Αναφορά 2ης Εργαστηριακής Άσκησης Νευρωνικά Δίκτυα

Κωνσταντίνος Καλλάς 03112057
Τζίνης Ευθύμιος 03112007

15 Ιανουαρίου 2017

Εισαγωγή

Σε αυτή την εργαστηριακή άσκηση υλοποιήσαμε έναν αυτο-οργανώμενο χάρτη ($Self\ Organizing\ Map\ -SOM$. Πειραματιστήκαμε σε διάφορα προβλήματα όπως $classification$, $clustering$, $TSP\ solving$ και είδαμε την ικανότητα γενίκευσης και περιγραφής δεδομένων, των χαρτών αυτών. Παρακάτω περιγράφουμε τα σημεία τα οποία μας κέντρισαν το ενδιαφέρον καθόλη την διάρκεια πραγμάτωσης και πειραματισμού και παραθέτουμε αντίστοιχα διαγράμματα που βοηθούν στην κατανόηση και των αποτελεσμάτων μας. Ο κώδικας παρέχεται στον αντίστοιχο φάκελο και είναι αρκετά επεξηγηματικός μαζί με τα απαραίτητα σχόλια. Παρέχουμε επίσης και ένα φάκελο με διάφορα διαγράμματα που εξήχθησαν κατά τη διάρκεια της άσκησης.

1

Σε αυτό το κομμάτι της άσκησης υλοποιήσαμε τις ζητούμενες συναρτήσεις ώστε να ακολουθούν τις δοθείσες περιγραφές. Παρόλα αυτά θέλουμε να επισημάνουμε κάποια σημεία:

- Στην συνάρτηση $somActivation(pattern, neighborDist)$ δεν θα ήταν το καλύτερο να έχουμε μόνο δύο διακριτές τιμές 0.5 και 0 για τους γειτονικούς νευρώνες. Το πιο σωστό θα ήταν να παρέχουμε μία συνάρτηση η οποία θα είναι φθίνουσα ως προς την απόσταση των νευρώνων από τον νευρώνα νικητή. Με αυτόν τον τρόπο θα μπορούσαμε να δώσουμε μιά πιο δίκαιη μετρική για την ανανέωση των βαρών των γειτωνικών νευρώνων κατά την διάρκεια της **Ανταμοιβής**.
- Στην συνάρτηση $somTrain(patterns)$ χρησιμοποιήσαμε την συνάρτηση e^{-x} για την εκθετική μείωση της τιμής κατωφλίου για την απόσταση που εξαρτάται αν κάποιος νευρώνως είναι γείτονας ή όχι με τον νευρώνα νικητή. Πιστεύουμε ότι ίσως θα έπρεπε να γίνει κάποια ανάλυση για το ποιά θα ήταν η καταληλότερη συνάρτηση μείωσης αυτής της τιμής ώστε να μπορούμε να φτιάξουμε κάποιο δυναμικά μεταβαλλόμενο σύστημα που να κάνει $adapt$ στα δεδομένα που θέλει να αναπαραστήσει. Διαισθητικά, θεωρούμε ότι ο ρυθμός

μείωσης αυτής της απόστασης πρέπει να έχει κάποια σχέση με την διασπορά των δεδομένων μας στον Ευκλείδιο χώρο. (Όταν τα δεδομένα λεχουν μικρή διασπορά σε κάποια διάσταση τότε αυτό σημαίνει ότι ένας χάρτης θα συγκλίνει πολύ γρήγορα στο κεντροειδές αυτής της κατανομής και άρα δεν θα ήταν καλύτερο η εκθετική μείωση να είναι ακόμη πιο γρήγορη σε σχέση με τον αριθμό των επαναλήψεων, και αντιστρόφως για όταν τα δεδομένα μας έχουν μεγάλη διασπορά).

Δεν υλοποιήσαμε τις παραπάνω ιδέες απλά την αναφέρουμε για πληρότητα και για πιθανή δουλειά που θα θέλαμε να συνεχίσουμε να εργαζόμαστε.

2

Α΄

Σε αυτό το κομμάτι της άσκησης είδαμε το πόσο καλά μπορεί να αναπαραστήσει ένας αυτο-οργανούμενος χάρτης τα δισδιάστατα δεδομένα που του βάζουμε ως είσοδο. Επειδή τα δεδομένα μας βρίσκονται στον δισδιάστατο χώρο είναι πολύ εύκολο να τα παραστήσουμε στο ίδιο επίπεδο μαζί με τις τελικές θέσεις των νευρώνων. Πειραματιστήκαμε με όλες τις δυνατές τοπολογικές διατάξεις του πλέγματος των νευρώνων καθώς επίσης και με αρκετές μετρικές αποστάσεων του Ευκλείδιου χώρου. Πιο κάτω παραθέτουμε διάφορες παραμετροποιήσεις και επιλέγουμε κάποιο άριθμό νευρώνων. 1, 2

Παρατηρούμε ότι για τα δεδομένα 8 τα μονοδιάστατα πλέγματα δεν αρκούν για την αναπαράσταση των δεδομένων. Αντίθετα για τα δεδομένα του ερωτηματικού τα μονοδιάστατα πλέγματα απεικονίζουν αρκετά καλά τα δεδομένα. Όσον αφορά τις διάφορες παραμετροποιήσεις τοπολογίας και αποστάσεων παρατηρούμε ότι η εξαγωνική τοπολογία τείνει να επιστρέφει καλύτερα αποτελέσματα.

