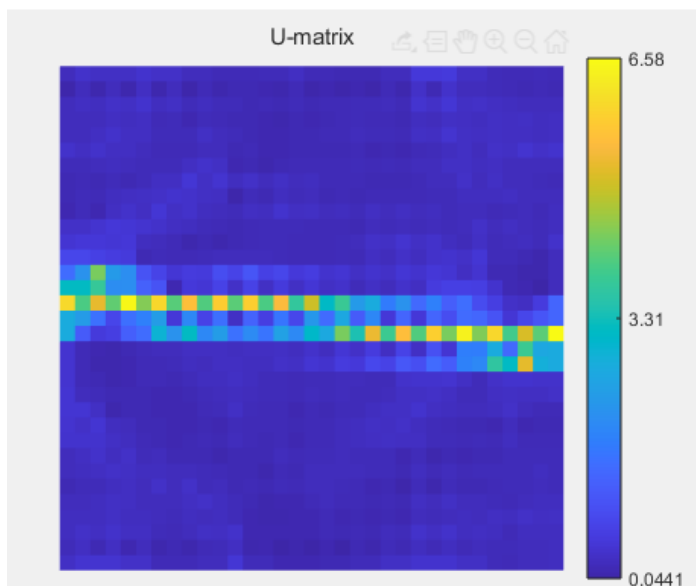
5.4.6 Διαστάσεις χάρτη 17x17 σε κυλινδρικό σχήμα σε κανονικό som

Προχωρώντας στο κυλινδρικό παράδειγμα βλέπουμε (εικόνα 98) ότι σε σχέση με το 10x10 αρχίζουμε να βλέπουμε το όριο των δύο ομάδων να ξεκινάει να παίρνει το σχήμα που είδαμε και στο επίπεδο σχήμα. Με της δύο ομάδες να είναι ακόμα ξεκάθαρες και παρατηρώντας παρόμοια αλλοίωση με το επίπεδο σχήμα επίσης στο όριο των ομάδων.



(Εικόνα 98) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιοτήτας , με μέγεθος χάρτη 17χ17, σε κανονικό som , σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα .

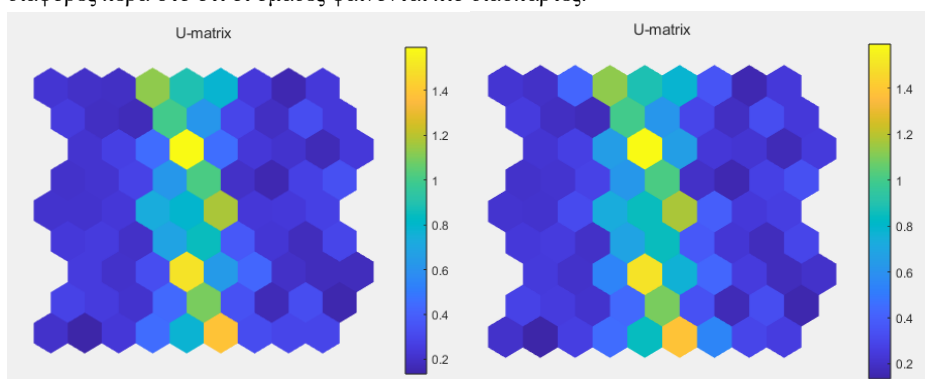
Αντιθέτως με το εξαγωνικό πλέγμα όπου είδαμε ότι έδειξε πιο πολύ το σχήμα που βλέπαμε στα παραδείγματα με επίπεδο σχήμα στο τετραγωνικό πλέγμα βλέπουμε (εικόνα 99) ότι το σχήμα παραμένει αρκετά πιο ίσιο όπως και στο προηγούμενο παράδειγμα. Ταυτόχρονα η αλλοίωση του ορίου μεταξύ των ομάδων μπορεί επίσης να παρατηρηθεί ότι είναι αρκετά μικρότερη.



(Εικόνα 99) Σχεδιάγραμμα umatrix σετ δεδομένων ανομοιότητας , με μέγεθος χάρτη 17x17, σε κανονικό som , σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα .

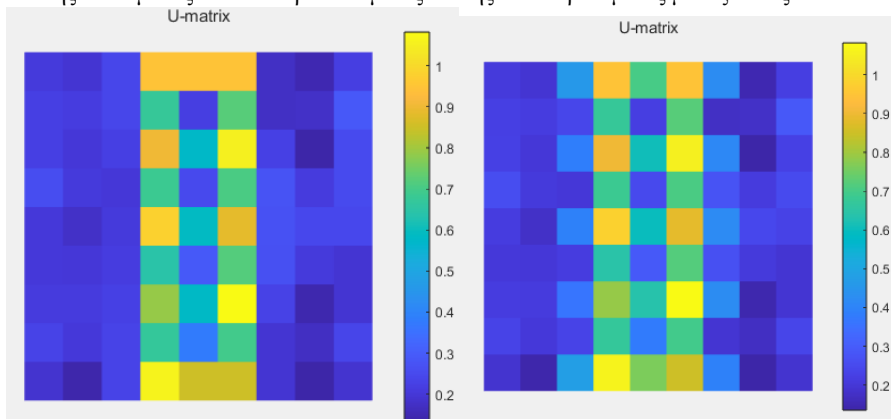
5.4.7 Διαστάσεις χάρτη 5x5 σε επίπεδο σχήμα σε rsom

Παρόμοια με τον κανονικό som βλέπουμε (εικόνα 100) ότι ο αλγόριθμος διαχωρίζει επίσης της δύο ομάδες αρκετά ξεκάθαρα σε μικρές διαστάσεις 5x5 μάλιστα τα αποτελέσματα που παίρνουμε τουλάχιστον προς το όριο θα έλεγε κάποιος ήταν σχεδόν όμοια χωρίς ιδιαίτερες διαφορές πέρα στο ότι οι ομάδες φαίνονται πιο διάσπαρτες.



(Εικόνα 100) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα, σε εξαγωνικό πλέγμα.

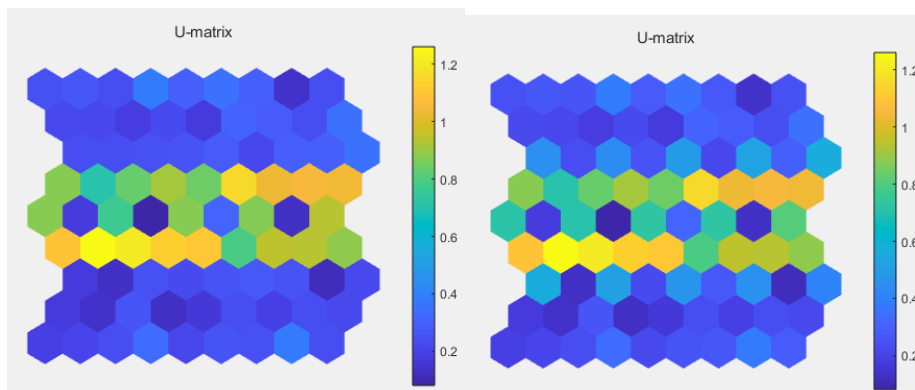
Από την άλλη στο τετραγωνικό πλέγμα βλέπουμε (εικόνα 101) ένα πάρα πολύ ξεκάθαρο όριο σε αντίθεση με το εξαγωνικό πλέγμα είναι πολύ πιο ξεκάθαρο όριο σε σχέση με τον κανονικό som. Με της δύο ομάδες να είναι αρκετά όμοιος επίσης διασκορπισμένες μεταξύ τους.



(Εικόνα 101) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε επίπεδο σχήμα, σε τετραγωνικό πλέγμα.

