ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ

«ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΑ ΣΥΣΤΗΜΑΤΑ»

ΜΕΤΑΠΤΥΧΙΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΟΠΤΙΚΟΠΟΙΗΣΗ ΠΟΛΥΔΙΑΣΤΑΤΩΝ ΔΕΔΟΜΕΝΩΝ»

ΣΟΡΟΒΟΥ ΕΙΡΗΝΗ (Α.Μ: 3262016121)

Επιβλέπων καθηγητής: Σταματάτος Ευστάθιος

Σάμος, Ιούνιος 2019

Περιεχόμενα

Περίληψη

Ένα από τα σημεία της τεχνολογικής ανάπτυξης είναι η ραγδαία αύξηση των δεδομένων που παράγονται καθημερινά. Δεδομένα πρωτοφανή σε όγκο, ποικιλομορφία, ταχύτητα παραγωγής και ανανέωσης. Η φύση τους μας οδήγησε στην ανάπτυξη τεχνικών ανάλυσης και απεικόνισης τους ώστε να γίνονται αντιληπτές οι σχέσεις και οι δομές που κρύβονται μέσα σε αυτά.

Στο πλαίσιο της παρούσας εργασίας αναλύονται, μέσα από βιβλιογραφική έρευνα, τεχνικές οπτικοποίησης δεδομένων εμβαθύνοντας και εφαρμόζοντας δυο εξ αυτών σε μια βάση δεδομένων 757 αρχείων.

Abstract

One of the technological development effects is the rapid growth of the daily produced data. Data unprecedented in volume, diversity, production speed and renewal. Their nature has led us to the development of analysis and visualization techniques in order to understand the hidden relationships and structures within them.

This thesis presents data visualization techniques and in the end deepening and applying two of them to a database of 757 files.

Εισαγωγή

Τα "μεγάλα δεδομένα" είναι ένα πεδίο που διαχειρίζεται τους τρόπους ανάλυσης και τη συστηματική απόσπαση πληροφοριών ή διαφορετικά διαχειρίζεται τα πακέτα δεδομένων που είναι πολύ μεγάλα ή περίπλοκα για να τα διαχειριστεί ένα παραδοσιακό λογισμικό εφαρμογών επεξεργασίας δεδομένων. Δεδομένα με πολλές πεδία (σειρές) προσφέρουν μεγαλύτερη στατιστική ισχύ, ενώ δεδομένα με μεγαλύτερη πολυπλοκότητα (περισσότερα χαρακτηριστικά ή στήλες) μπορεί να οδηγήσουν σε υψηλότερο ποσοστό ψευδών ανακαλύψεων. Οι μεγάλες προκλήσεις για τα δεδομένα περιλαμβάνουν τη συλλογή, την αποθήκευση, την ανάλυση, την αναζήτηση, την κοινή χρήση, τη μεταφορά, την απεικόνιση, τη δημιουργία ερωτημάτων, την ενημέρωση, την ιδιωτικότητα των πληροφοριών και την πηγή τους. Η τρέχουσα χρήση του όρου μεγάλα δεδομένα τείνει να αναφέρεται στη χρήση αναλυτικών στοιχείων πρόβλεψης, αναλύσεων συμπεριφοράς χρηστών ή ορισμένων άλλων προηγμένων μεθόδων ανάλυσης δεδομένων που εξάγουν αξία από δεδομένα και σπανίως σε ένα συγκεκριμένο μέγεθος συνόλου δεδομένων. Λόγω του πλήθους των χαρακτηριστικών που διέπουν τα μεγάλα δεδομένα είναι δύσκολη η απεικόνιση του συνόλου των χαρακτηριστικών και εργασία πάνω σε αυτά. Έτσι δημιουργήθηκε η ανάγκη ανάπτυξης αλγορίθμων οι οποίοι βοηθούν την επεξεργασία και την οπτικοποίηση πολυδιάστατων δεδομένων.

1. Δεδομένα Μεγάλου Όγκου

1.1. Ερμηνεία

Τα μεγάλα δεδομένα συνδέθηκαν αρχικά με τρεις βασικές έννοιες, τον όγκο, την ποικιλία και την ταχύτητα (Laney, 2001). Άλλες έννοιες που συνδέθηκαν αργότερα με την έννοια των μεγάλων δεδομένων είναι η εγκυρότητα (δηλαδή το ποσοστό του θορύβου ή της χαοτικής πληροφορίας που υπάρχει στα δεδομένα) (Goes, 2014) και η αξία τους (Rieder & Simon, 2017).

"Δεν υπάρχει αμφιβολία ότι οι διαθέσιμες ποσότητες δεδομένων είναι πράγματι μεγάλες, αλλά αυτό δεν είναι το πιο σχετικό χαρακτηριστικό αυτού του νέου οικοσυστήματος δεδομένων." (Boyd & Crawford, 2011). Η ανάλυση των συνόλων δεδομένων μπορεί να βρει νέους συσχετισμούς για να εντοπίσει τις τάσεις των επιχειρήσεων, να προλάβει τις ασθένειες, να καταπολεμήσει την εγκληματικότητα κ.ο.κ. (Power, 2014). Οι επιστήμονες, τα στελέχη επιχειρήσεων, οι γιατροί, οι διαφημιστές και οι κυβερνήσεις αντιμετωπίζουν πολύ συχνά δυσκολίες με τη διαχείριση μεγάλων συνόλων δεδομένων σε τομείς όπως η αναζήτηση στο Διαδίκτυο, η χρήση νέων τεχνολογιών στον χρηματοπιστωτικό κλάδο (FinTech), η αστική αλλά και η επιχειρησιακή πληροφορική. Οι επιστήμονες αντιμετωπίζουν περιορισμούς στις εργασίες που σχετίζονται με την υπολογιστική εντατική επιστήμη που εκτελείται σε υψηλά κατανεμημένα περιβάλλοντα δικτύου, όπως η μετεωρολογία, η γονιδιωματική, η σύνθετη φυσική προσομοίωση, η βιολογία και η περιβαλλοντική έρευνα (Reichman et al., 2011).

Ο όρος χρησιμοποιείται από τη δεκαετία του '90, με πολλούς να πιστεύουν ότι ο John Mashey ήταν αυτός που έκανε τον όρο δημοφιλή (Lohr, 2013). Τα μεγάλα δεδομένα περιλαμβάνουν συνήθως σύνολα δεδομένων με μεγέθη πέρα από την ικανότητα των εργαλείων λογισμικού που χρησιμοποιούνται συνήθως για τη συλλογή, επεξεργασία, διαχείριση και επεξεργασία δεδομένων εντός ενός αποδεκτού χρόνου (Snijders et al., 2012). Η φιλοσοφία των μεγάλων δεδομένων περιλαμβάνει μη δομημένα, ημιδομημένα και δομημένα δεδομένα, ωστόσο η κύρια εστίαση είναι στα μη δομημένα δεδομένα (Dedić & Stanier, 2016). Το μέγεθος των μεγάλων δεδομένων είναι μια ρευστή έννοια, καθώς από το 2012 κυμαίνεται από μερικές δεκάδες terabytes έως πολλά exabytes δεδομένων (Wang, 2017). Τα μεγάλα δεδομένα απαιτούν ένα σύνολο τεχνικών και τεχνολογιών με νέες ολοκληρωμένες μορφές που να είναι ικανές να

αποκαλύψουν πληροφορίες από σύνολα δεδομένων που είναι ποικίλα, πολύπλοκα και μαζικής κλίμακας (Hashem et al., 2015).