Εκτός από την απλή εποπτεία των αποτελεσμάτων των νευρώνων είναι χρήσιμο να ορίσουμε μία πιο μαθηματική έκφραση του αν ένα δίκτυο περιγράφει αρκετά καλά τα δεδομένα μας. Όπως περιγράφεται και στο [1] μια πολύ καλή μετρική που δείχνει ακριβώς αυτό, είναι η εξης:

$$QE = \frac{1}{P}\sum_{j=1}^{P} ||p_j - m_{c(j)}||^2$$

Όπου το x_j αναπαριστά το j-οστό πρότυπο από τα P συνολικά πρότυπα και τα $m_{c(j)}$ είναι τα αντίστοιχιζόμενα καλύτερα διανύσματα από όλους τους νευρώνες που περιγράφουν καλύτερα το πρότυπο j. όσο μικρότερη τιμή έχει η συγκεκριμένη μετρική τόσο καλύτερα αναπαρίσταται το σύνολο δεδομένων από το δίκτυοΚατά αυτόν τον τρόπο υπολογίζουμε αυτή την ποσότητα για όλες τις παραμετροποιήσεις μας για να έχουμε μιά καλύτερη και πιο μαθηματική αντίληψη για το πόσο καλά ο χάρτης μας περιγράφει την τοπολογική διάταξη αλλά και την πιθανοτική κατανομή των δεδομένων εισόδου. Αυτή η μετρική είναι πάρα πολύ χρήσιμη κυρίως σε προβλήματα με διάσταση μεγαλύτερη του 3 αφού για τέτοιου είδου δεδομένα χάνουμε την δυνατότητα καθολικής εποπτείας και μια τέτοια μετρική είναι χρήσιμη για την επιλογή των παραμέτρων του χάρτη μας.

Παραθέτουμε τα αποτελέσματα αυτής της μετρικής για όλες τις παραμετροποιήσεις που εξετάσαμε.

QE για τα δεδομένα 8						
	Αριθμός Νευρώνων					
	5-1	10-1	20-1	5-5	10-5	10-10
QE	1.51	1.11	0.673	0.631	0.349	0.246

QE για τα δεδομένα 8					
	Είδος Αποστάσεων				
Τοπολογία	*boxdist*	*dist*	*linkdist*	*mandist*	$ring_distance$
gridtop	0.678	0.563	0.576	0.576	0.576
hextop	0.59	0.637	0.65	0.527	0.527
randtop	0.725	0.618	0.644	0.55	0.55
hexagonaltopology	0.59	0.631	0.65	0.527	0.527

QE για τα δεδομένα ερωτηματικό						
	Αριθμός Νευρώνων					
	5-1	10-1	20-1	5-5	10-5	10-10
QE	0.572	0.298	0.195	0.241	0.146	0.0988

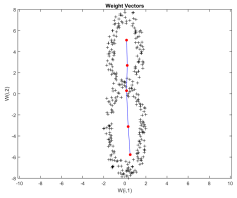
QE για τα δεδομένα ερωτηματικό					
	Είδος Αποστάσεων				
Τοπολογία	*boxdist*	*dist*	*linkdist*	*mandist*	$ring_distance$
gridtop	0.25	0.202	0.205	0.205	0.205
hextop	0.208	0.241	0.217	0.179	0.2
randtop	0.221	0.212	0.218	0.192	0.192
hexagonaltopology	0.208	0.241	0.217	0.179	0.2

Β′

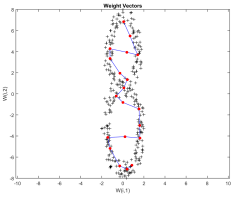
Σε αυτό το κομμάτι της άσκησης χρησιμοποιούμε τον χάρτη ΣΟΜ για την επίλυση του $NP-Hard$ προβλήματος $TravellingSalesmanProblem(TSP)$, προφανώς η λύση είναι ευρεστική και δεν εγγυάται προφανώς την πλήρη επίλυσή του (λύση με μικρότερο κόστος συνολικά). Παρατηρούμε ότι προκειμένου να έχουμε μία καλή ευρεστική λύση του προβλήματος, αρκεί να θέσουμε τον αριθμό των νευρώνων περίπου ίσο με το διπλάσιο αριθμό των πόλεων.

Πιο συγκεκριμένα για να γίνει πιο κατανοητό ότι η λύση του προβλήματος είναι ευρεστική και εξαρτάται τόσο από τις αρχικές τιμές των παραμέτρων του πλέγματος μπορούμε να δούμε στις εικόνες 3 (πάνω αριστερά) ότι για διαφορετικό αριθμό νευρώνων ο τρόπος που επισκεπτόμαστε τις πόλεις αλλάζει. Αυτό είναι το πιο χαρακτηριστικό παράδειγμα του ότι η λύση που μας δίνει το SOM δεν μπορεί να ανταποκριθεί σε κανένα κριτήριο την βέλτιστη λύση. Παρακάτω φαίνεται και ένας πίνακας με τα QE αποτελέσματα του χάρτη για τις διάφορες παραμετροποιήσεις

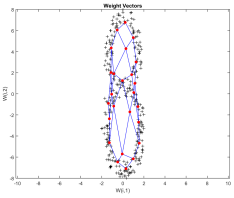
QE για το TSP						
	Αριθμός Νευρώνων					
	5	10	20	50	100	200
QE	0.206	0.136	0.0744	0.00804	$4.99 \cdot 10^{-16}$	$4.99 \cdot 10^{-16}$



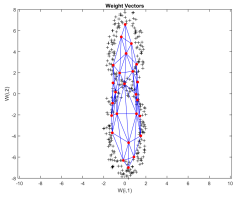
(α′) Ένα SOM με 5 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



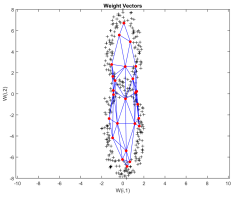
(β′) Ένα SOM με 20 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