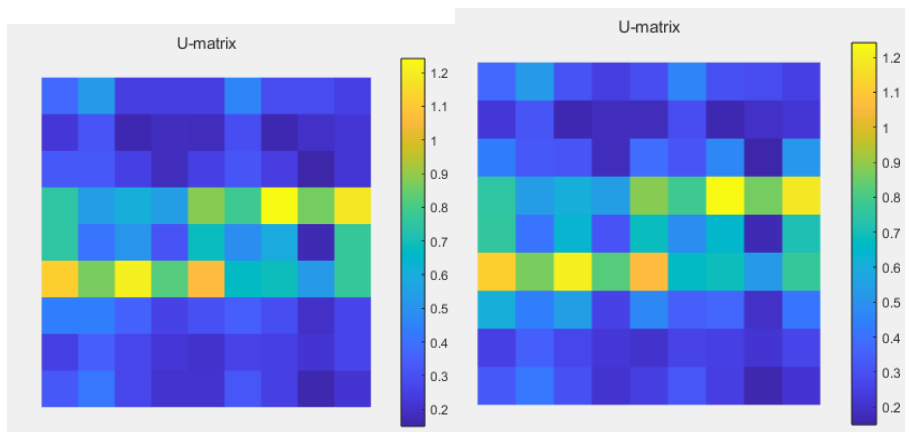
5.4.8 Διαστάσεις χάρτη 5x5 σε κυλινδρικό σχήμα σε rsom

Προχωρώντας στο κυλινδρικό σχήμα του παραδείγματος βλέπουμε (εικόνα 102) μερικές ενδιαφέρων εξελίξεις αρχικά βλέπουμε το όριο των δύο ομάδων μιμείται σε σχήμα το όριο στο εξαγωνικό πλέγμα του επίπεδου σχήματος σε σχέση με το προηγούμενο παράδειγμα του εξαγωνικού πλέγματος το οποίο έμοιαζε πάρα πολύ με το παράδειγμα του κανονικού som.



(Εικόνα 102) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

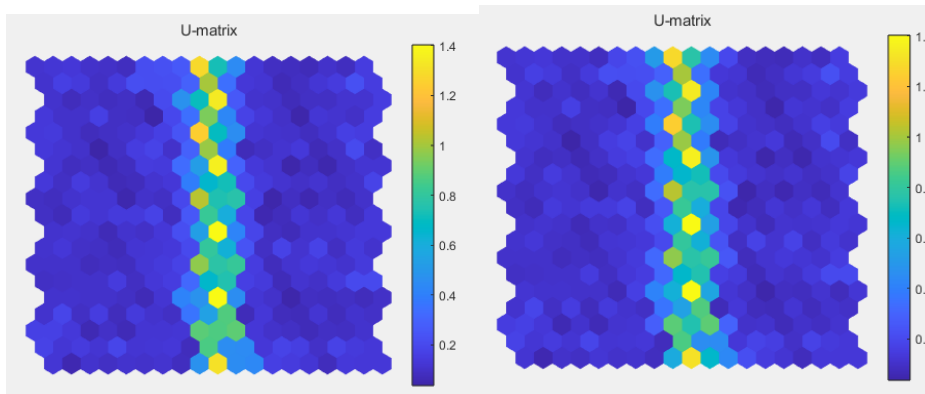
Από την άλλη όμως βλέπουμε (εικόνα 103) το τετραγωνικό πλέγμα να ξεκινάει να εμφανίζει το σχήμα που είδαμε να εμφανίζεται στο παράδειγμα 10x10 διαστάσεων του κανονικού som σε επίπεδο σχήμα όπου η δύο ομάδες διακατείχαν ένα μεγαλύτερο κομμάτι στα άκρα του αλγορίθμου. Βέβαια να τονιστεί ότι το όριο συνεχίζει να είναι αρκετά ξεκάθαρο και να εμφανίζει το σχήμα μιας ευθείας γραμμής που στα άκρα της ξεκινάει να μειώνει της αποστάσεις της.



(Εικόνα 103) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 5x5, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

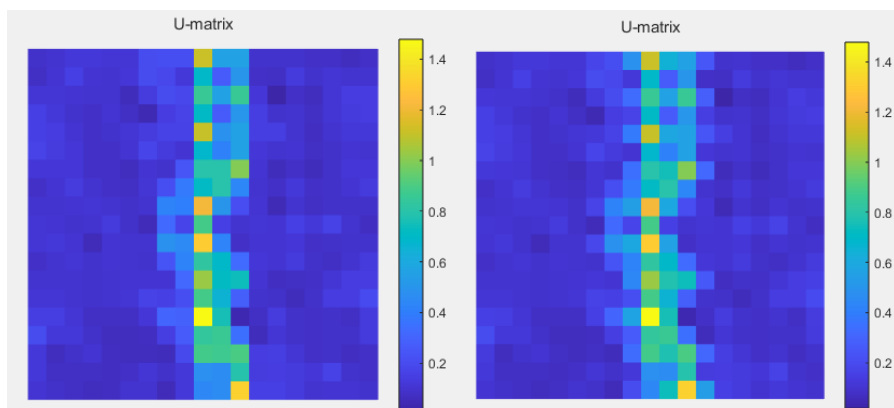
5.4.9 Διαστάσεις χάρτη 10x10 σε επίπεδο σχήμα σε rsom

Προχωρώντας στο παράδειγμα με διαστάσεις χάρτη νευρώνων 10x10 βλέπουμε ότι δεν υπάρχει κάποια αλλαγή στο όριο σε σχέση με το προηγούμενο παράδειγμα σε επίπεδο σχήμα. Στο εξαγωνικό πλέγμα του umatrix συνεχίζουμε να βλέπουμε (εικόνα 104) ένα ξεκάθαρο ευθείς όριο το οποίο διαχωρίζει της δύο ομάδες ομοιόμορφα ενώ η ομάδες επίσης είναι ξεκάθαρες και ενιαίες μεταξύ τους.



(Εικόνα 104) Σχεδιαγράμματα umatrix σε δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε επίπεδο σχήμα, σε εξαγωνικό πλέγμα.

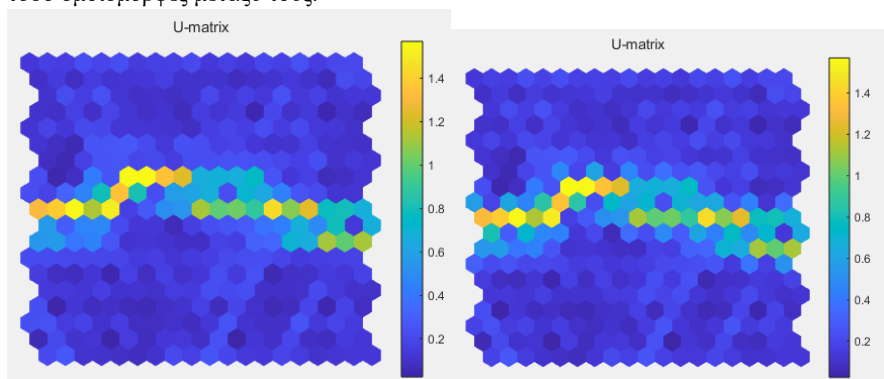
Ταυτόχρονα στο τετραγωνικό πλέγμα του παραδείγματος βλέπουμε (εικόνα 105) ότι δεν ακολουθείτε το ίδιο παράδειγμα με το εξαγωνικό και παρατηρούμε ότι ξεκινάει να υπάρχει μια διαφορά στο όριο. Ξεκινάει να εμφανίζεται περισσότερη λεπτομέρεια στο όριο αναμεσα των δύο ομάδων το οποίο θυμίζει λίγο το όριο που είδαμε στο κυλινδρικό σχήμα αλλά συνεχίζει να είναι ένα ευθείων όριο το οποίο χωρίζει της δύο ομάδες με της ομάδες να συνεχίζουν αν είναι ίσες αναμεσα τους.



(Εικόνα 105) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε επίπεδο σχήμα, σε τετραγωνικό πλέγμα.

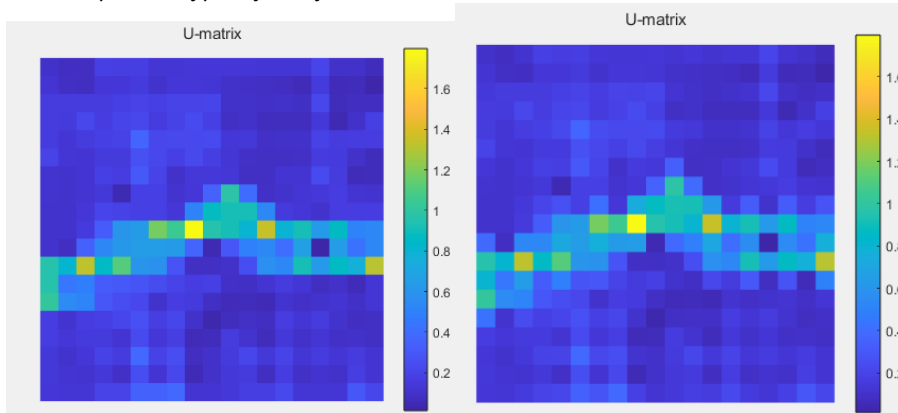
5.4.10 Διαστάσεις χάρτη 10x10 σε κυλινδρικό σχήμα σε rsom

Στο εξαγωνικό πλέγμα του κυλινδρικού σχήματος μπορούμε επίσης να δούμε ότι το όριο έχει ξεκινήσει να δείχνει παραπάνω σχήμα αλλά βλέπουμε (εικόνα 106) επίσης για πρώτη φορά ότι το όριο μεταξύ των ομάδων δεν είναι πια ευθείς. Αλλά έχει ξεκινήσει να διαμορφώνεται πιο πολύ στις ομάδες ταυτόχρονα βλέπουμε ότι η δύο ομάδες αναμεσά τους ξεκινάνε να μην είναι τόσο ομοιόμορφες μεταξύ τους.