Το 2016 δόθηκε ένας ορισμός που δήλωνε ότι "τα μεγάλα δεδομένα αντιπροσωπεύουν τα στοιχεία της πληροφορίας που χαρακτηρίζονται από τόσο μεγάλο όγκο, ταχύτητα και ποικιλία που απαιτούν συγκεκριμένη τεχνολογία και αναλυτικές μεθόδους για τη μετατροπή τους σε αξία" (De Mauro et al., 2016). Ομοίως, οι Kaplan και Haenlein ορίζουν τα μεγάλα δεδομένα ως "σύνολα δεδομένων που χαρακτηρίζονται από τεράστια ποσά δεδομένων (όγκος) με μεγάλη συχνότητα επικαιροποίησης (ταχύτητα) σε διάφορες μορφές, όπως αριθμητικά δεδομένα, δεδομένα κειμένου ή δεδομένα εικόνων / βίντεο (ποικιλία)" (Kaplan & Haenlein, 2019). Επιπλέον, από ορισμένες οργανώσεις, προστίθεται ως χαρακτηριστικό περιγραφής τους και η έννοια της αξιοπιστίας, ενώ ο αναθεωρητισμός αμφισβητείται από ορισμένες βιομηχανικές αρχές (Grimes, 2013). Τα τρία Vs (Ο όγκος, η ποικιλία και η ταχύτητα αποτελούν τα βασικά χαρακτηριστικά των μεγάλων δεδομένων, γνωστά και ως τα τρία Vs (Volume, Variety, Velocity), έχουν επεκταθεί περαιτέρω σε άλλα συμπληρωματικά χαρακτηριστικά των μεγάλων δεδομένων (Hilbert, 2016).

1.2. Χαρακτηριστικά

Τα μεγάλα δεδομένα μπορούν να περιγραφούν από τα ακόλουθα χαρακτηριστικά (Hilbert, 2016):

Όγκος

Η ποσότητα των παραγόμενων και αποθηκευμένων δεδομένων. Το μέγεθος των δεδομένων καθορίζει την αξία και τη δυνητική γνώση και αν μπορεί βάσει αυτού να θεωρηθούν μεγάλα δεδομένα ή όχι.

Ποικιλία

Ο τύπος και η φύση των δεδομένων. Αυτό βοηθά τα άτομα που τα αναλύουν να χρησιμοποιήσουν αποτελεσματικά την προκύπτουσα γνώση. Τα μεγάλα δεδομένα αντλούνται

από κείμενο, εικόνες, ήχο, βίντεο και επιπλέον, τα κομμάτια που λείπουν ολοκληρώνονται μέσω της σύντηξης δεδομένων.

Ταχύτητα

Η ταχύτητα με την οποία παράγονται και επεξεργάζονται τα δεδομένα για να ανταποκριθούν στις απαιτήσεις και τις προκλήσεις που βρίσκονται στην πορεία της αύξησης και της ανάπτυξής τους. Μεγάλα δεδομένα είναι συχνά διαθέσιμα σε πραγματικό χρόνο. Σε σύγκριση με τα μικρά δεδομένα, τα μεγάλα δεδομένα παράγονται συνεχώς. Τα δύο είδη ταχύτητας που σχετίζονται με τα μεγάλα δεδομένα είναι η συχνότητα παραγωγής και η συχνότητα χειρισμού, καταγραφής και δημοσίευσης (Kitchin & McArdle, 2016).

Εγκυρότητα

Είναι ο εκτεταμένος ορισμός για μεγάλα δεδομένα, ο οποίος αναφέρεται στην ποιότητα των δεδομένων και την τιμή των δεδομένων. Η ποιότητα των συλλεγόμενων δεδομένων μπορεί να ποικίλει σημαντικά, επηρεάζοντας την ακριβή ανάλυση τους.

Τα δεδομένα πρέπει να υποβάλλονται σε επεξεργασία με προηγμένα εργαλεία (αναλυτικά στοιχεία και αλγόριθμους) για την αποκάλυψη σημαντικών πληροφοριών. Για παράδειγμα, για τη διαχείριση ενός εργοστασίου, πρέπει να λαμβάνονται υπόψη τόσο τα ορατά όσο και τα αόρατα ζητήματα με μια ποικιλία στοιχείων. Οι αλγόριθμοι παραγωγής πληροφοριών πρέπει να ανιχνεύουν και να αντιμετωπίζουν τα αόρατα ζητήματα όπως η υποβάθμιση του μηχανήματος, η φθορά εξαρτημάτων κ.λπ. ακόμα και από το κατώτατο επίπεδο του εργοστασίου (Lee et al., 2014).

1.3. Παραδείγματα Δεδομένων Μεγάλου Όγκου στον Πραγματικό κόσμο

Εφαρμογές

Ενώ πολλοί προμηθευτές προσφέρουν λύσεις για μεγάλα δεδομένα, οι εμπειρογνώμονες συνιστούν την ανάπτυξη εσωτερικών λύσεων ειδικά προσαρμοσμένων για την επίλυση των προβλημάτων της εταιρείας, εφόσον η εταιρεία διαθέτει επαρκείς τεχνικές δυνατότητες (Sabarmathi & Chinnaiyan, 2017).

Κυβέρνηση

Η χρήση και η υιοθέτηση μεγάλων δεδομένων στο πλαίσιο κυβερνητικών διαδικασιών επιτρέπει την αποδοτικότητα όσον αφορά το κόστος, την παραγωγικότητα και την καινοτομία, αλλά όλα αυτά συνοδεύονται από ελαττώματα. Η ανάλυση δεδομένων συχνά απαιτεί τη συνεργασία πολλών κυβερνητικών τμημάτων, τόσο κεντρικών όσο και τοπικών, ώστε να δημιουργήσουν νέες και καινοτόμες διαδικασίες για την επίτευξη του επιθυμητού αποτελέσματος.

Η υπηρεσία ¨Πολιτικών Εγγραφών και Κρίσιμων Στατιστικών (Civil Registration and Vital Statistics - CRVS)¨ συλλέγει όλα τα πιστοποιητικά από τη γέννηση μέχρι το θάνατο. Η CRVS αποτελεί μια σημαντική πηγή μεγάλων δεδομένων για τις κυβερνήσεις.

Διεθνής Ανάπτυξη

Η έρευνα σχετικά με την αποτελεσματική χρήση των ¨Τεχνολογιών που σχετίζονται με τις Πληροφορίες και τις Επικοινωνίες για την Ανάπτυξη (Information and Communication Technologies for Development – ICT4D)¨, υποδηλώνει ότι η τεχνολογία μεγάλων δεδομένων μπορεί να συμβάλει σημαντικά αλλά και να φέρει στο προσκήνιο μοναδικές προκλήσεις για τη διεθνή ανάπτυξη. Οι πρόοδοι στη μεγάλη ανάλυση δεδομένων προσφέρουν οικονομικά αποδοτικές ευκαιρίες για τη βελτίωση στη λήψη αποφάσεων σε κρίσιμους τομείς ανάπτυξης όπως η υγειονομική περίθαλψη, η απασχόληση, η οικονομική παραγωγικότητα, το έγκλημα, η ασφάλεια και η διαχείριση φυσικών πόρων και φυσικών καταστροφών (Hilbert, 2013). Επιπλέον, τα δεδομένα που δημιουργούνται από τους χρήστες προσφέρουν νέες ευκαιρίες για να δώσουν φωνή σε ότι δεν ακούγεται και σε άγνωστε κοινωνικές πτυχές (Yang et al., 2017). Ωστόσο, οι μακροχρόνιες προκλήσεις για τις αναπτυσσόμενες περιφέρειες, όπως η ανεπαρκής τεχνολογική υποδομή και η έλλειψη οικονομικών και ανθρώπινων πόρων, επιτείνουν τις υφιστάμενες ανησυχίες που σχετίζονται με τα μεγάλα δεδομένα, όπως η ιδιωτικότητα, η ελλιπής μεθοδολογία και τα θέματα διαλειτουργικότητας (Hilbert, 2013).

Βιομηχανία

Με βάση τη Μελέτη Παγκόσμιων Τάσεων TCS 2013, οι βελτιώσεις τόσο στον προγραμματισμό της προσφοράς, όσο και στην ποιότητα των παραγόμενων προϊόντων παρέχουν το μεγαλύτερο όφελος από την ανάλυση των μεγάλων δεδομένων, όσον αφορά τη βιομηχανία. Τα μεγάλα δεδομένα παρέχουν μια υποδομή διαφάνειας στην κατασκευαστική βιομηχανία, η οποία μεταφράζεται στη δυνατότητα να ξεδιαλύνονται αβεβαιότητες όπως η ασυνεπείς αποδόσεις και

η διαθεσιμότητα διαφόρων εξαρτημάτων. Η προγνωστική παραγωγή ως εφαρμόσιμη προσέγγιση για την επίτευξη ενός σχεδόν μηδενικού χρόνου διακοπής της παραγωγής και η διαφάνεια απαιτούν τεράστιες ποσότητες δεδομένων και προηγμένα εργαλεία πρόβλεψης για τη συστηματική επεξεργασία των δεδομένων και τη μετατροπή τους σε χρήσιμες πληροφορίες (Lee et al., 2014). Ένα εννοιολογικό πλαίσιο πρόβλεψης παραγωγής ξεκινά με την απόκτηση δεδομένων, όπου διατίθενται διαφορετικοί τύποι αισθητήριων δεδομένων, όπως δεδομένα ακουστικής, δόνησης, πίεσης, ρεύματος, τάσης και ελέγχου. Μια μεγάλη ποσότητα αισθητήριων δεδομένων σε συνδυασμό με τα ιστορικά δεδομένα δομούν τα μεγάλα δεδομένα στη βιομηχανία. Τα παραγόμενα μεγάλα δεδομένα λειτουργούν ως εισροές σε εργαλεία πρόληψης και προληπτικές στρατηγικές όπως η ¨Προγνωστική και η Διαχείριση της Υγείας (Prognostics and Health Management - PHM)¨.