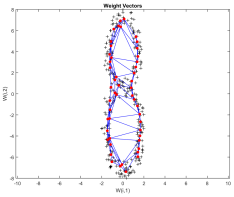
(γ′) Ένα SOM με 5-5 νευρώνες, τετραγωνικό πλέγμα και ευκλ. αποστάσεις



(δ′) Ένα SOM με 5-5 νευρώνες, τετραγωνικό πλέγμα και αποστάσεις $boxdist$

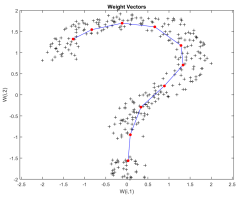


(ε′) Ένα SOM με 5-5 νευρώνες, τυχαίο πλέγμα και αποστάσεις $boxdist$

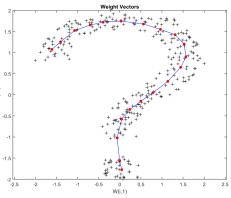


(ϛ′) Ένα SOM με 10-5 νευρ., εξαγ. πλέγμα και ευκλ. αποστάσεις $boxdist$

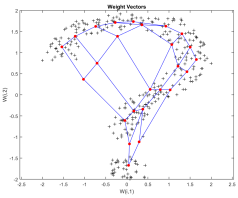
Σχήμα 1: Τα βάρη του αυτοοργανώμενου χάρτη για τα δεδομένα 8



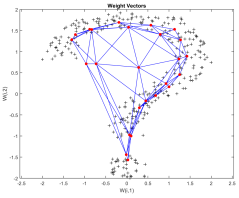
(α′) Ένα SOM με 10 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



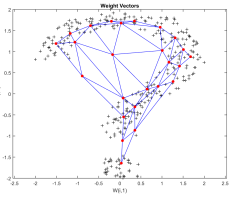
(β′) Ένα SOM με 20 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



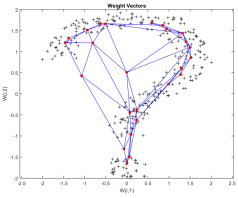
(γ′) Ένα SOM με 5-5 νευρώνες, τετραγωνικό πλέγμα και ευκλ. αποστάσεις



(δ′) Ένα SOM με 5-5 νευρώνες, τετραγωνικό πλέγμα και αποστάσεις $boxdist$



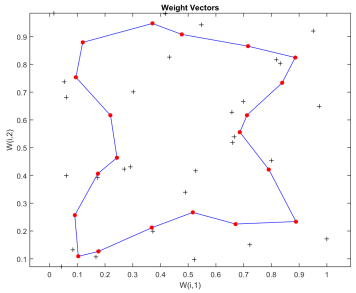
(ε′) Ένα SOM με 5-5 νευρώνες, εξαγωνικό πλέγμα και αποστάσεις $boxdist$



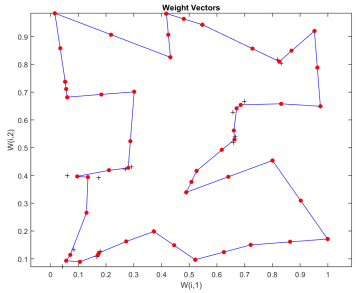
(ϛ′) Ένα SOM με 5-5 νευρ., εξαγ. πλέγμα και ευκλ. αποστάσεις $boxdist$

Σχήμα 2: Τα βάρη του αυτοοργανώμενου χάρτη για τα δεδομένα Ερωτημαικό

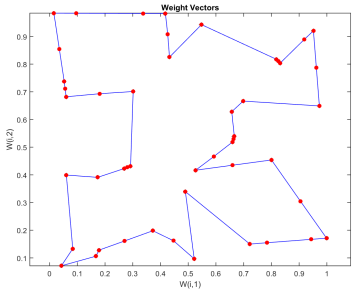
Φαίνεται λοιπόν ότι με 50-100 νευρώνες το δίκτυο αναπαριστά εντελώς τις 30 πόλεις. Παρόλαυτα για 100 και 200 νευρώνες το δίκτυο δεν δίνει το ίδιο αποτέλεσμα μονοπατιού λόγω τυχαιότητας όπως εξηγήσαμε και παραπάνω.