(Εικόνα 106) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε κυλινδρικό σχήμα, σε εξαγωνικό πλέγμα.

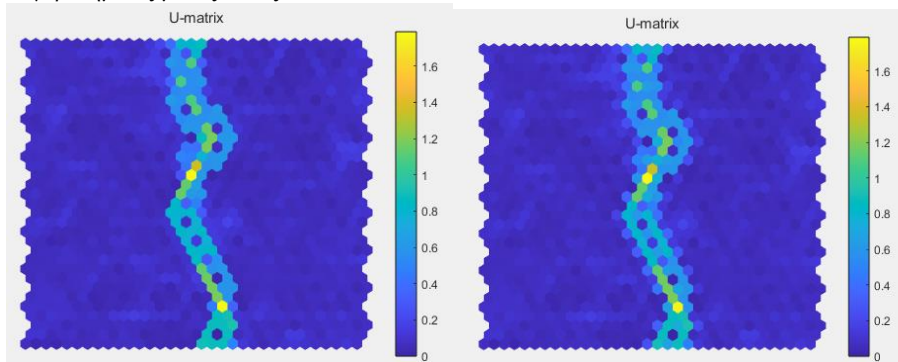
Παρόμοια αποτελέσματα μπορούμε να δούμε και στο εξαγωνικό πλέγμα ,μάλιστα στον rsom είναι η πρώτη φορά που μπορούμε να δούμε (εικόνα 107) τα δύο πλέγματα να συμφωνούν στα αποτελέσματα τους μεταξύ τους.



(Εικόνα 107)Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 10x10, σε rsom, σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα.

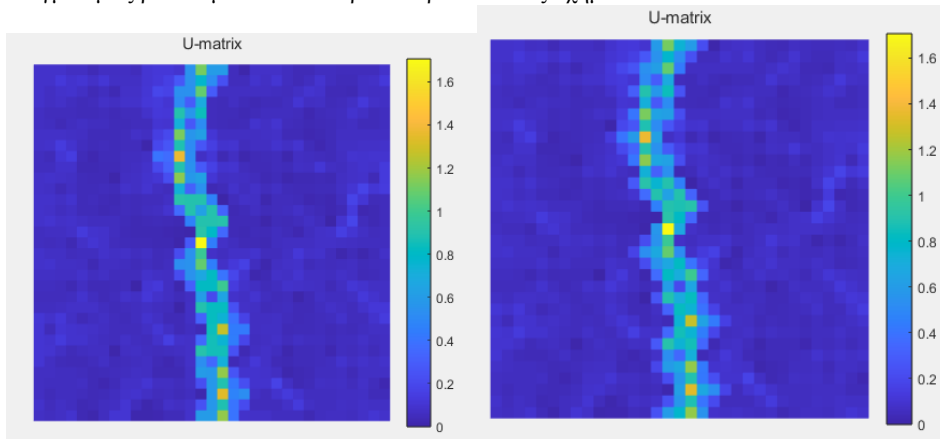
5.4.11 Διαστάσεις χάρτη 17x17 σε επίπεδο σχήμα σε rsom

Προχωρώντας στο μεγαλύτερο σε διαστάσεις του rsom μπορούμε να δούμε (εικόνα 108) ότι ξεκινάμε και στο επίπεδο σχήμα στο εξαγωνικό πλέγμα να βλέπουμε το όριο μεταξύ των ομάδων να μην είναι πια ευθείς αλλά να παίρνει σχήμα όπως είχαμε δει να δείχνουν τα κυλινδρικά σχήματα στα παραδείγματα του αλγορίθμου. Ταυτόχρονα όμως βλέπουμε ότι η ομάδες είναι συγκροτημένες μεταξύ τους .



(Εικόνα 108) Σχεδιαγράμματα umatrix σε δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 17x17, σε rsom, σε επίπεδο σχήμα, σε εξαγωνικό πλέγμα.

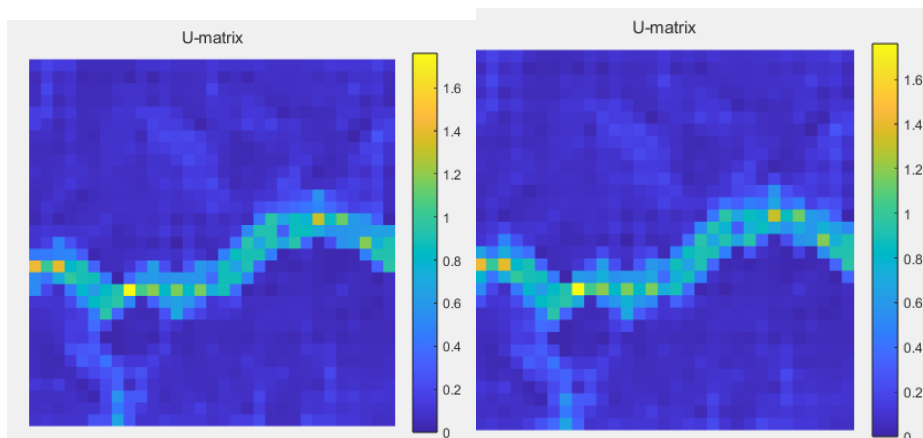
Στο τετραγωνικό πλέγμα επίσης βλέπουμε (εικόνα 109) το όριο να εμφανίζει περισσότερη λεπτομέρεια με τις δύο ομάδες να είναι ομοιόμορφες μέσα τους. Σε σχέση με το εξαγωνικό πλέγμα όμως βλέπουμε ότι είναι περισσότερο σε ευθείς σχήμα.



(Εικόνα 109) Σχεδιαγράμματα umatrix σε δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 17x17, σε rsom, σε επίπεδο σχήμα, σε τετραγωνικό πλέγμα.

5.4.12 Διαστάσεις χάρτη 16x16 σε κυλινδρικό σχήμα σε rsom

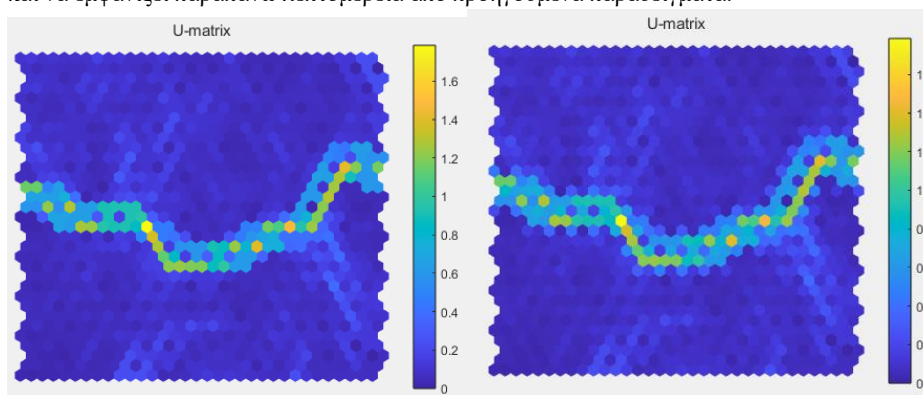
Προχωρώντας στο παράδειγμα διαστάσεων 16x16 σε τετραγωνικό πλέγμα που όπως προαναφέρθηκε είναι σε μικρότερη διάσταση λόγω limitation του rsom και προβλημάτων που εμφάνιζε ξεκινάμε να βλέπουμε κάτι πολύ ενδιαφέρον. Αρχικά βλέπουμε (εικόνα 110) ότι το όριο μεταξύ των δύο ομάδων έχει πάρει ένα αρκετά περιγραφικό σχήμα όμως βλέπουμε ότι σε μια από τις ομάδες ξεκινάει να εμφανίζεται μια τρίτη μικρότερη ομάδα. Αυτό είναι κάτι που βλέπουμε πρώτη φορά και στους δύο αλγόριθμους όχι μόνο στο τετραγωνικό αλλά και από το εξαγωνικό πλέγμα.



(Εικόνα 110) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος χάρτη 16x16, σε rsom, σε κυλινδρικό σχήμα, σε τετραγωνικό πλέγμα.