Υγειονομική Περίθαλψη

Η ανάλυση των μεγάλων δεδομένων έχει βοηθήσει στη βελτίωση της υγειονομικής περίθαλψης παρέχοντας εξατομικευμένη ιατρική βοήθεια, παρεμβάσεις σε περιπτώσεις μεγάλου κλινικού ρίσκου και πρόβλεψη αναλύσεων, μείωση της αστάθειας και της μεταβλητότητας όσον αφορά στη φροντίδα των ασθενών και στα νοσοκομειακά απόβλητα, αυτοματοποιημένη εξωτερική και εσωτερική αναφορά δεδομένων ασθενών, τυποποίηση σε διάφορους ιατρικούς όρους και μητρώα ασθενών και λύσεις που βασίζονται σε κατακερματισμένα σημεία (Huser & Cimino, 2016). Σε ορισμένους τομείς υπάρχουν ακόμα πολύ μεγάλα περιθώρια βελτίωσης απ’ό,τι στην πραγματικότητα υλοποιούνται. Το επίπεδο των δεδομένων που παράγονται στα συστήματα υγειονομικής περίθαλψης δεν είναι ασήμαντο. Με την προστιθέμενη υιοθέτηση των τεχνολογιών που υποστηρίζουν την ιατρική φροντίδα μέσω φορητών συσκευών (mHealth), μέσω διαδικτύου (eHealth) και συσκευών που φέρουν στο σώμα τους οι ασθενείς, ο όγκος των δεδομένων θα συνεχίσει να αυξάνεται. Αυτό περιλαμβάνει τα ηλεκτρονικά δεδομένα καρτών υγείας, τα δεδομένα απεικόνισης, τα δεδομένα που παράγονται από τον ασθενή, τα δεδομένα των αισθητήρων και άλλες μορφές δεδομένων που είναι δύσκολο να επεξεργαστούν. Στις μέρες μας υπάρχει ακόμη μεγαλύτερη η ανάγκη για τέτοια περιβάλλοντα που θα δίνουν μεγαλύτερη προσοχή στην ποιότητα των δεδομένων και των πληροφοριών (Sejdic & Falk, 2018). "Τα μεγάλα δεδομένα πολύ συχνά σημαίνουν «βρώμικα δεδομένα» και το ποσοστό των ανακριβών δεδομένων αυξάνεται με την αύξηση του όγκου δεδομένων». Ο ανθρώπινος έλεγχος στην

κλίμακα των μεγάλων δεδομένων είναι αδύνατος και υπάρχει μια απελπιστική ανάγκη στην υπηρεσία υγείας για έξυπνα εργαλεία που θα είναι ικανά να ελέγχουν με ακρίβεια την πιστότητα και τη διαχείριση των πληροφοριών που χάθηκαν (Raghupathi & Raghupathi, 2014). Καθώς οι εκτεταμένες πληροφορίες στην υγειονομική περίθαλψη είναι πλέον ηλεκτρονικές, αυτές ταξινομούνται ως μεγάλα δεδομένα, καθώς οι περισσότερες είναι αδόμητες και δύσκολες στη χρήση (Viceconti et al., 2015). Η χρήση των μεγάλων δεδομένων στην υγειονομική περίθαλψη έχει προκαλέσει σημαντικές ηθικές προκλήσεις που θα μπορούσαν να μεταφραστούν ως οι κίνδυνοι παραβίασης των ατομικών δικαιωμάτων, της ιδιωτικότητας και αυτονομίας και της διαφάνειας των διαδικασιών (O'donoghue & Herbert, 2012).

Εκπαίδευση

Μια μελέτη του Διεθνούς Ινστιτούτου McKinsey διαπίστωσε έλλειψη 1,5 εκατομμυρίων επαγγελματιών και διαχειριστών μεγάλων δεδομένων με υψηλό επίπεδο εκπαίδευσης, καθώς και μια σειρά πανεπιστημίων, όπως το Πανεπιστήμιο του Τενεσί και το Πανεπιστήμιο του Μπέρκλεϋ, όπου δημιούργησαν προγράμματα μάστερ για την κάλυψη αυτής της ζήτησης. Τα ιδιωτικά κέντρα εκπαίδευσης έχουν επίσης αναπτύξει προγράμματα για την κάλυψη αυτής της ζήτησης, συμπεριλαμβανομένων δωρεάν προγραμμάτων όπως ¨Το Εκκολαπτήριο Δεδομένων (The Data Incubator)¨ ή πληρωμένα προγράμματα όπως η ¨Γενική Συνέλευση (General Assembly)¨ (Murdoch & Detsky, 2013). Στο συγκεκριμένο τομέα μάρκετινγκ, ένα από τα προβλήματα που τονίζουν οι Wedel και Kannan (Vayena et al., 2015) είναι ότι το μάρκετινγκ έχει αρκετούς υποτομείς όπως η διαφήμιση, οι προσφορές, η ανάπτυξη προϊόντων και το branding που όλοι τους χρησιμοποιούν διαφορετικά είδη δεδομένων. Επειδή δεν είναι επιθυμητές οι αναλυτικές λύσεις ενός μεγέθους, τα σχολεία επιχειρήσεων θα πρέπει να προετοιμάσουν τους διαχειριστές μάρκετινγκ να έχουν ευρεία γνώση για όλες τις διαφορετικές τεχνικές που χρησιμοποιούνται σε αυτούς τους υποτομείς ώστε να αποκτήσουν μια πιο ευρεία εικόνα και να συνεργαστούν αποτελεσματικότερα με τους αναλυτές.

Μέσα Μαζική Ενημέρωσης

Για να γίνει κατανοητός ο τρόπος με τον οποίο τα μέσα μαζικής ενημέρωσης χρησιμοποιούν τα μεγάλα δεδομένα, πρέπει πρώτα να δοθεί κάποιο πλαίσιο στο μηχανισμό που χρησιμοποιείται για τη διαδικασία των μέσων ενημέρωσης. Έχει προταθεί από τους Nick Couldry και Joseph Turow ότι οι ασχολούμενοι με τα Μέσα και τη Διαφήμιση προσεγγίζουν τα μεγάλα δεδομένα,

όπως πολλά σημεία ενεργητικής πληροφόρησης για εκατομμύρια άτομα. Η βιομηχανία φαίνεται να απομακρύνεται από την παραδοσιακή προσέγγιση της χρήσης συγκεκριμένων μέσων μαζικής ενημέρωσης, όπως εφημερίδες, περιοδικά ή τηλεοπτικές εκπομπές, και αντί αυτών στοχεύει σε καταναλωτές με τεχνολογίες που φτάνουν σε στοχευμένους ανθρώπους σε βέλτιστες χρονικές στιγμές και τοποθεσίες. Ο απώτερος στόχος είναι να υπηρετήσουν ή να μεταδώσουν ένα μήνυμα ή ένα περιεχόμενο που στατιστικά είναι σύμφωνο με τη νοοτροπία του καταναλωτή. Για παράδειγμα, τα περιβάλλοντα εκδόσεων προσαρμόζουν όλο και περισσότερο τα μηνύματα - διαφημίσεις και το περιεχόμενο - άρθρα για να απευθύνονται σε καταναλωτές που έχουν συλλεχθεί αποκλειστικά μέσω διαφόρων δραστηριοτήτων εξόρυξης δεδομένων.