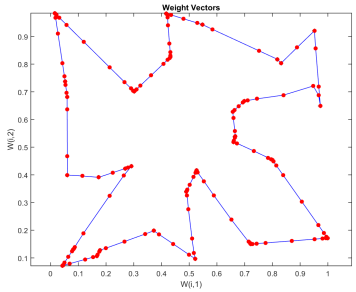
(α′) Ένα SOM με 20 νευρώνες στο TSP



(β′) Ένα SOM με 50 νευρώνες στο TSP



(γ′) Ένα SOM με 100 νευρώνες στο TSP



(δ′) Ένα SOM με 200 νευρώνες στο TSP

Σχήμα 3: Τα βάρη του αυτοοργανώμενου χάρτη για το πρόβλημα του Πλανόδιου πωλητή

Γ′

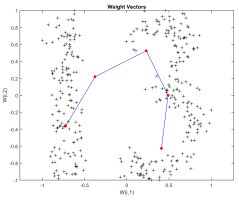
Αυτό το πρόβλημα είναι ένα χαρακτηριστικό πρόβλημα $classification$, συνεπώς για την εκπαίδευση του δικτύου μας δεν χρησιμοποιήσαμε καθόλου τα δεδομένα της τρίτης στήλης (κλασση 0 ή 1). Αν δούμε τα δεδομένα μας $GroupData.m$ μπορούμε να δούμε ότι οι 2 πρώτες στήλες αναπαριστούν τον αριθμό 13 στο δισδιάστατο επίπεδο. Αφου εκπαιδεύσουμε το δίκτυό μας και οι νευρώνες του καταλήξουν κάπου στον δισδιάστατο χώρο, τότε αφού τα δεδομένα εκπαίδευσης ειναι γραμμικά διαχωρίσιμα, στην οριζόντια διάσταση, μπορούμε να θεωρήσουμε ότι θα πάρουμε τις γραμμές $x = -0.35, x = -0.25$ για όρια απόφασης. Αριστερά από το $x = -0.35$ τα δεδομένα ανήκουν στην **κλάσση 0** ενώ δεξιά του $x = -0.25$ ανήκουν στην **κλάσση 1**. Ανάμεσα σε αυτές τις δύο γραμμές θεωρούμε ότι δεν μπορούμε να αποφανθούμε για το ποιόν του νευρώνα καθώς αυτός βρίσκεται κάπου ενδιάμεσα στις δύο κλάσσεις.

Σημείωση: Θα μπορούσαμε σε πιο δύσκολα προβλήματα να βρούμε τις επιφάνειες απόφασης με κάποιο ποιο εκλεπτυσμένο τρόπο όπως για παράδειγμα κάνει και ο SVM. Καθώς όμως στα δεδομένα μας είναι πολύ εύκολο εποπτικά να βρούμε τις επιφάνειες αυτές δεν υπήρχε λόγος να περιπλέξουμε την ανάλυσή μας.

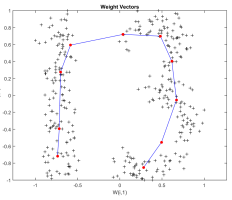
Επιστρέφοντας στον ορισμό του προβλήματος $classification$, τώρα που έχουμε

εκπαιδεύση το μοντέλο μας πάνω στα δεδομένα εισόδου και ξέρουμε σε ποιά κλάση ανήκει πρακτικά ο κάθε νευρώνας και με αυτόν τον τρόπο μπορούμε με κάποια μετρική στον Ευκλείδιο χώρο των *features* να κατατάξουμε το *test* δεδομένο μας στον καταλληλότερο νευρώνα για να αντιπροσωπεύσει αυτό το δεδομένο. Κατα αυτόν τον τρόπο η κλάσση που ανήκει ο νευρώνας αντιπρόσωπος θα ανατεθέι και στο *test* δεδομένο.

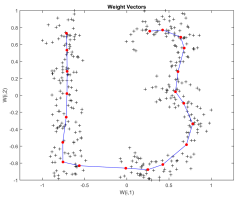
Παραθέτουμε *umatrices* αλλά και δισδιάστατη απεικόνιση των βαρών του δικτύου για διάφορες παραμετροποιήσεις.



(α′) Ένα *SOM* με 5 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



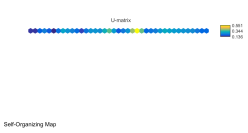
(β′) Ένα *SOM* με 10 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



(γ′) Ένα *SOM* με 20 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις

(δ′) Το *umatrix* ενός *SOM* με 5 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις

(ε′) Το *umatrix* ενός *SOM* με 10 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



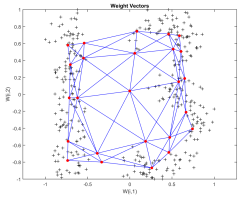
(ϛ′) Το *umatrix* ενός *SOM* με 20 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις

Σχήμα 4: Τα βάρη του αυτοοργανώμενου χάρτη για τα δεδομένα 13

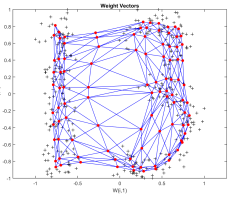
Με βάση όλα τα παραπάνω, δίνουμε απαντήσεις στα ερωτήματα που τίθενται στην εκφώνηση της άσκησης.

- Όπως φαίνεται και στο διάγραμμα 5 το μέγεθος των ομάδων στο *umatrix* είναι ανάλογο και του αριθμού των δεδομένων κάθε ομάδας. Για την ακρίβεια τα δεδομένα του 1 είανι 144 ενώ του 3 206. Η αναλογία 3/1 είναι 1.4306.
- Παρατίθονται πίνακες με τις αναλογίες νευρώνων που ανατίθονται σε κάθε ομάδα. (Αυτές οι αναλογίες υπολογίστηκαν με τον τρόπο που εξηγείται παραπάνω αφού τα δεδομένα είναι γραμμικά διαχωρίσιμα)

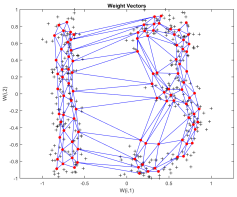
Αναλογία πλήθους δεδομένων σε κάθε ομάδα						
	Αριθμός Νευρώνων					
	10-1	20-1	5-5	10-5	5-10	10-10
Αναλογία	1.5	1.5	1.5	1.33	1.45	1.5



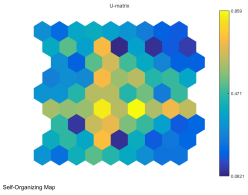
(α′) Ένα *SOM* με 5-5 νευρώνες, τετραγωνικό πλέγμα και *boxdist* αποστάσεις



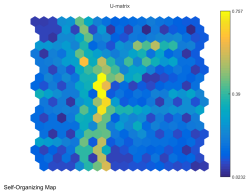
(β′) Ένα *SOM* με 10-10 νευρώνες, τετραγωνικό πλέγμα και *boxdist* αποστάσεις



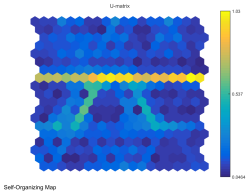
(γ′) Ένα *SOM* με 10-10 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις



(δ′) Ένα *SOM* με 5-5 νευρώνες, τετραγωνικό πλέγμα και *boxdist* αποστάσεις



(ε′) Ένα *SOM* με 10-10 νευρώνες, τετραγωνικό πλέγμα και *boxdist* αποστάσεις



(ϛ′) Ένα *SOM* με 10-10 νευρώνες, εξαγωνικό πλέγμα και ευκλείδιες αποστάσεις

Σχήμα 5: Τα βάρη του αυτοοργανώμενου χάρτη για τα δεδομένα 13

Αναλογία πλήθους δεδομένων σε κάθε ομάδα			
	Είδος Αποστάσεων		
Τοπολογία	*boxdist*	*dist*	*mandist*
gridtop	1.77	1.63	1.58
hexagonaltopology	1.65	1.5	1.63

Παρατηρούμε ότι συνολικά η αναλογία νευρώνων και προτύπων σε κάθε ομάδα είναι παρόμοια άσχετα με την παραμετροποίηση του συστήματος.

- Οι διαφοροποιήσεις στα σύνορα διαχωρισμού όπως φαίνονται στο *umatrix* εξαρτόνται στο πόσο καλά αναπαρίστανται τα δεδομένα από κάθε παραμετροποίηση.

3

3Α′, 3Β′

Σε αυτό το κομμάτι της άσκησης δοκιμάζουμε τις δυνατότητες του δικτύου *SOM* στο πρόβλημα του *clustering*. Όπως θα δούμε και παρακάτω το δίκτυο *SOM* μπορεί να ανακαλύψει κάποια συσχέτιση μεταξύ των δεδομένων εισόδου και να τα χωρήσει σε συστάδες-*clusters*. Προφανώς δεν υπάρχει κάποια μετρική για να μπορέσουμε να δούμε αν πραγματικά έχουμε καταφέρει να κάνουμε μαι καλή συσταδοποίηση ή όχι. Για αυτό τον λόγο, η μόνη αξιολόγηση για διαφορετικές παραμετροποιήσεις βασίζεται μόνο στις παρατηρήσεις μας και στην διαίσθησή μας. Στην περιπτωσή μας τα δεδομένα εισόδου είναι τα έγγραφα-πρότυπα τα οποία έχουν

ένα διακριτό τίτλο και αναπαριστόνται στον χώρο των χαρακτηριστικών με ένα διάνυσμα από την συχνότητα εμφάνισης της κάθε λέξης-διάστασης.

Έχει αποδειχτεί πειραματικά ότι η αναλυση $tf - idf$ αν και πάρα πολύ παλιά είναι ένας πολύ ικανοποιητικός τρόπος για να μπορέσουμε να επεξεργαστούμε την πληροφορία που περιέχεται στην υφολογία ενός κειμένου. Αυτή η μέθοδος κανονικοποιεί κατάλληλα τον αριθμό της συχνότητας μίας λέξης που εμφανίζεται σε ένα έγγραφο με τον αριθμό των φορών που εμφανίζεται σε όλη την συλλογή των κειμένων μας προκειμένου να απαλείψουμε τις κοινότυπες λέξεις και να δώσουμε βάρος στις ειδικές, $highly - discriminant$ λέξεις. Παρόλα αυτά, θεωρήσαμε όλη αυτή την διαδικασία κυρίως σαν ένα μαύρο κουτί και απλά χρησιμοποιήσαμε τις παραγόμενες τιμές για όλες τις λέξεις των εγγράφων.

Επειδή η διαδικασία της εκπαίδευση του χάρτη μας είναι πολύ χρονοβόρα επιλέξαμε πιο μικρά πλέγματα σε μέγεθος και λιγότερες επαναλήψεις στις καταστάσεις του $Ordering$ και του $Tuning$. Αυτό το κάναμε διότι είδαμε εποπτικά απο τα αποτελέσματα των $U - Matrices$ 6 ότι η συμπεριφορά των χαρτών ακόμη και με λίγους νευρώνες (συνολικά 100 ή 400) είναι πάρα πολύ ικανοποιητικά και δείχνουν το διαισθητικά αναμενόμενο αποτέλεσμα του διαχωρισμού των νευρώνων σε 3 κυρίρχες συστάδες. Ο κυρίαρχος λόγος που συμβαίνει αυτό είναι ότι ο αυτοοργανούμενος χάρτης εξαιτίας της δομής του μπορεί να καταλήξει σε πολύ καλές προσεγγίσεις των προτύπων στον $\mathbf{R}^N$ με πολύ λίγες επαναλήψεις. Αυτό συμβαίνει διότι στην φάση του $Ordering$ ο χάρτης μας προσπαθεί γρήγορα να φτάσει στα δεδομένα μας ενώ στην φάση του $Tuning$ κάνει πολύ μικρές κινήσεις. Προφανώς σε έναν χώρο 8296 διαστάσεων αυτό που μετράει και ειδικά στο πρόβλημα του $clustering$ είναι μια σχετικά καλή προσέγγιση και όχι ένα $fine - tuned$