5.4.13 Διαστάσεις χάρτη 17x17 σε κυλινδρικό σχήμα σε rsom

Τέλος έχουμε το τελευταίο παράδειγμα σε διαστάσεις νευρώνων 17x17 (εικόνα 111) σε κυλινδρικό σχήμα και μπορούμε και σε αυτό το παραδειγμα την τρίτη ομάδα να φαίνεται ελάχιστα. Ταυτόχρονα όμως μπορούμε να παρατηρήσουμε και άλλες μικρότερες ομάδες να ξεκινάνε να διαχωρίζονται. Το κεντρικό όριο μεταξύ των ομάδων συνεχίζει να είναι ξεκάθαρο και να εμφανίζει παραπάνω λεπτομέρεια από προηγούμενα παραδείγματα.

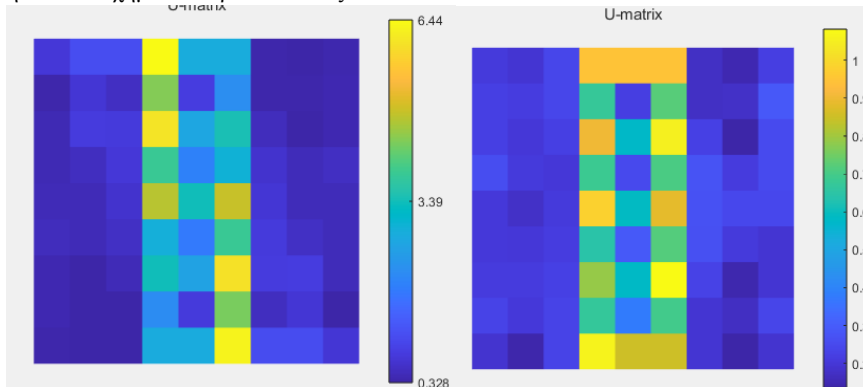


(Εικόνα 111) Σχεδιαγράμματα umatrix σετ δεδομένων υπολειμμάτων αμινοξέων στα αριστερά σε διάμεσες αποστάσεις και στα δεξιά σε μέσες αποστάσεις μεταξύ των νευρώνων, με μέγεθος

χάρτη 17x17, σε rsom, σε κυλινδρικό σχήμα ,σε εξαγωνικό πλέγμα.

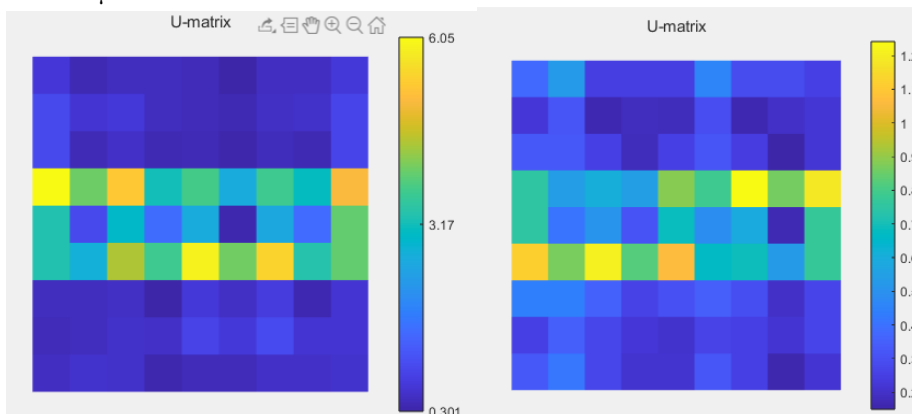
5.5.14 Σύγκριση των αποτελεσμάτων

Γενικότερα από τα αποτελέσματα είδαμε ότι και οι δύο αλγόριθμοι κατάφεραν να ομαδοποιήσουν της ομάδες με επιτυχία και ξεκάθαρα. Η κύρια διαφορά στα αποτελέσματα τους ήταν στο σχήμα του ορίου και πως το επιδείκνυναν στο umatrix .



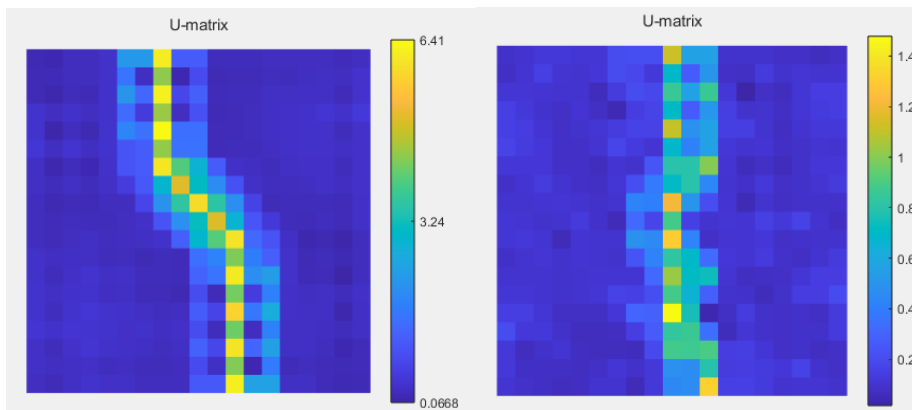
(Εικόνα 112) Σχεδιάγραμμα umatrix μεγέθους χάρτη 5x5 κανονικού som από τα αριστερά και rsom από τα δεξιά σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα .

Αρχικά από τα πρώτα παραδείγματα διαστάσεων 5x5 βλέπουμε (εικόνα 112) ότι και οι δύο αλγόριθμοι κάνουν μια πολύ καλή διαχωροποίηση των ομάδων με τον κανονικό som στα αριστερά να δείχνει ελάχιστα καλύτερα το σχήμα των ομάδων σε επίπεδο σχήμα . Ενώ το rsom βλέπουμε ότι κατάφερε να απλά ομαδοποίηση της ομάδες αλλά με πολύ παρόμοια αποτελέσματα.



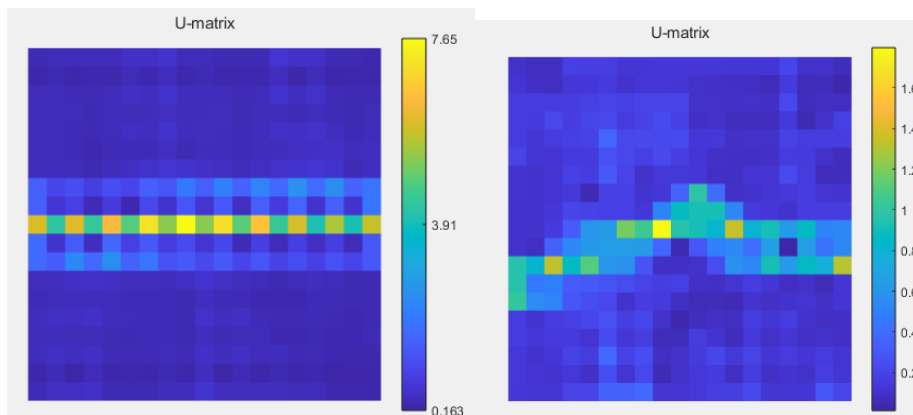
(Εικόνα 113) Σχεδιάγραμμα umatrix μεγέθους χάρτη 5x5 κανονικού som από τα αριστερά και rsom από τα δεξιά σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα

Όμως μετά στο κυλινδρικό σχήμα (εικόνα 113) στο τετραγωνικό πλέγμα των δύο αλγορίθμων είδαμε ότι το αντίθετο άρχισε να γίνεται. Κάτι που παρατηρήθηκε στους δύο αλγορίθμους ήταν ότι τα αποτελέσματα τους σε μικρότερους χάρτες μοιάζανε αλλά όχι για παρόμοια σχήματα, αλλά αντιθέτως μοιάζανε με το διαφορετικό τους σχήμα μεταξύ τους. Ο κανονικός som για παράδειγμα είχε μια τάση να δείχνει περισσότερο σχήμα σε επίπεδα σχήματα χάρτη και πιο ευθεία όρια και ξεκάθαρα στο κυλινδρικό σχήμα ενώ ο rsom έκανε το ακριβώς αντίθετο δηλαδή έδειχνε περισσότερο σχήμα στο όριο στο κυλινδρικό σχήμα και πιο ευθεία όρια στο επίπεδο σχήμα.