Ασφάλιση

Οι πάροχοι ασφάλισης υγείας συγκεντρώνουν δεδομένα σχετικά με τους καθοριστικούς κοινωνικούς παράγοντες που σχετίζονται με την υγεία όπως η κατανάλωση τροφίμων, η οικογενειακή κατάσταση, το μέγεθος των ενδυμάτων και οι αγοραστικές συνήθειες, από τις οποίες προβαίνουν σε προβλέψεις για το κόστος υγείας, προκειμένου να εντοπίζουν τα θέματα υγείας των πελατών τους. Είναι αμφιλεγόμενο αν αυτές οι προβλέψεις χρησιμοποιούνται επί του παρόντος για τιμολόγηση (Couldry & Turow, 2014).

2. Μείωση Διαστάσεων

2.1. Ορισμός

Στα προβλήματα κατηγοριοποίησης μηχανικής μάθησης, υπάρχουν συχνά πάρα πολλοί παράγοντες βάσει των οποίων γίνεται η τελική κατηγοριοποίηση. Αυτοί οι παράγοντες είναι βασικές μεταβλητές που ονομάζονται χαρακτηριστικά. Όσο μεγαλύτερος είναι ο αριθμός των χαρακτηριστικών, τόσο πιο δύσκολη γίνεται η απεικόνιση του συνόλου των χαρακτηριστικών και στη συνέχεια η εργασία σε αυτά. Μερικές φορές, τα περισσότερα από αυτά τα χαρακτηριστικά συσχετίζονται και επομένως είναι περιττά. Σε αυτό το σημείο αναλαμβάνουν δράση οι αλγόριθμοι μείωσης της διαστατικότητας. Η μείωση της διαστατικότητας είναι η διαδικασία μείωσης του αριθμού των τυχαίων μεταβλητών που εξετάζονται, επιτυγχάνοντας ένα σύνολο κύριων μεταβλητών. Αυτή μπορεί να χωριστεί σε επιλογή χαρακτηριστικών και σε εξαγωγή χαρακτηριστικών.

2.2. Ανάγκη για μείωση της διαστατικότητας

Ένα ευκολονόητο παράδειγμα μείωσης της διαστατικότητας μπορεί να συζητηθεί μέσω ενός απλού προβλήματος κατηγοριοποίησης μηνυμάτων ηλεκτρονικού ταχυδρομείου, όπου πρέπει να κατηγοριοποιηθούν τα μηνύματα σχετικά με το αν είναι spam ή όχι. Αυτό μπορεί να περιλαμβάνει έναν μεγάλο αριθμό χαρακτηριστικών, όπως με το αν το ηλεκτρονικό εισερχόμενο μήνυμα έχει ή δεν έχει έναν γενικό τίτλο, με το περιεχόμενο του ηλεκτρονικού μηνύματος, με το εάν το ηλεκτρονικό μήνυμα χρησιμοποιεί ένα συγκεκριμένο πρότυπο κλπ. Ωστόσο, ορισμένα από αυτά τα χαρακτηριστικά μπορεί να επικαλύπτονται. Σε μια άλλη συνθήκη, ένα πρόβλημα κατηγοριοποίησης που βασίζεται τόσο στην υγρασία όσο και στις βροχοπτώσεις μπορεί να καταρρεύσει σε ένα μόνο υποκείμενο χαρακτηριστικό, αφού αμφότερα τα παραπάνω συσχετίζονται σε υψηλό βαθμό. Ως εκ τούτου, μπορούμε να μειώσουμε τον αριθμό των χαρακτηριστικών σε τέτοια προβλήματα. Ένα πρόβλημα κατηγοριοποίησης τριών διαστάσεων μπορεί να είναι δύσκολο να απεικονιστεί, ενώ ένα πρόβλημα δύο διαστάσεων μπορεί να χαρτογραφηθεί σε ένα απλό χώρο δυο διαστάσεων και ένα πρόβλημα μιας διάστασης θα μπορούσε να απεικονιστεί σε μια απλή γραμμή.

Πλεονεκτήματα της Μείωσης των Διαστατικότητας

- Βοηθά στη συμπίεση δεδομένων και επομένως στη μείωση του αποθηκευτικού χώρου.
- Μειώνει το χρόνο υπολογισμού.
- Βοηθά επίσης στην αφαίρεση περιττών χαρακτηριστικών, εάν υπάρχουν.

Μειονεκτήματα της Μείωσης της Διαστατικότητας

- Μπορεί να οδηγήσει σε μερική απώλεια δεδομένων.
- Η ¨Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA)¨ τείνει να βρει γραμμικούς συσχετισμούς μεταξύ μεταβλητών, κάτι που είναι μερικές φορές ανεπιθύμητο.
- Η PCA αποτυγχάνει σε περιπτώσεις όπου ο μέσος όρος και η συνδιακύμανση δεν επαρκούν για τον ορισμό συνόλων δεδομένων.
- Μπορεί να μην είναι γνωστός ο βασικός αριθμός συνιστωσών που πρέπει να διατηρηθούν - στην πράξη, εφαρμόζονται ορισμένοι εμπειρικοί κανόνες.

Η κατάρα της διαστατικότητας αναφέρεται σε διάφορα φαινόμενα που προκύπτουν όταν αναλύονται και οργανώνονται δεδομένα σε χώρους μεγάλης διαστάσεως (συχνά με εκατοντάδες ή χιλιάδες διαστάσεις) που δεν εμφανίζονται σε περιβάλλοντα μικρών διαστάσεων όπως ο τρισδιάστατος φυσικός χώρος της καθημερινής εμπειρίας. Η έκφραση επινοήθηκε από τον Richard E. Bellman κατά την εξέταση προβλημάτων στη δυναμική βελτιστοποίηση (Abo-Sinna et al., 2014), (Bellman, 2015).

Τα παραπάνω φαινόμενα εμφανίζονται σε τομείς όπως η αριθμητική ανάλυση, η δειγματοληψία, η συνδυαστική, η μηχανική μάθηση, η εξόρυξη δεδομένων και οι βάσεις δεδομένων. Το κοινό γνώρισμα αυτών των προβλημάτων είναι ότι όταν αυξάνεται η διαστασιμότητα, ο όγκος του χώρου αυξάνεται τόσο γρήγορα ώστε τα διαθέσιμα δεδομένα καθίστανται αραιά. Αυτή η ανεπάρκεια είναι προβληματική για οποιαδήποτε μέθοδο που απαιτεί στατιστική σημασία. Προκειμένου να επιτευχθεί ένα στατιστικώς ηχηρό και αξιόπιστο αποτέλεσμα, η ποσότητα των δεδομένων που απαιτούνται για την υποστήριξη του αποτελέσματος αυξάνεται συχνά εκθετικά με τη διαστασιμότητα. Επίσης, η οργάνωση και η αναζήτηση δεδομένων συχνά βασίζεται στον εντοπισμό περιοχών όπου αντικείμενα σχηματίζουν ομάδες με παρόμοιες ιδιότητες. Σε δεδομένα

μεγάλων διαστάσεων, ωστόσο, όλα τα αντικείμενα φαίνεται να είναι αραιά και ανόμοια με πολλούς τρόπους, γεγονός που εμποδίζει την αποτελεσματικότητα των κοινών στρατηγικών οργάνωσης δεδομένων.

2.3. Η Κατάρα της Διαστατικότητας

Η αύξηση του αριθμού των διαστάσεων ενός συνόλου δεδομένων σημαίνει ότι υπάρχουν περισσότερες καταχωρήσεις στο φορέα των χαρακτηριστικών που αντιπροσωπεύουν κάθε παρατήρηση στον αντίστοιχο Ευκλείδειο χώρο. Μετριέται η απόσταση σε ένα διανυσματικό χώρο χρησιμοποιώντας την Ευκλείδεια απόσταση.