3Γ′

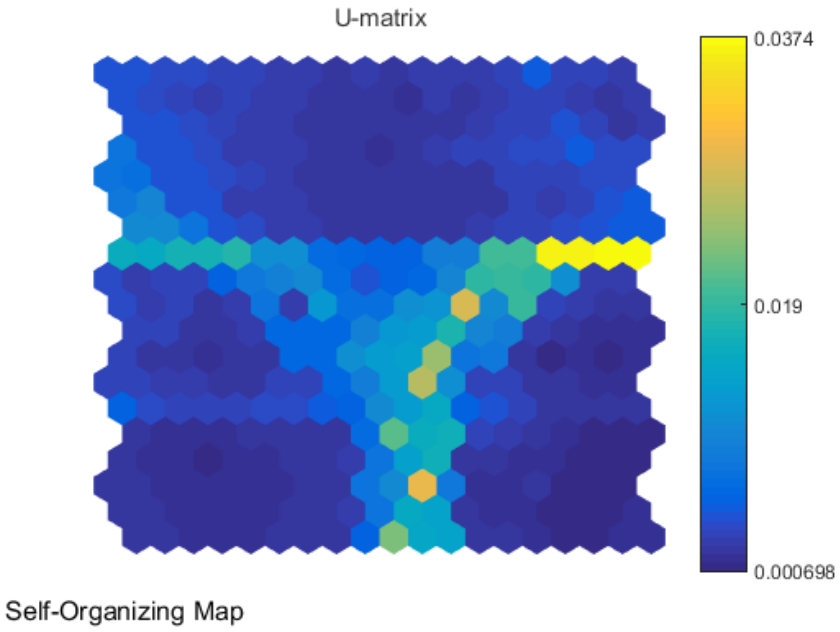
Παρακάτω απαντάμε σε όλες τις ερωτήσεις της άσκησης αφού έχουμε εκπαιδεύσει τον χάρτη μας πάνω στο επιλεγμένο μοντέλο (Εξαγωνικό Πλέγμα, Ευκλείδια απόσταση και πλέγμα 10×10).

i

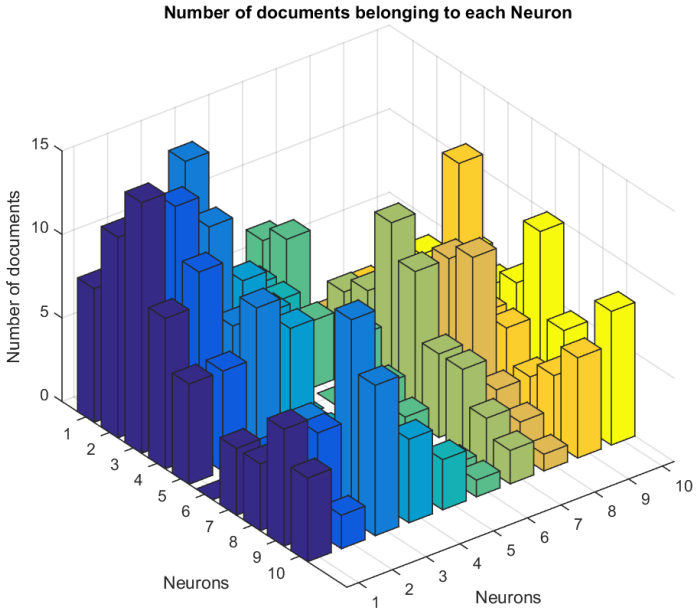
Υπολογίζουμε όλες τις αποστάσεις από όλα τα πρότυπα προς τις τελικές θέσεις των νευρώνων. Με βάση αυτές βρίσκουμε για κάθε πρότυπο ποιός είναι ο νευρώνας που είναι πιο κοντά του και αυξάνουμε στον αντίστοιχο νευρώνα τον μετρητή του. Στο 7 βλέπουμε τα τον αριθμό εγγράφων που αντιστοιχίζεται σε κάθε νευρώνα από τους 10-10 του πλέγματος. Παρατηρούμε ότι υπάρχουν 3 ΄βουνά' σε αυτό το διάγραμμα που ανάμεσα τους έχουν ξεκάθαρο χώρισμα. Αυτό μας οδηγεί στο συμπέρασμα ότι τα έγγραφα χωρίζονται σε 3 ομάδες ανάλογα με το περιεχόμενο τους.

ii

Αυτή την φορά υπολογίζουμε τις αποστάσεις από τις τελικές θέσεις των νευρώνων προς όλα τα πρότυπα εισόδου. Με βάση αυτές, υπολογίζουμε για κάθε νευρώνα την ελάχιστη απόσταση του από τα πρότυπα και αναθέτουμε σε αυτόν ως τον πιό αντιπροσωπευτικό τίτλο, τον τίτλο του εγγράφου-προτύπου που βρίσκεται πιο

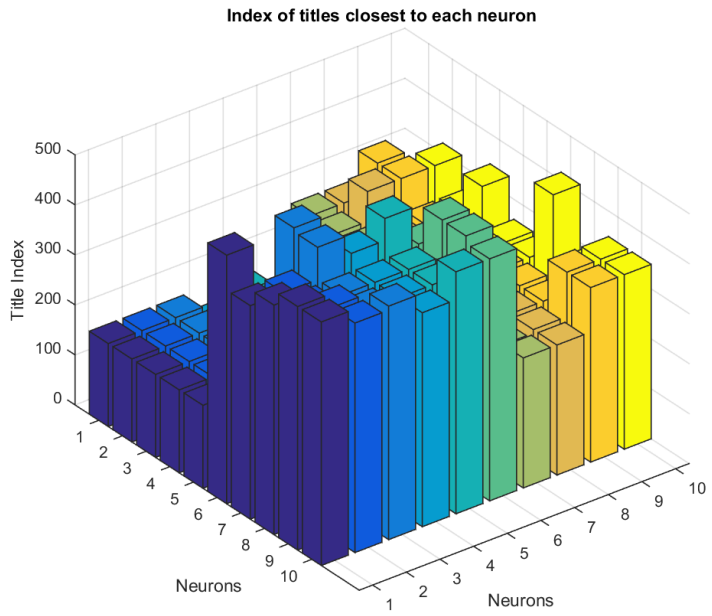


Σχήμα 6: Το *umatrix* μετά από την εκπαίδευση στα δεδομένα $NIPS$



Σχήμα 7: Αριθμός εγγράφων που αντιστοιχούν σε κάθε νευρώνα

κοντά σε αυτόν. Παραθέτονται 2 διαγράμματα για καλύτερη οπτικοποίηση των αποτελεσμάτων 8, 9 Συνολικά παρατηρούμε ότι και από το 8 φαίνεται ο διαχωρι-