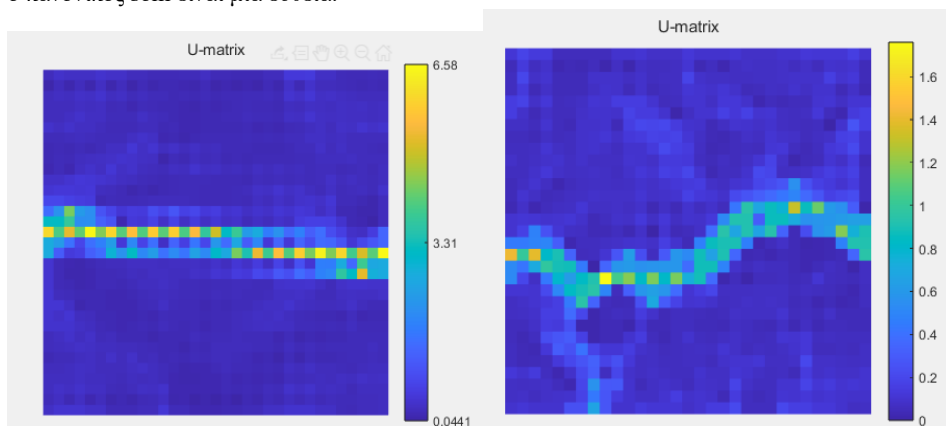
(Εικόνα 114) Σχεδιάγραμμα umatrix μεγέθους χάρτη 10x10 κανονικού som από τα αριστερά και rsom από τα δεξιά σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα.

Παραπάνω (εικόνα 114) μπορούμε να παρατηρήσουμε το φαινόμενο αυτό και είναι και η πρώτη φορά που βλέπουμε μεγάλες διαφορές μεταξύ των αλγορίθμων. Ξεκινώντας από διαστάσεις 10x10 βλέπουμε ότι το κανονικό som στα αριστερά έχει πολύ ξεκάθαρα διαχωρίσει αλλά και δώσει ένα σχήμα στο όριο ενώ το rsom παρόλο που και αυτό ξεκινάει ελάχιστα να δίνει ένα σχήμα το όριο παραμένει για το μεγαλύτερο μέρος ευθείς.



(Εικόνα 115) Σχεδιάγραμμα umatrix μεγέθους χάρτη 10x10 κανονικού som από τα αριστερά και rsom από τα δεξιά σε κυλινδρικό σχήμα ,σε τετραγωνικό πλέγμα

Ταυτόχρονα μπορούμε να δούμε(εικόνα 115) ότι στο παράδειγμα με ίδιες διαστάσεις αλλά σε κυλινδρικό σχήμα το ακριβές αντίθετο συμβαίνει με το rsom να δείχνει περισσότερο σχήμα ενώ ο κανονικός som είναι μια ευθεία.



(Εικόνα 116) Σχεδιάγραμμα umatrix μεγέθους χάρτη 17x17 κανονικού som από τα αριστερά και μεγέθους χάρτη 16x16 rsom από τα δεξιά σε επίπεδο σχήμα ,σε τετραγωνικό πλέγμα

Όμως το πιο ενδιαφέρον παράδειγμα είναι στο παράδειγμα με της μεγαλύτερες διαστάσεις συγκεκριμένα στο κυκλικό σχήμα. Οπού βλέπουμε ότι ο rsom ξεκινάει να δείχνει μια τρίτη ομάδα κάτι που βλέπουμε πρώτη φορά . Βλέπουμε επίσης πιο ξεκάθαρα ότι όσο μεγαλώνουμε της διαστάσεις η ομάδες σταματάνε να είναι τόσο ομοιόμορφες στον rsom σε σχέση με τον κανονικό som . Κάτι που παρατηρήθηκε και στο παράδειγμα με τα δεδομένα iris οπού όσο μεγάλωναν η διαστάσεις τόσο χειρότερα αποτελέσματα παίρναμε.

6 Συμπεράσματα

Τα τελικά συμπεράσματα είναι ότι βλέπουμε ότι και η δύο αλγόριθμοι μπορούν να διαχειριστούν τα δεδομένα . Η διαφορά τους φαίνεται κυρίως στα μεγέθη του χάρτη όπου ο rsom μας έδειχνε καλύτερα αποτελέσματα σε μικρότερες διαστάσεις ενώ ο κανονικό som μας έδειχνε καλύτερα αποτελέσματα σε μεγαλύτερες διαστάσεις. Είναι επίσης σημαντικό να αναφερθεί ότι ο rsom εμφάνισε προβλήματα με τους χάρτες σε μεγάλες διαστάσεις που δεν επέτρεπαν να διεξαχθούν πειράματα. Γενικότερα όμως ο rsom παρείχε πιο καθαρά αποτελέσματα από τον κανονικό som και πολλές φορές παρείχε αποτελέσματα σε σημεία όπου ο κανονικός som δυσκολευόταν να μας διηξεί κάτι. Οπότε ο κανονικός som ήταν πιο προσαρμόσιμος στις ρυθμίσεις και τα δεδομένα που του δίνουμε αλλά ο rsom μας έδειχνε πολύ πιο ξεκάθαρα αποτελέσματα σε σχέση με τον som παρόλο τον τεχνικών δυσκολιών που εμφάνισε κάποιες φορές. Επίσης ο rsom φάνηκε να διαχειρίζεται καλύτερα μικρότερα σε διαστάσεις δεδομένα ενώ ο som δυσκολεύτηκε σε μικρότερα δεδομένα για παράδειγμα τα δεδομένα με της αποστάσεις grantham.

• ΒΙΒΛΙΟΓΡΑΦΙΑ

- (1) Fernando Bação, Victor Lobo, Marco Painho, The self-organizing map, the Geo-SOM, and relevant variants for geosciences, *Computers & Geosciences*, Volume 31, Issue 2, 2005, Pages 155-163, ISSN 0098-3004
- (2) Bação, Fernando & Lobo, Victor. (2005). GEO-SOM and ITS integration with geographic information systems. *WSOM 2005 - 5th Workshop on Self-Organizing Maps*.
- (3) Agarwal, Pragya & Skupin, André. (2008). Self-Organising Maps: Applications in Geographic Information Science. 10.1002/9780470021699.
- (4) Painho, M., Bacao, F. and Lobo, V. (2004) "The Self-Organizing Map and it's variants as tools for geodemographical data analysis: the case of Lisbon's Metropolitan Area."
- (5) Cottrell, Marie & Olteanu, Madalina & Rossi, Fabrice & Villa-Vialaneix, Nathalie. (2018). *Self-Organizing Maps, theory and applications*.
- (6) Energy functions for self-organizing maps Tom Heskes a RWCP, Theoretical Foundation SNN, University of Nijmegen Geert Grooteplein 21, 6252 EZ, Nijmegen, The Netherlands
- (7) La Rosa, Massimo & Di Fatta, Giuseppe & Gaglio, Salvatore & Giammanco, Giovanni & Rizzo, Riccardo & Urso, Alfonso. (2007). Soft Topographic Map for Clustering and Classification of Bacteria. 10.1007/978-3-540-74825-0_30.
- (8) La Rosa, Massimo & Rizzo, Riccardo & Urso, Alfonso. (2011). Soft Topographic Maps for Clustering and Classifying Bacteria Using Housekeeping Genes. *Advances in Artificial Neural Systems*. 2011. 10.1155/2011/617427.
- (9) *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016*. Γερμανία, Springer International Publishing, 2016.
- (10) Christopher M. Bishop, Markus Svensén, Christopher K.I. Williams, Developments of the generative topographic mapping, *Neurocomputing*, Volume 21, Issues 1–3, 1998, Pages 203-224, ISSN 0925-2312, [https://doi.org/10.1016/S0925-2312\(98\)00043-5](https://doi.org/10.1016/S0925-2312(98)00043-5).
- (11) Schmitt, C., Böck, M., & Kramer, S. (2011). SOM Biclustering of Gene Expression Data. *Machine Learning in Systems Biology*, 78.

(12) Marie Cottrell, Barbara Hammer, Alexander Hasenfuß, Thomas Villmann, Batch and median neural gas, *Neural Networks*, Volume 19, Issues 6–7, 2006, Pages 762-771, ISSN 0893-6080, <https://doi.org/10.1016/j.neunet.2006.05.018> (<https://www.sciencedirect.com/science/article/pii/S0893608006000815>)

(13) Rossi, Fabrice, et al. "Usage Guided Clustering of Web Pages with the Median Self Organizing Map." *ESANN*. 2005.

(14) Oja, Merja, et al. "Clustering of human endogenous retrovirus sequences with median self-organizing map." *Proc. WSOM*. Vol. 3. 2003.