Η Ευκλείδεια απόσταση μεταξύ δύο διανυσμάτων του δισδιάστατου χώρου με Καρτεσιανές συντεταγμένες *p = (p1, p2, ..., pn)* και *q = (q1, q2, ..., qn)* υπολογίζεται χρησιμοποιώντας το γνωστό τύπο:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Επομένως, κάθε νέα διάσταση προσθέτει ένα μη αρνητικό όρο στο άθροισμα, έτσι ώστε η απόσταση να αυξάνεται με τον αριθμό των διαστάσεων για διακεκριμένα διανύσματα. Με άλλα λόγια, καθώς ο αριθμός των χαρακτηριστικών αυξάνεται για ένα δεδομένο αριθμό παρατηρήσεων, ο χώρος χαρακτηριστικών γίνεται όλο και πιο αραιός. δηλαδή, λιγότερο πυκνός και περισσότερο άδειος. Από την άλλη πλευρά, η χαμηλότερη πυκνότητα δεδομένων απαιτεί περισσότερες παρατηρήσεις για να διατηρηθεί η ίδια μέση απόσταση μεταξύ των σημείων των δεδομένων.

Παρόμοια με τα παραπάνω στοιχεία, για τη διατήρηση της μέσης απόστασης 10 παρατηρήσεων ομοιόμορφα κατανεμημένων σε μια γραμμή, η απόσταση αυξάνει εκθετικά από 10^1 σε μία μόνο διάσταση σε 10^2 για τις δύο διαστάσεις και σε 10^3 για τις τρεις διαστάσεις, δεδομένου ότι τα δεδομένα πρέπει να επεκταθούν κατά ένα συντελεστή ίσο με 10, κάθε φορά που προστίθεται μια νέα διάσταση.

2.4. Γιατί η Μείωση της Διαστατικότητας είναι Σημαντική;

Στη μηχανική μάθηση, για να συλλεχθούν χρήσιμοι δείκτες και να υπάρχει ένα πιο ακριβές αποτέλεσμα, υπάρχει η τάση να προστίθενται πρώτα όσο το δυνατόν περισσότερα στοιχεία. Ωστόσο, μετά από ένα συγκεκριμένο σημείο, η απόδοση του μοντέλου αρχίζει να μειώνεται με τον αυξανόμενο αριθμό στοιχείων. Αυτό το φαινόμενο αναφέρεται συχνά ως "Η κατάρα της Διαστατικότητας".

Η κατάρα της διαστατικότητας συμβαίνει επειδή η πυκνότητα του δείγματος μειώνεται εκθετικά με την αύξηση της διαστατικότητας. Όταν συνεχίζουμε να προσθέτουμε λειτουργίες χωρίς να αυξάνουμε τον αριθμό των δειγμάτων κατάρτισης, η διάσταση του χώρου των χαρακτηριστικών αυξάνεται και γίνεται ολοένα και πιο αραιή. Λόγω αυτής της ιδιαιτερότητας, γίνεται πολύ πιο εύκολο να βρεθεί μια "τέλεια" λύση για το μοντέλο της μηχανικής μάθησης, η οποία κατά πάσα πιθανότητα οδηγεί σε υπερπροσαρμογή.

Η υπερπροσαρμογή συμβαίνει όταν το μοντέλο αντιστοιχεί πολύ κοντά σε ένα συγκεκριμένο σύνολο δεδομένων και δεν μπορεί εύκολα να γενικευτεί. Ένα υπερπροσαρμοσμένο μοντέλο θα λειτουργούσε πάρα πολύ καλά στο συγκεκριμένο σύνολο δεδομένων κατάρτισης, ωστόσο θα αποτύγχανε στα μελλοντικά δεδομένα με αποτέλεσμα να κάνει την πρόβλεψη αναξιόπιστη.

Με ποιο τρόπο, λοιπόν, θα μπορούσε να ξεπεραστεί η κατάρα της διαστατικότητας και να αποφευχθεί η υπερπροσαρμογή, ειδικά όταν έχουμε πολλά χαρακτηριστικά και συγκριτικά λίγα δείγματα κατάρτησης; Μια δημοφιλής προσέγγιση είναι η μείωση της διαστατικότητας.

Η μείωση της διαστατικότητας είναι η διαδικασία μείωσης της διαστατικότητας του χώρου των χαρακτηριστικών, λαμβάνοντας υπόψη την απόκτηση ενός συνόλου βασικών χαρακτηριστικών. Η μείωση της διαστατικότητας μπορεί να διαχωριστεί περαιτέρω στην επιλογή χαρακτηριστικών και στην εξαγωγή χαρακτηριστικών.

Η επιλογή των δυνατοτήτων προσπαθεί να επιλέξει ένα υποσύνολο των αρχικών χαρακτηριστικών για χρήση στο μοντέλο μηχανικής μάθησης. Με αυτόν τον τρόπο, θα μπορούσαν να καταργηθούν περιττές και άσχετες λειτουργίες χωρίς την ύπαρξη πολλών απωλειών πληροφοριών.

Η εξαγωγή των χαρακτηριστικών ονομάζεται επίσης προβολή των χαρακτηριστικών. Ενώ η επιλογή των χαρακτηριστικών επιστρέφει ένα υποσύνολο των αρχικών χαρακτηριστικών, η εξαγωγή χαρακτηριστικών δημιουργεί νέες λειτουργίες, προβάλλοντας τα δεδομένα ενός χώρου υψηλών διαστάσεων σε ένα χώρο με λιγότερες διαστάσεις. Αυτή η προσέγγιση μπορεί επίσης να αντλήσει πληροφοριακά και μη πλεονάζοντα χαρακτηριστικά.

Είναι δυνατό να χρησιμοποιηθούν μαζί η επιλογή χαρακτηριστικών και η εξαγωγή χαρακτηριστικών. Η εξαγωγή χαρακτηριστικών μπορεί να εκτελεστεί σε επιλεγμένα στοιχεία που περιέχουν σχετικές πληροφορίες αντί για χρήση των αρχικών χαρακτηριστικών. Εκτός από την αποφυγή της υπερπροσαρμογής και του πλεονασμού, η μείωση της διαστατικότητας, με την απλοποίηση των μοντέλων, οδηγεί επίσης σε καλύτερες ανθρώπινες ερμηνείες και σε λιγότερο υπολογιστικό κόστος.

3. Τεχνικές για μείωση της διαστατικότητας

Το πρόβλημα της (μη γραμμικής) μείωσης των διαστάσεων μπορεί να οριστεί ως εξής. Υποθέτουμε ότι έχουμε ένα σύνολο δεδομένων που παριστάνεται σε ένα πίνακα **X,** n × *D* που αποτελείται από n διανύσματα δεδομένων $\mathbf{x}_i$ (I ϵ {1, 2, …,n}) με τη διάσταση *D.* Υποτίθεται περαιτέρω ότι αυτό το σύνολο δεδομένων έχει εγγενή διάσταση *d* (όπου *d* <*D* , και συχνά *d*<<*D*). Εδώ, από μαθηματικούς όρους, η εγγενής διάσταση σημαίνει ότι τα σημεία στο σύνολο δεδομένων **X** βρίσκονται σε ή κοντά σε ένα πολλαπλό χωρίο με διάσταση *d* που είναι ενσωματωμένος στον *D* - διαστάσεων χώρο. Σημειώστε ότι δεν κάνουμε υποθέσεις σχετικά με τη δομή αυτού του πολλαπλού χωρίου: το πολλαπλό χωρίο μπορεί να είναι μη Riemann λόγω των ασυνεχειών (π.χ.., το πολλαπλό χωρίο μπορεί να αποτελείται από έναν αριθμό μη συνδεδεμένων πολλαπλών υποχωρίων). Οι τεχνικές μείωσης διαστάσεων μετασχηματίζουν το σύνολο δεδομένων **X** με διάσταση *D* σε ένα νέο σύνολο δεδομένων **Y** με διάσταση *d,* διατηρώντας τη γεωμετρία των δεδομένων όσο το δυνατόν περισσότερο. Γενικά, δεν είναι γνωστή ούτε η γεωμετρία των δεδομένων του πολλαπλού χωρίου ούτε η εγγενής διάσταση *d* του συνόλου δεδομένων X. Επομένως, η μείωση των διαστάσεων είναι ένα δύσκολα τιθέμενο πρόβλημα που μπορεί να λυθεί μόνο υποθέτοντας ορισμένες ιδιότητες των δεδομένων (όπως η εγγενής διάσταση τους). Στην εργασία των Maaten & Postma (2009), δηλώνεται ένα πολλών διαστάσεων σημείο δεδομένων με το $\mathbf{x_i}$, όπου $\mathbf{x_i}$ είναι η i-οστή γραμμή του πίνακα δεδομένων **X** διάστασης *D*. Το λίγων διαστάσεων αντίστοιχο δεδομένο του $\mathbf{x_i}$ συμβολίζεται με το $\mathbf{y_i}$, όπου $\mathbf{y_i}$ είναι η i-οστή γραμμή του πίνακα δεδομένων **Y** διαστάσεως των *d*. Στην εργασία των Maaten & Postma (2009) υιοθετείται η συμβολική παρουσίαση που παρουσιάστηκε παραπάνω και υποτίθεται ότι το σύνολο δεδομένων **X** είναι μηδενικού- μέσου.