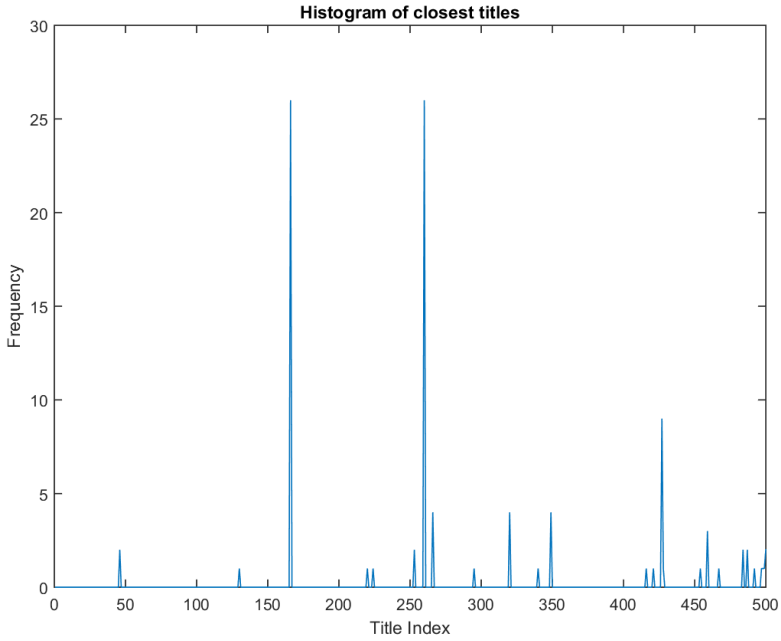


Σχήμα 8: Οι δείκτες των τίτλων που είναι πιο κοντά σε κάθε νευρώνα

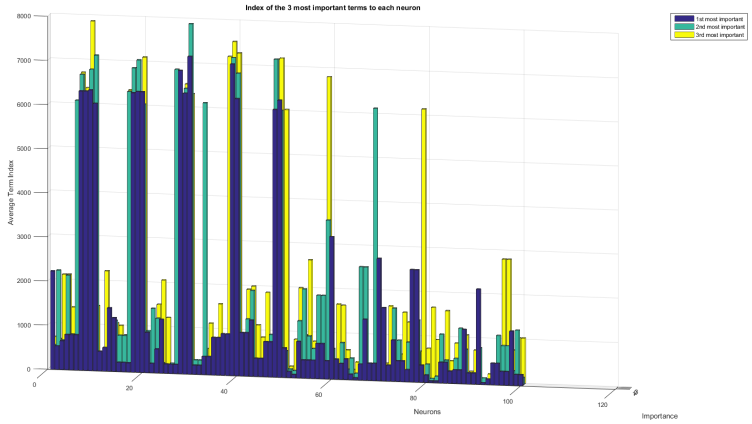
σμός των κειμένων σε 3 θεματικές ομάδες ενώ κάθε μία από αυτές τις ομάδες έχει ένα αντιπρόσωπο (όπως φαίνεται στο 9). Οι αντιπρόσωπο αυτόι είναι A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-test Split, Learning Curves for Gaussian Processes , Variational Inference for Bayesian Mixtures of Factor AnaIysers.

iii

Ο πίνακας *terms* περιέχει την αντιστοίχηση των δεικτών του διανύσματος χαρακτηριστικών, που έχει προκύψει για κάθε νευρώνα του δικτύου, με όλες τις λέξεις-διαστάσεις-χαρακτηριστικά. Συνεπώς βλέπουμε ποιές είναι οι πιο εμφανιζόμενες λέξεις για όλους τους εκπαιδευμένους νευρόνες απλά παίρνοντας τις 3 μέγιστες τιμές από το διάνυσμα χαρακτηριστικών τους (αντίστοιχη γραμμή στον IW) και βρίσκοντας από τον πίνακα *terms* την λέξη που αντιστοιχεί στους μέγιστους δείκτες. Παρατίθεται 10 με τους δείκτες των 3 πιο σημαντικών λέξεων για κάθε νευρώνα. Σε αυτό το διάγραμμα δεν φαίνεται ξεκάθαρα ο διαχωρισμός των λέξεων σε 3 ομάδες όπως φαινόταν ο διαχωρισμός των κειμένων πριν επειδή ο διαχωρισμός με βάση τις λέξεις είναι πιο θορυβώδεις. Παρόλαυτά πάλι μπορούμε να παρατηρήσουμε πως κοντινοί νευρώνες έχουν και σχετικά κοντινούς δείκτες λέξεων.