(15) Madalina Olteanu, Nathalie Villa-Vialaneix, On-line relational and multiple relational SOM, *Neurocomputing*, Volume 147, 2015, Pages 15-30, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2013.11.047>.

(16) Olteanu, M., Villa-Vialaneix, N., Cottrell, M. (2013). On-Line Relational SOM for Dissimilarity Data. In: Estévez, P., Principe, J., Zegers, P. (eds) *Advances in Self-Organizing Maps. Advances in Intelligent Systems and Computing*, vol 198. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35230-0_2

(17) M. Khalilia and M. Popescu, "Fuzzy relational self-organizing maps," 2012 IEEE International Conference on Fuzzy Systems, Brisbane, QLD, Australia, 2012, pp. 1-6, doi: 10.1109/FUZZ-IEEE.2012.6250833.

keywords: {Neurons; Vectors; Clustering algorithms; Prototypes; Training data; Data visualization; Lattices; Fuzzy relational self-organizing maps; self organizing maps; relational data; clustering; fuzzy sets}

(18) R.D. Pascual-Marqui, A.D. Pascual-Montano, K. Kochi, J.M. Carazo, Smoothly distributed fuzzy c-means: a new self-organizing map, *Pattern Recognition*, Volume 34, Issue 12, 2001, Pages 2395-2402, ISSN 0031-3203, [https://doi.org/10.1016/S0031-3203\(00\)00167-9](https://doi.org/10.1016/S0031-3203(00)00167-9) (<https://www.sciencedirect.com/science/article/pii/S0031320300001679>)

(19) Sheng-Chai Chi, Ren-Jien Kuo and Po-Wen Teng, "A fuzzy self-organizing map neural network for market segmentation of credit card," *Smc 2000 conference proceedings. 2000 IEEE international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions'* (cat. no.0, Nashville, TN, USA, 2000, pp. 3617-3622 vol.5, doi: 10.1109/ICSMC.2000.886571.

keywords: {Fuzzy neural networks; Neural networks; Fuzzy set theory; Unsupervised learning; Artificial neural networks; Uncertainty; Humans; Clustering methods; Backpropagation; Intelligent systems},

- (20) J. C. Bezdek, "Fuzzy mathematics in pattern classification," 1973.
<https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar> HYPERLINK
["https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y"](https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y) & HYPERLINK
["https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y"](https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y) cbl=18750 HYPERLINK
["https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y"](https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y) & HYPERLINK
["https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y"](https://www.proquest.com/openview/24b294ff26a04fe9c5ad83f61fece5c0/1?pq-origsite=gscholar&cbl=18750&diss=y) diss=y
- (21) J. C. Bezdek, Pattern recognition with fuzzy objective function algorithms: Kluwer Academic Publishers, 1981.
- (22) Romain Boulet, Bertrand Jouve, Fabrice Rossi, Nathalie Villa, Batch kernel SOM and related Laplacian methods for social network analysis, *Neurocomputing*, Volume 71, Issues 7–9, 2008, Pages 1257-1273, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2007.12.026>.
- (23) Nathalie Villa-Vialaneix, Fabrice Rossi. A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph. WSOM'07 - Workshop on Self-Organizing Maps, Sep 2007, Bielefeld, Germany. pp.WeP-1. fhal-00674189
- (24) Leskovec, J., & Krevl, A. (2014). SNAP Datasets: Stanford Large Network Dataset Collection. Retrieved from <https://snap.stanford.edu/data/egonets-Facebook.html>
- (25) Mariette, J., Olteanu, M., Boelaert, J., Villa-Vialaneix, N. (2014). Bagged Kernel SOM. In: Villmann, T., Schleif, FM., Kaden, M., Lange, M. (eds) *Advances in Self-Organizing Maps and Learning Vector Quantization. Advances in Intelligent Systems and Computing*, vol 295. Springer, Cham. https://doi.org/10.1007/978-3-319-07695-9_4
- (26) "Self Organizing Maps (SOM) in NetLogo." cs.us.es — University of Seville, https://www.cs.us.es/~fsancho/Blog/posts/SOM_NetLogo.md
- (27) Ravikant Mordia, Arvind Kumar Verma, Visual techniques for defects detection in steel products: A comparative study, *Engineering Failure Analysis*, Volume 134, 2022, 106047, ISSN 1350-6307, <https://doi.org/10.1016/j.engfailanal.2022.106047>. (<https://www.sciencedirect.com/science/article/pii/S1350630722000218>)

- (28) The Wizard of Ads Brain Map - Wernicke and Broca ,
<https://www.wizardofads.com.au/brain-map-brocas-area>
- (29) Teuvo Kohonen, Essentials of the self-organizing map, Neural Networks, Volume 37, 2013, Pages 52-65, ISSN 0893-6080,
<https://doi.org/10.1016/j.neunet.2012.09.018>.(<https://www.sciencedirect.com/science/article/pii/S0893608012002596>)
- (30) D. Miljković, "Brief review of self-organizing maps," 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2017, pp. 1061-1066, doi: 10.23919/MIPRO.2017.7973581.
- (31) Διαμαντάρας, Κωνσταντίνος. Τεχνητά νευρωνικά δίκτυα. Αθηνά: Κλειδάριθμος, 2007.
- (32) Τεχνητά Νευρωνικά Δίκτυα - Σ. Π. Αδάμ μαθημα - 5
- (33) Tu, Le. (2019). Improving Feature Map Quality of SOM Based on Adjusting the Neighborhood Function. 10.5772/intechopen.89233.
- (34) "SOM algorithm implementation." SOM Toolbox Documentation, CIS,
<http://www.cis.hut.fi/somtoolbox/documentation/somalg.shtml>
- (35) Mohd Zin, Zalian & Khalid, Maather & Mesbahi, Ehsan & Yusof, Rubiyah. (2012). Data Clustering and Topology Preservation Using 3D Visualization of Self Organizing Maps. Lecture Notes in Engineering and Computer Science. 2198.
- (36) V. J. A.S. Lobo, "Application of Self-Organizing Maps to the Maritime Environment ", Proc. IF 2009, 20 May 2009, St. Petersburg, Russia, pp. 19–36.
https://link.springer.com/chapter/10.1007/978-3-642-00304-2_2
- (37) Kohonen, T., MATLAB Implementations and Applications of the Self-Organizing Map Unigrafia Oy, Helsinki, Finland, 2014.
- (38) S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed., Prentice Hall PTR Upper Saddle River, NJ, USA, 1998.

- (39) Introduction To Neural Networks: Revision Lectures: © John A. Bullinaria, 2004,
<https://www.scribd.com/document/362832608/ANN-Introduction>
- (40) Iris Flower Classification, [Peace Ikeoluwa Adegbite](#) , 2020 ,
<https://peaceadegbite1.medium.com/iris-flower-classification-60790e9718a1>
- (41) What is a Distance Matrix, Tim Bock, <https://www.displayr.com/what-is-a-distance-matrix/>
- (42) R. Grantham ,Amino Acid Difference Formula to Help Explain Protein Evolution.Science185,862-864(1974).DOI:10.1126/science.185.4154.862
- (43)<https://www.jstor.org/stable/1739007?seq=2>
- (44)<https://www.geeksforgeeks.org/machine-learning/principal-coordinates-analysis-pcoa-a-comprehensive-guide/>
- (45)<https://www.mathworks.com/help/stats/introduction-to-classical-multidimensional-scaling.html>
- (46) <http://www.cis.hut.fi/somtoolbox/>
- (47) Dunham, M. H. (2006). Data mining: Εισαγωγικά και προηγμένα θέματα εξόρυξης γνώσης από δεδομένα (Β. Βερύκιος & Γ. Θεοδωρίδης, Επιμ.). Εκδόσεις Νέων Τεχνολογιών.
- (48) El Naqa, I., Murphy, M.J. (2015). What Is Machine Learning?. In: El Naqa, I., Li, R., Murphy, M. (eds) Machine Learning in Radiation Oncology. Springer, Cham.
https://doi.org/10.1007/978-3-319-18305-3_1
- (49) Alexander L. Fradkov, Early History of Machine Learning, IFAC-PapersOnLine ,Volume53 ,Issue 2, 2020,Pages 1385-1390, ISSN 2405-8963 ,
<https://doi.org/10.1016/j.ifacol.2020.12.1888>.(<https://www.sciencedirect.com/science/article/pii/S2405896320325027>)
- (50) A Brief History of Deep Learning, Keith D. Foote, 2022,
<https://www.dataversity.net/articles/brief-history-deep-learning/>
- (51) Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193-202.