Οι τεχνικές για τη μείωση των διαστάσεων υποδιαιρούνται σε κυρτές και μη κυρτές τεχνικές. Οι κυρτές τεχνικές βελτιστοποιούν μια αντικειμενική συνάρτηση που δεν περιέχει κανένα τοπικό βέλτιστο, ενώ οι μη κυρτές τεχνικές βελτιστοποιούν αντικειμενικές συναρτήσεις που περιέχουν τοπικά βέλτιστα. Οι περαιτέρω υποδιαιρέσεις στην ταξινόμηση συζητούνται στα ακόλουθα τμήματα.

3.1. Ανάλυση βασικών συνιστωσών (PCA)

Η ανάλυση βασικών συνιστωσών (PCA) (Pearson, 1901) είναι μια γραμμική τεχνική για τη μείωση των διαστάσεων, πράγμα που σημαίνει ότι εκτελεί τη μείωση των διαστάσεων με την ενσωμάτωση των δεδομένων σε ένα γραμμικό υποχωρίο χαμηλότερης διάστασης. Αν και υπάρχουν διάφορες τεχνικές για να γίνει αυτό, η PCA είναι μακράν η πιο δημοφιλής (χωρίς εποπτεία) γραμμική τεχνική. Ως εκ τούτου, στη σύγκριση, συμπεριλαμβάνεται μόνο η PCA.

Η PCA κατασκευάζει μια λίγων διαστάσεων αναπαράσταση των δεδομένων που περιγράφουν όσο το δυνατόν περισσότερο τη διακύμανση τους. Αυτό γίνεται με την εύρεση μιας γραμμικής βάσης μειωμένης διάστασης για τα δεδομένα, στην οποία το μέγεθος της διακύμανσης στα δεδομένα είναι μέγιστο.

Σε μαθηματικούς όρους, η PCA επιχειρεί να βρει μια γραμμική χαρτογράφηση **M** που μεγιστοποιεί την συνάρτηση κόστους ίχνους **($\mathbf{M}^T$cov(X)M)**, όπου cov (**X**) είναι ο πίνακας συνδιακύμανσης του δείγματος των δεδομένων **X**. Μπορεί να φανεί ότι η γραμμική χαρτογράφηση δημιουργείται από τα βασικά *d* ιδιοδιανύσματα (π.χ. κύρια στοιχεία) του δείγματος πίνακα συνδιακύμανσης των δεδομένων μηδενικού μέσου[1]. Ως εκ τούτου, το PCA λύνει το πρόβλημα ιδιοτιμών

$$\mathbf{cov(X)M} = \lambda\mathbf{M} \qquad (1)$$

Το πρόβλημα ιδιοτιμών λύνεται για τις *d* κύριες ιδιοτιμές λ. Οι αναπαραστάσεις με δεδομένα λίγων διαστάσεων $\mathbf{y}_i$ των σημείων δεδομένων $\mathbf{x}_i$ υπολογίζονται με τη χαρτογράφηση τους στη γραμμική βάση **M**, δηλ. **Y** = **XM**.

Η PCA είναι πανομοιότυπη με την παραδοσιακή τεχνική για πολυδιάστατη κλιμάκωση που ονομάζεται κλασσική κλιμάκωση (Torgerson, 1952). Η εισαγωγή στην κλασσική κλιμάκωση είναι, όπως οι εισροές στις περισσότερες άλλες τεχνικές πολυδιάστατης κλιμάκωσης, ένας Ευκλείδειος πίνακας ζευγών αποστάσεων **D** των οποίων οι καταχωρήσεις d_{ij} αντιπροσωπεύουν

[1] Η PCA μεγιστοποιεί το $\mathbf{M}^T cov(\mathbf{X})\,\mathbf{M}$ ως προς το **M,** υπό τον περιορισμό ότι η 1.2 νόρμα της κάθε στήλης $\mathbf{m}_j$ του **M** είναι 1, π.χ, ότι $\|\mathbf{m}_j\|^2 = 1$. Αυτός ο περιορισμός μπορεί να ενισχυθεί εισάγοντας έναν πολλαπλασιαστή Lagrange λ. Έτσι γίνεται μία χωρίς περιορισμούς μεγιστοποίηση του $\mathbf{m}_j^T cov(\mathbf{X})\mathbf{m}_j + \lambda(1 - \mathbf{m}_j^T \mathbf{m}_j)$. Τα σταθερά σημεία αυτής της ποσότητας βρίσκονται όταν $cov(\mathbf{X})\mathbf{m}_j = \lambda\mathbf{m}_j$

την ευκλείδεια απόσταση μεταξύ των σημείων δεδομένων πολλών διαστάσεων $\mathbf{x}_i$ και $\mathbf{x}_j$. Η κλασσική κλιμάκωση βρίσκει τη γραμμική χαρτογράφηση **M** που ελαχιστοποιεί[2] τη συνάρτηση κόστους στην οποία $\|\mathbf{y}_i - \mathbf{y}_j\|^2$ είναι η τετραγωνική ευκλείδεια απόσταση μεταξύ των σημείων δεδομένων λίγων διαστάσεων $\mathbf{y}_i$ και $\mathbf{y}_j$, και το $\mathbf{y}_i$ περιορίζεται να είναι $\mathbf{x}_i\mathbf{M}$, και $\|\mathbf{m}_j\|^2=1$ για το $\forall j$

$$\varphi(\mathbf{Y}) = \sum_{ij}(d_{ij}^2 - \|\mathbf{y}_i - \mathbf{y}_j\|^2) \qquad (2)$$

Μπορούμε να δειχθεί (Torgerson, 1952) ότι το ελάχιστο αυτής της συνάρτησης δίνεται από την ιδιο-αποσύνθεση του Gram πίνακα $\mathbf{K} = \mathbf{X}\mathbf{X}^\mathbf{T}$ των δεδομένων πολλών διαστάσεων. Οι καταχωρίσεις του Gram πίνακα μπορούν να ληφθούν με διπλό κεντράρισμα του τετραγωνικού Ευκλείδειου πίνακα ζευγών απόστασης, δηλ. με υπολογισμό του

$$k_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{n}\sum_l d_{il}^2 - \frac{1}{n}\sum_l d_{jl}^2 + \frac{1}{n^2}\sum_{lm} d_{lm}^2\right) \qquad (3)$$

Το ελάχιστο της συνάρτησης κόστους στην Εξίσωση (2) μπορεί τώρα να ληφθεί πολλαπλασιάζοντας τα κύρια ιδιοδιανύσματα του διπλά κεντραρισμένου τετραγωνικού ευκλείδειου πίνακα απόστασης (δηλ. Των κύριων ιδιοδιανύσματων του πίνακα Gram) με την τετραγωνική ρίζα των αντίστοιχων ιδιοτιμών τους. Η ομοιότητα της κλασσικής κλιμάκωσης με την PCA οφείλεται σε μία σχέση μεταξύ των ιδιοδιανυσμάτων του πίνακα συνδιακύμανσης και του πίνακα Gram των δεδομένων πολλών διαστάσεων: μπορεί να δειχθεί ότι τα ιδιοδιανύσματα $\mathbf{u}_i$ και $\mathbf{v}_i$ των πινάκων $\mathbf{X}^\mathbf{T}\mathbf{X}$ και $\mathbf{X}\mathbf{X}^\mathbf{T}$ σχετίζονται μέσω $\sqrt{\lambda_i}\mathbf{v}_i = \mathbf{X}\mathbf{u}_i$ (Chatfield & Collins, 1980). Η σύνδεση μεταξύ της PCA και της κλασικής κλιμάκωσης περιγράφεται λεπτομερέστερα στα π.χ., (Williams, 2002).