Σχήμα 9: Ιστόγραμμα των πιο κοντινών τίτλων



Σχήμα 10: Δείκτες των πιο σημαντικών λέξεων για κάθε νευρώνα

iv

Για να βρούμε τους αντίστοιχους δείκτες για τις λέξεις "$network$","$fuction$" απλά παίρνουμε τις τιμές των δεικτών του πίνακα $terms$ για αυτές τις λέξεις. Η εκφώνηση της άσκησης δεν είναι τόσο ξεκάθαρη για το πού αναφέρεται η μέγιστη τιμή. Η πρώτη περίπτωση είναι η μέγιστη τιμή να ανφέρεται στην μέγιστη τιμή του διανύσματος του ίδιου νευρώνα. Αυτή όμως η αναζήτηση στο ποιοί νευρώνες περιέχουν τις ζητούμενες λέξεις τουλάχιστον στο 30% της μέγιστης τιμής του

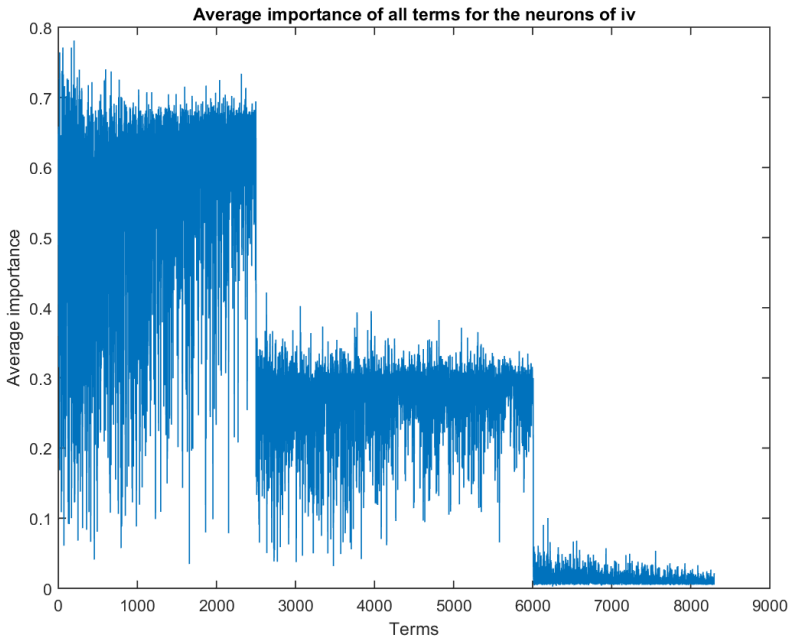
ίδιου διανύσματος δεν μας έδινε κανέναν νευρώνα που να ικανοποιεί την συνθήκη αυτή. Για υτό τον λόγο καταλήξαμε στην δεύτερη δυνατή ερμηνεία της μέγιστης τιμής που αναφέρεται στην μέγιστη τιμή αυτών των δύο λέξεων σε όλα τα εκπαιδευμένα διανύσματα των νευρώνων. Με αυτόν τον τρόπο βρήκαμε για όλους τους νευρώνες ποιές είναι οι τιμές των λέξεων "$network$","$fuction$" στην αντίστοιχη γραμμή του IW και βρήκαμε τις μέγιστες τιμές από όλες αυτές τις τιμές. Έστω αυτές οι τιμές $max_{network}, max_{function}$. Συνεπώς, οι ζητούμενοι νευρώνες θα είναι αυτοί που έχουν στην αντίστοιχη θέση της γραμμής του IW τιμές μεγαλύτερες από $0.3max_{network}, 0.3max_{function}$. Αυτό θα μπορούσε να χρησιμοποιηθεί για να βρίσκουμε μέσα από μία βάση δεδομένων τα πιο σχετικά έγγραφα με την αναζήτηση μας που την κάνουμε μέσω $key-words$. Οι νευρώνες που ικανοποιούν το παραπάνω κριτήριο τελικά είναι 36 από τους 100.

v

Η άσκηση και πάλι έχει λανθασμένη διατύπωση και εννοεί τους νευρώνες του ερωτήματος iv. Με βάση την προηγούμενη διαδικασία βρίσκουμε για όλα τα εκπαιδευμένα διανύσματα νευρώνων και για όλες τις διαστάσεις-λέξεις τις μέγιστες τιμές. Με βάση αυτές διαιρούμε για να κανονικοποιήσουμε όλες τις τιμές στο [0,1]. Τώρα όλες οι τιμές που έχουν τα διανύσματα των νευρώνων είναι κανονικοποιημένες ως προς το μέγιστο της κάθε λέξης. Τώρα παίρνουμε τους προηγούμενους 36 νευρώνες πυο βρήκαμε στο ερώτημα iv και από αυτά τα 36 διανύσματα βρίσκουμε τον μέσο όρο για την κάθε λέξη. Συνεπώς τώρα έχουμε ένα διάνυσμα διάστασης $1 \times \#words$ που περιέχει όλους αυτούς τους μέσους όρους. Παρατίθεται τελικό διάγραμμα αυτών των μέσων όρων 11 Στο 11 μπορούμε να παρατηρήσουμε καθαρά ότι οι λέξεις χωρίζονται σε 3 θεματικές ομάδες. Οι ομάδες 1-2 μοιάζουν περισσότερο από ότι η 1 και η 3 και ούτω καθεξής. Αυτό ενισχύει την αρχική μας υπόθεση ότι τα κείμενα που εξετάσαμε μπορούν να χωριστούν σε τρεις ομάδες ανάλογα με το περιεχόμενο τους. Αυτό το αποτέλεσμα είναι μεγαλειώδες και καταδεικνύει την αξία των αυτοοργανώμενων χαρτών για $unsupervisedclassification$ αφού καταφέραμε με τη χρήση τους να ταξινομήσουμε ένα τόσο μεγάλο και θορυβώδες σύνολο δεδομένων σε 3 ξεκάθαρες κλάσεις.

Αναφορές

[1] Peter Sarlin, (2011) "Evaluating a Self-Organizing Map for Clustering and Visualizing Optimum Currency Area Criteria", Economics Bulletin, Vol. 31 no.2 pp. 1483-1495.



Σχήμα 11: Μέσοι όροι του βάρους όλων των λέξεων για τους επιλεγμένους νευρώνες