Commented [GU4]: αναφορές στο τέλος

- (52) Greene, D., Cunningham, P., Mayer, R. (2008). Unsupervised Learning and Clustering. In: Cord, M., Cunningham, P. (eds) Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_3
- (53) Kodinariya, Trupti & Makwana, Prashant. (2013). Review on Determining of Cluster in K-means Clustering. International Journal of Advance Research in Computer Science and Management Studies. 1. 90-95.
- (54) Bradley, P. S., Fayyad, U., & Reina, C. (1998). Scaling EM (expectation-maximization) clustering to large databases. Microsoft Research, 0-25.
- (55) D. Deng, "DBSCAN Clustering Algorithm Based on Density," 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, 2020, pp. 949-953, doi: 10.1109/IFEEA51475.2020.00199.
- keywords: {Machine learning algorithms;Clustering algorithms;Machine learning;Big Data;Prediction algorithms;Data mining;Unsupervised learning;DBSCAN Algorithm;Density Clustering;Machine Learning;Algorithm Research},
- (56) Van Der Maaten, L., Postma, E. O., & Van Den Herik, H. J. (2009). Dimensionality reduction: A comparative review. Journal of machine learning research, 10(66-71), 13.
- (57) Abdi, H. and Williams, L.J. (2010), Principal component analysis. WIREs Comp Stat, 2: 433-459. <https://doi.org/10.1002/wics.101>

appendix

appendix A κώδικας μετατροπής iris

Παρακάτω δείχνεται (A1) το script που χρησιμοποιήθηκε.

```
import pandas as pd
from sklearn.datasets import load_iris
from scipy.spatial import distance_matrix

#load iris dataset on x
iris = load_iris()
X = pd.DataFrame(iris.data, columns=iris.feature_names)

#transform to distances
dist_matrix = pd.DataFrame(distance_matrix(X.values, X.values))

#save to csv
dist_matrix.to_csv("iris_euclidean_distances.csv", index=False)
```

(Εικόνα A1) κώδικας που χρησιμοποιήθηκε για την μετατροπή του σετ δεδομένων iris σε πίνακα ανομοιότητας.

Στο συγκεκριμένο script χρησιμοποιούμε την βιβλιοθήκη pandas σαν ένα dataframe για τα δεδομένα μας, την βιβλιοθήκη sklearn για το σετ δεδομένων του iris και την scipy για την μετατροπή του iris dataset σε ένα πίνακα διαστάσεων. Όπου μετά το εξάγουμε σε ένα csv για μια τελευταία επεξεργασία.

Commented [GU5]: να αφαιρεθούν όλοι οι κώδικες μέσα από το κείμενο να μπουν αν θες Appendix

appendix B Κλάσεις που χρησιμοποιήθηκαν και τροποποιήσεις

Σε αυτό το κεφάλαιο θα μιλήσουμε περιληπτικά για της κλάσης που θα χρησιμοποιηθούν καθώς όμως και το βασικό script για τα παραδείγματα και των δυο αλγόριθμών. Επίσης θα αναφερθούν περαιτέρω τροποποιήσεις εάν έχουν γίνει σε κάποια κλάση για να βοηθηθεί η κατανόηση των παραδειγμάτων.

Ξεκινώντας με τον κανονικό SOM η κλάσης που χρησιμοποιούνται δεν έχουν πειραχτεί έως και καθόλου και κυρίως χρησιμοποιούνται τεχνικές για να δείξουμε καλύτερα τα δεδομένα και τα αποτελέσματά τους. Αρχικά τα δεδομένα που δέχεται ο κανονικός som πρέπει να βρίσκονται σε μορφή SOM_PAK . Αυτό σημαίνει ότι τα δεδομένα μας θα πρέπει να χωρίζονται από ένα κενό το ένα με το άλλο χωρίς κόμματα. Η μόνη εξαίρεση σε αυτό είναι στις πρώτες δυο γραμμές το αρχείου στις οποίες μπορούμε να ορίσουμε της διαστάσεις και το όνομα των χαρακτηριστικών των δεδομένων για την εμφάνιση τους. Η διαστάσεις δεν είναι κάτι υποχρεωτικό μιας και το πρόγραμμα μπορεί να αναγνωρίσει αυτόματα της διαστάσεις και επίσης σε περίπτωση ελλιπών δεδομένων μπορεί να σηματοδοτηθεί χρησιμοποιώντας το γράμμα x. Παρακάτω θα αναφερθούν περιληπτικά η κλάσης που χρησιμοποιήθηκαν και τυχών τροποποιήσεις

som_read_data: Η συγκεκριμένη κλάση χρησιμοποιείται για να διαβάσουμε τα δεδομένα μας από ένα .data αρχείο το οποίο όπως προαναφέρθηκε θα πρέπει να βρίσκεται σε SOM_PAK μορφή. Η μόνη διαφορά του όπως προαναφέρθηκε είναι ότι μπορεί να λίβεται της διαστάσεις των δεδομένων μιας και μπορεί να της υπολογίσει.

som_normalize: Η κλάση αυτή χρησιμοποιείται για να γίνει κανονικοποίηση των δεδομένων μας με σκοπό να γίνει ομαλή η εκπαίδευση των δεδομένων στον αλγόριθμο.

som_randinit: Σε αυτήν την κλάση δίνονται τυχαίες τιμές για τους νευρώνες του αλγορίθμου . Μια τεχνική που προαναφέρθηκε στην εξήγηση του αλγορίθμου με σκοπό να δοθούν αρχικές τιμές στου νευρώνες και είναι η πιο συνηθής τεχνική αρχικοποίησης των τιμών των νευρώνων. Επίσης σε περίπτωση που δεν έχει γίνει είδη δημιουργεί και ένα som struct το οποίο είναι το αντικείμενο του χάρτη som που περιέχει την δομή του χάρτη, της πληροφορίες των δεδομένων , των νευρώνων και την τοπολογία του. Στην ίδια κλάση επίσης μπορούν να δοθούν αρχικές τιμές για το σχήμα του χάρτη, του πλέγματος και της διαστάσεις του χάρτη.

som_set: Η συγκεκριμένη κλάση βρίσκεται μέσα στην som_randinit και τρέχει σε περίπτωση που δεν υπάρχει κάποιο som struct για να το δημιουργήσει αλλά και να βάλει της τιμές σε αυτό.

som_train_struct: Η συγκεκριμένη κλάση χρησιμοποιείται για να δώσουμε οδηγίες στο πως θέλουμε να γίνει η εκπαίδευση σε ένα som struct. Μπορεί να συνδυαστεί με προηγούμενα train struct για πιο πολύπλοκους τρόπους εκπαίδευσης επίσης.

som_batchtrain: Μια κλάση η οποία μας επιτρέπει να εκπαιδεύσουμε τον αλγόριθμο σε batch setting. Μας επιτρέπει να ελέγξουμε πόσο θα εκπαιδευτεί ο αλγόριθμος, αλλά και της ρυθμίσεις πάνω του όπως γειτονιές, σχήμα χάρτη και άλλες ρυθμίσεις

som_show: Η συγκεκριμένη κλάση χρησιμοποιείται για να απεικονίσουμε τα δεδομένα μας . Χρησιμοποιούνται για να δείξουμε τα πλαίσια των δεδομένων στο umatrix.

Παρακάτω (εικόνα B1) βλέπουμε το script που χρησιμοποιούμε για τον κανονικό som ο βασικός κώδικας είναι ο ίδιος με της μονές μεταβλητές που αλλάζουν μεταξύ παραδείγματα να είναι το μέγεθος του χάρτη και η διαστάσεις που δίνουμε σε οποία συνάρτηση την ζήτηση.

```
data='irisEuclidianData.data';%data for som training

msize = [5 5];%map size
sDiris = som_read_data(data);
sDiris = som_normalize(sDiris, 'range');
sMap = som_randinit(sDiris,'msize', msize);
```

(εικόνα B1) Κομμάτι κώδικα αρχικοποιήσεις κώδικα κλασσικού som .