Η PCA μπορεί επίσης να θεωρηθεί ως ένα μοντέλο λανθάνουσας μεταβλητής που ονομάζεται πιθανοτική PCA (Roweis, 1997). Αυτό το μοντέλο χρησιμοποιεί ένα Γκαουσιανό Μοντέλο (Gaussian) πάνω από τον λανθάνοντα χώρο και ένα μοντέλο γραμμικού- Γκαουσιανού θορύβου. Η πιθανοτική διατύπωση της PCA οδηγεί σε έναν αλγόριθμο EM που μπορεί να είναι

[2] Παρατηρήστε ότι οι αποστάσεις ζευγών μεταξύ των σημείων δεδομένων πολλών διαστάσεων είναι σταθερές σε αυτή τη συνάρτηση κόστους και έτσι δεν εξυπηρετούν στο σκοπό της βελτιστοποίησης. Η βελτιστοποίηση έτσι αποτελεί την μεγιστοποίηση του $\sum_{ij}\|\mathbf{M}\mathbf{x}_\mathbf{i} - \mathbf{M}\mathbf{x}_\mathbf{j}\|^2$, σαν αποτέλεσμα το οποίο μεγιστοποιεί τη διακύμανση στο χώρο λίγων διαστάσεων (έτσι είναι ίδια με την PCA)

υπολογιστικά πιο αποτελεσματικός για δεδομένα αρκετά πολλών διαστάσεων. Χρησιμοποιώντας Γκαουσιανές διεργασίες, η πιθανοτική PCA μπορεί επίσης να επεκταθεί για να κάνει μη γραμμικές χαρτογραφίες μεταξύ του χώρου πολλών διαστάσεων και λίγων διαστάσεων (Lawrence, 2005). Μια άλλη επέκταση της PCA περιλαμβάνει επίσης και τα δευτερεύοντα συστατικά (δηλ. Τα ιδιοδιανύσματα που αντιστοιχούν στις μικρότερες ιδιοτιμές) στη γραμμική χαρτογράφηση, καθώς τα δευτερεύοντα συστατικά μπορεί να έχουν σημασία στις ρυθμίσεις ταξινόμησης (Welling et al, 2004).

Η PCA και η κλασσική κλιμάκωση εφαρμόστηκαν επιτυχώς σε έναν μεγάλο αριθμό τομέων όπως η αναγνώριση προσώπου (Turk & Pentland, 1991), η ταξινόμηση κερμάτων (Huber et al, 2005) και η ανάλυση σεισμικών σειρών (Posadas et al, 1993). Η PCA και η κλασική κλιμάκωση υποφέρουν από δύο κύρια μειονεκτήματα

Πρώτον, στη PCA, το μέγεθος του πίνακα συνδιακύμανσης είναι ανάλογο με τη διάσταση των σημείων δεδομένων. Ως αποτέλεσμα, ο υπολογισμός των ιδιοδιανυσμάτων μπορεί να μην είναι εφικτός για δεδομένα αρκετά πολλών διαστάσεων. Σε σύνολα δεδομένων όπου $n < D$, αυτό το μειονέκτημα μπορεί να ξεπεραστεί με την εκτέλεση κλασικής κλιμάκωσης αντί για PCA, επειδή η κλασική κλιμάκωση κλιμακώνει με τον αριθμό των σημείων δεδομένων αντί με τον αριθμό των διαστάσεων στα δεδομένα. Εναλλακτικά, μπορούν να χρησιμοποιηθούν επαναληπτικές τεχνικές όπως η απλή PCA (Partridge & Calvo, 1997) ή η πιθανοτική PCA (Roweis, 1997).

Δεύτερον, η συνάρτηση κόστους στην εξίσωση (2) αποκαλύπτει ότι η PCA και η κλασσική κλιμάκωση εστιάζουν κυρίως στη διατήρηση μεγάλων αποστάσεων ζευγών d_{ij}^2, αντί να εστιάζουν στη διατήρηση των μικρών αποστάσεων ζευγών, κάτι που είναι πιο σημαντικό.

3.2. Isomap

Η κλασική κλιμάκωση έχει αποδειχθεί επιτυχής σε πολλές εφαρμογές, αλλά πάσχει από το γεγονός ότι αποσκοπεί κυρίως στη διατήρηση των ζευγών ευκλείδειων αποστάσεων και δεν λαμβάνει υπόψη τη κατανομή των γειτονικών σημείων δεδομένων. Εάν τα δεδομένα πολλών διαστάσεων βρίσκονται πάνω ή κοντά σε ένα καμπύλο πολλαπλό χωρίο, όπως στο ελβετικό ρολό σειράς δεδομένων (Tenenbaum et al, 2000), η κλασική κλιμάκωση μπορεί να θεωρήσει δύο

σημεία δεδομένων ως κοντινά σημεία, ενώ η απόσταση τους πάνω στο πολλαπλό χωρίο είναι πολύ μεγαλύτερη από την τυπική απόσταση μεταξύ σημείων. Η Isomap (Tenenbaum et al, 2000) είναι μια τεχνική που επιλύει αυτό το πρόβλημα προσπαθώντας να διατηρήσει γεωδεσιακά ζεύγη (ή καμπυλόγραμμες) αποστάσεις μεταξύ των σημείων δεδομένων. Η γεωδεσιακή απόσταση είναι η απόσταση μεταξύ δύο σημείων που μετρείται πάνω στο πολλαπλό χωρίο.

Στην Isomap (Tenenbaum et al, 2000), οι γεωδεσιακές αποστάσεις μεταξύ των σημείων δεδομένων $\mathbf{x}_i$ (i = 1, 2, ..., n) υπολογίζονται με την κατασκευή ενός γραφήματος γειτνίασης *G*, όπου κάθε σημείο δεδομένων $\mathbf{x}_i$ συνδέεται με τους *k* πλησιέστερους γείτονες $\mathbf{x}_{ij}$ (j = 1, 2, ..., k) στο σύνολο δεδομένων **X**. Η συντομότερη διαδρομή μεταξύ δύο σημείων στο γράφημα αποτελεί μια εκτίμηση της γεωδεσιακής απόστασης μεταξύ αυτών των δύο σημείων και μπορεί εύκολα να υπολογιστεί χρησιμοποιώντας τον αλγόριθμο συντομότερης διαδρομής Dijkstra (1959) ή Floyd (1962). Οι γεωδεσικές αποστάσεις μεταξύ όλων των σημείων δεδομένων στο **X** υπολογίζονται, σχηματίζοντας έτσι έναν πίνακα με ζεύγη γεωδεσιακών αποστάσεων. Οι μειωμένων διαστάσεων αναπαραστάσεις $\mathbf{y}_i$ των σημείων δεδομένων $\mathbf{x}_i$ στο χώρο των λίγων διαστάσεων **Y** υπολογίζονται εφαρμόζοντας την κλασσική κλιμάκωση στον προκύπτοντα πίνακα με τα ζεύγη γεωδεσιακών αποστάσεων.

Μια σημαντική αδυναμία του αλγορίθμου Isomap είναι η τοπολογική του αστάθεια (Balasubramanian & Schwartz, 2002). Το Isomap μπορεί να κατασκευάσει λανθασμένες συνδέσεις στο γράφημα γειτονίασης G. Αυτό το βραχυκύκλωμα (Lee & Verleysen, 2005) μπορεί να επηρεάσει σοβαρά την απόδοση του Isomap. Έχουν προταθεί αρκετές προσεγγίσεις για να ξεπεραστεί το πρόβλημα του βραχυκυκλώματος, π.χ. με την αφαίρεση σημείων δεδομένων με μεγάλες συνολικές ροές στον αλγόριθμο της μικρότερης διαδρομής (Choi & Choi, 2007) ή με την απομάκρυνση των πλησιέστερων γειτόνων που παραβιάζουν την τοπική γραμμικότητα του γραφήματος γειτονίασης (Saxena et al, 2004). Μια δεύτερη αδυναμία είναι ότι το Isomap μπορεί να υποφέρει από «τρύπες» στο πολλαπλό χωρίο. Το πρόβλημα αυτό μπορεί να αντιμετωπιστεί με τη διάσπαση με οπές του πολλαπλού χωρίου (Lee & Verleysen, 2005). Μια τρίτη αδυναμία του Isomap είναι ότι μπορεί να αποτύχει εάν το πολλαπλό χωρίο είναι μη κυρτό (Tenenbaum, 1998). Παρά τις τρεις αυτές αδυναμίες, το Isomap εφαρμόστηκε επιτυχώς σε εργασίες όπως η επιθεώρηση ξύλου (Niskanen & Silven, 2003), η απεικόνιση βιοϊατρικών δεδομένων (Lim et al, 2003) και η εκτίμηση πόζας κεφαλιού (Raytchev et al, 2004).