Στο παραπάνω κομμάτι κώδικα βλέπουμε αρχικά την μεταβλητή data την οποία θα χρησιμοποιήσουμε στις μεταβλητές που εξηγήθηκαν προηγούμενος για να διαβάσουμε τα δεδομένα και έπειτα θα τα κανονικοποιήσουμε για να τα χρησιμοποιήσουμε στην εκπαίδευση το som read data γενικά επιστρέφει ένα struct από τιμές οι οποίες δημιουργούν το τελικό som struct που περιέχει όλα τα δεδομένα του χάρτη. Ο χάρτης όπως προαναφέρθηκε δημιουργείται στην συνάρτηση som_randinit η οποία δέχεται 2 τιμές το sDiris οπου είναι τα δεδομένα που χρησιμοποιούμε και το msize οπου είναι το μέγεθος των νευρώνων που θα εκπαιδευτούν στον χάρτη .

```
lengthdata= length(sDiris.data);%data length for train struct
sT1 = som_train_struct('rough','dlen',lengthdata,sMap); % rough training
sT2 = som_train_struct('finetune','previous',sT1); % finetuning
for i=1:100
    if i < 60%first we train rough then we move to fine
        disp('rough')
        sMap = som_batchtrain(sMap,sDiris,'tracking',1,'trainlen',5,...
            'train',sT1);
    else
        disp('fine')
        sMap = som_batchtrain(sMap,sDiris,'tracking',1,'trainlen',5,...
            'train',sT2);
    end
end
title('Sequential training after 500 steps')
```

(Εικόνα B2) Κομμάτι κώδικα εκπαίδευσης του κλασσικού som.

Έπειτα παραπάνω (εικόνα B2)βλέπουμε που προχωράει ο αλγόριθμος με το να ξεκινάμε την εκπαίδευση του. Χρησιμοποιούμε την κλάση som_train_struct για να ορίσουμε πως θα γίνει η εκπαίδευση του αλγορίθμου. Συγκεκριμένα φτιάχνουμε δυο struct το πρώτο το ρυθμίζουμε σε rough setting όπου θα κάνει της μεγαλύτερες κινήσεις στην εκπαίδευση και το δεύτερο σε fine setting όπου θα κάνει πιο μικρές κινήσεις στον χάρτη για να είναι πιο ακριβής. Έπειτα δίνουμε μια λούπα 100 βήματα όπου για το 60% της εκπαίδευσης θα γίνεται ένα rough training σε 5 βήματα την φορά και στο υπόλοιπο fine training . Σκοπός είναι να γίνει για το πρώτο μεγαλύτερο κομμάτι η εφαρμογή των νευρώνων στον χάρτη και μετά να τον ακριβαίνουμε στα δεδομένα όσο μπορούμε. Συνολικά θα έχουμε μια εκπαίδευση 500 βημάτων την οποία μπορούμε να πειράξουμε εάν θέλουμε στην αρέσκεια μας.

```
som_show(sMap,'umat','all')%umatrix trained
pause
```

(εικόνα B3)

Τέλος(εικόνα B3) χρησιμοποιώντας τις συναρτήσεις `som_show` και `som_show_add` δείχνουμε το `umatrix` του χάρτη προς τα δεδομένα τα οποία χρησιμοποιούμε. Παρακάτω με παρόμοιο τρόπο θα εξηγηθεί το `script` και η συναρτήσεις για τον RSOM.

RSOM

Η συναρτήσεις που χρησιμοποιούμε στον `rsom` λειτουργούν με παρόμοιο τρόπο μια όμως από της κυρίες διάφορες τους είναι στα δεδομένα τα οποία δέχονται. Μιας και τα δεδομένα τώρα πρέπει να είναι πίνακας της `matlab` και να έχει προ φορτωθεί . Συγκεκριμένα θα παρατηρηθεί στο `script` το οποίο θα δείχθει ότι τα δεδομένα μας θα έχουν το όνομα `Dissim` το οποίο είναι το όνομα του χάρτη `mat` το οποίο φορτώνουμε με τα δεδομένα. Παρακάτω όπως και με τον κανονικό `som` θα εξηγηθούν η κλάσης που χρησιμοποιήθηκαν καθώς και θα επεξηγηθεί το βασικό `script` που θα χρησιμοποιήσουμε.

Rsom_init: Αυτή η συνάρτηση αρχικοποιεί τον χάρτη μας με βάση της ομοιοτιμές των δεδομένων μας. Ουσιαστικά η αρχικοποίηση γίνεται γραμμικά και επίσης δέχεται της αρχικές ρυθμίσεις του χάρτη για να δημιουργήσει ένα παρόμοιο με τον SOM struct .Η τιμές που δέχεται είναι τα δεδομένα και το μέγεθος του χάρτη υποχρεωτικά μαζί με μη υποχρεωτικές τιμές όπως για παράδειγμα το σχήμα του χάρτη , αν θα είναι τετραγωνικός η εξαγωνικός αλλά και το όνομα του αν επιλέξουμε ένα. Είναι σημαντικό να αναφερθεί ότι παρόλο που είναι όμοια τα structs πρέπει να δημιουργηθεί ένα διαφορετικό εάν κάποιος θέλει να χρησιμοποιήσει της λειτουργίες του αλλού και ακόμα και τότε η λειτουργίες που μπορούν να χρησιμοποιηθούν είναι περιορισμένες.

rsom_batchtrain: Όπως συνιστά και το όνομα της αυτή η συνάρτηση χρησιμοποιείται για να εκπαιδεύσουμε τον αλγόριθμο σε ρύθμιση batch . Επίσης μπορεί να οριστεί για πόσο θα εκπαιδευτεί ο αλγόριθμος φυσικά και δέχεται σαν είσοδο τον είδη αρχικοποιημένο χάρτη, τα δεδομένα και σαν δευτερεύων ρυθμίσεις τον αριθμό των επαναλήψεων που θα εκπαιδευτεί αλλά για το λαμδα.

rsom_show: Όμοιος με τον κανονικό som χρησιμοποιείται για να δείξουμε το umatrix του χάρτη.

mysomshow: Μια παραλλαγή το rsomshow το οποίο καλή να δείξει της αποστάσεις στο umatrix του χάρτη σε μέσης αποστάσεις το οποίο σημαίνει ότι θα μας δείξει της αποστάσεις στον αριθμητικό μέσω σε σχέση με την κανονική ρύθμιση του rsom_show όπου μας δείχνει την διάμεσο .Γενικότερα τα όρια μεταξύ των δεδομένων φαίνονται πιο καθαρά.

Όπως και με τον κανονικό SOM παρακάτω εξηγείτε το βασικό script του αλγόριθμου rsom που θα χρησιμοποιηθεί .Όπου όπως και με τον κλασσικό SOM τα μόνα πράγματα που αλλάζουν μεταξύ παραδειγμάτων είναι η μεταβλητές για τα δεδομένα , το μέγεθος του χάρτη αλλά και η διαστάσεις όπου και αν χρειαστεί.

```
load matlabIrisDissim.mat;

data='irisEuclidianData.data';
sDiris = som_read_data(data);
sDiris = som_normalize(sDiris);
```

(εικόνα B4)Κομμάτι κώδικα αρχικοποιήσεις των δεδομένων στον rsom.

Αρχικά (εικόνα B4) όπως και με τον κανονικό som ξεκινάμε με τα δεδομένα μας. Όπως προαναφέρθηκε τα δεδομένα μας πρέπει να βρίσκονται σε μορφή numeric matrix.mat. Κάτι που επίσης πρέπει να αναφερθεί είναι η χρήση των συναρτήσεων για τα δεδομένα από τον κανονικό som τα οποία θα χρησιμοποιηθούν από τον άλλες συναρτήσεις του κανονικού som.

```
sMap = rsom_lininit(Dissim, [5 5], 'lattice', 'hexa'); % train the RSOM
sMap = rsom_batchtrain(sMap, Dissim, 'epochs', 100);
```

(εικόνα B5)(κομμάτι κώδικα εκπαίδευσης στον rsom)

Έπειτα (εικόνα B5) προχωράμε στην αρχικοποίηση και εκπαίδευση του αλγόριθμου που όπως αναφέραμε παραπάνω χρησιμοποιεί μια γραμμική αρχικοποίηση των δεδομένων. Στην εικόνα ουσιαστικά χρησιμοποιούμε τα δεδομένα Dissim για την αρχικοποίηση σε μέγεθος χάρτη νευρώνων 5x5 και σε πλέγμα το οποίο είναι εξαγωνικό. Έπειτα εκπαιδεύουμε τον αλγόριθμο σε ρύθμιση batch για 100 επαναλήψεις.

```
figure;
rsom_show(sMap, Dissim);

pause % Strike any key to continues...
clc;

myrsomshow(sMap, Dissim);
```

(εικόνα B6)

Αφού τελειώσει την εκπαίδευση ο χάρτης φυσικά προχωράμε στο να δείξουμε (εικόνα B6) τα δεδομένα, έχουμε και της δυο κλάσης για το umatrix έτσι ώστε να δείξουμε την διαφορά στην ποιότητα μεταξύ τους.