3.3. Kernel PCA

Η Kernel PCA (KPCA) είναι η αναδιατύπωση της παραδοσιακής γραμμικής PCA σε ένα χώρο πολλών διαστάσεων που κατασκευάζεται χρησιμοποιώντας μια συνάρτηση kernel (Scholkopf et al, 1998). Τα τελευταία χρόνια, η αναδιατύπωση των γραμμικών τεχνικών που χρησιμοποιούν το «κόλπο kernel» οδήγησε στην πρόταση επιτυχών τεχνικών, όπως η παλινδρόμηση της κορυφογραμμής Kernel και οι Μηχανές Υποστήριξης Διανύσματος (Shawe – Taylor & Christianini, 2004). Η Kernel PCA υπολογίζει τα κυριότερα ιδιοδιανύσματα του πίνακα Kernel, αντί αυτών του πίνακα συνδιακύμανσης. Η αναδιατύπωση του PCA στον χώρο Kernel είναι απλή, αφού ένας πίνακας Kernel είναι παρόμοιος με το παραπροϊόν των σημείων δεδομένων στον χώρο πολλών διαστάσεων που κατασκευάζεται χρησιμοποιώντας τη συνάρτηση Kernel. Η εφαρμογή της PCA στον χώρο Kernel παρέχει στην Kernel PCA την ιδιότητα να κατασκευάζει μη γραμμικές χαρτογραφήσεις.

Η Kernel PCA υπολογίζει τον Kernel πίνακα K των σημείων δεδομένων $\mathbf{x}_i$. Οι καταχωρήσεις στον πίνακα Kernel ορίζονται από το

$$k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \qquad (4)$$

όπου το κ είναι μια συνάρτηση Kernel (Shawe-Taylor & Christianini, 2004), η οποία μπορεί να είναι οποιαδήποτε συνάρτηση που δημιουργεί ένα θετικό-ημιορισμένο Kernel πίνακα **K**. Στη συνέχεια, ο kernel πίνακας **K** διπλοκεντράρεται χρησιμοποιώντας την ακόλουθη τροποποίηση των καταχωρήσεων

$$k_{ij} = -\frac{1}{2}\left(k_{ij} - \frac{1}{n}\sum_{l} k_{il} - \frac{1}{n}\sum_{l} k_{jl} + \frac{1}{n^2}\sum_{lm} k_{lm}\right) \qquad (5)$$

Η λειτουργία κεντραρίσματος αντιστοιχεί στην αφαίρεση του μέσου όρου των χαρακτηριστικών στην παραδοσιακή PCA: αφαιρεί το μέσο όρο των δεδομένων στο χώρο χαρακτηριστικών που ορίζεται από τη kernel συνάρτηση κ. Ως αποτέλεσμα, τα δεδομένα στον χώρο των λειτουργιών που ορίζονται από τη kernel συνάρτηση είναι μηδενικού μέσου. Ακολούθως, υπολογίζονται τα κύρια d ιδιοδιανύσματα $\mathbf{v}_i$ του κεντραρισμένου πίνακα kernel. Τα ιδιοδιανύσματα του πίνακα συνδιακύμανσης $\mathbf{a}_i$ (στο χώρο χαρακτηριστικών που κατασκευάζεται από το κ) μπορούν τώρα να

υπολογιστούν, επειδή σχετίζονται με τα ιδιοδιανύσματα του kernel πίνακα $\mathbf{v}_i$ (βλέπε π.χ. (Chatfield & Collins, 1980)) μέσω

$$\mathbf{a}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{v}_i \qquad (6)$$

Προκειμένου να επιτευχθεί η αναπαράσταση των δεδομένων λίγων διαστάσεων, τα δεδομένα προβάλλονται στα ιδιοδιανύσματα του πίνακα συνδιακύμανσης $\mathbf{a}_i$. Το αποτέλεσμα της προβολής (δηλ. Η αναπαράσταση δεδομένων χαμηλής διάστασης **Y**) δίνεται από

$$\mathbf{y}_i = \left\{ \sum_{j=1}^{n} a_1^{(j)} k(\mathbf{x}_j, \mathbf{x}_i), \ldots, \sum_{j=1}^{n} a_d^{(j)} k(\mathbf{x}_j, \mathbf{x}_i \right\} \qquad (7)$$

όπου $a_1^{(j)}$ υποδεικνύει την j-οστή τιμή στο διάνυσμα $\mathbf{a}_1$ και *κ* είναι η kernel συνάρτηση που χρησιμοποιήθηκε επίσης στον υπολογισμό του kernel πίνακα. Δεδομένου ότι η Kernel PCA είναι μια μέθοδος που βασίζεται στην Kernel, η χαρτογράφηση που εκτελείται από τη Kernel PCA βασίζεται στην επιλογή της Kernel συνάρτησης *κ*. Πιθανές επιλογές για τη Kernel συνάρτηση περιλαμβάνουν την γραμμική Kernel (καθιστώντας την Kernel PCA ισοδύναμη με την παραδοσιακή PCA), την πολυωνύμική Kernel και την Gaussian kernel που δίνεται στην Εξίσωση 8 (Shawe-Taylor & Christianini, 2004). Παρατηρήστε ότι όταν χρησιμοποιείται η γραμμική Kernel, ο πίνακας kernel *K* είναι ίσος με τον πίνακα Gram και η διαδικασία που περιγράφεται παραπάνω είναι ίδια με την κλασσική κλιμάκωση (βλέπε 3.1.1).

Μια σημαντική αδυναμία της Kernel PCA είναι ότι το μέγεθος του πίνακα kernel είναι ανάλογο με το τετράγωνο του αριθμού των περιπτώσεων στο σύνολο δεδομένων. Μια προσέγγιση για την επίλυση αυτής της αδυναμίας προτείνεται στην εργασία του Tipping (2000). Επίσης, η Kernel PCA επικεντρώνεται κυρίως στη διατήρηση μεγάλων αποστάσεων ανά ζεύγος (παρόλο που αυτές μετριούνται τώρα στο χώρο των χαρακτηριστικών). Η Kernel PCA εφαρμόστηκε με επιτυχία, π.χ. στην αναγνώριση προσώπων (Kim et al, 2002), στην αναγνώριση ομιλίας (Lima et al, 2004) και στην ανίχνευση καινοτομίας (Hoffman, 2007).

Όπως και η Kernel PCA, το λανθάνον Μεταβλητό Μοντέλο της Γκαουσιανής Διαδικασίας (GPLVM) χρησιμοποιεί επίσης kernel συναρτήσεις για να κατασκευάσει μη γραμμικές

παραλλαγές της (πιθανοτικής) PCA (Lawrence, 2005). Ωστόσο, το GPLVM δεν είναι απλώς το πιθανοτικό αντιστάθμισμα του Kernel PCA: στο GPLVM, η kernel συνάρτηση ορίζεται πάνω από τον λανθάνοντα χώρο λίγων διαστάσεων, ενώ στην Kernel PCA, η συνάρτηση kernel ορίζεται πάνω από τον χώρο δεδομένων πολλών διαστάσεων.

3.4. Χάρτες διάχυσης

Το πλαίσιο χαρτών διάχυσης (DM) (Lafon & Lee, 2006) προέρχεται από το πεδίο των δυναμικών συστημάτων. Οι χάρτες διάχυσης βασίζονται στον ορισμό τυχαίου πεδίου Markov στο γράφημα των δεδομένων. Πραγματοποιώντας το τυχαίο πεδίο για μια σειρά χρονικών βημάτων, λαμβάνεται ένα μέτρο για την εγγύτητα των σημείων δεδομένων. Χρησιμοποιώντας αυτό το μέτρο, καθορίζεται η αποκαλούμενη απόσταση διάχυσης. Στην λίγων διαστάσεων αναπαράσταση των δεδομένων, οι αποστάσεις διάχυσης ανά ζεύγος διατηρούνται όσο το δυνατόν καλύτερα. Η βασική ιδέα πίσω από την απόσταση διάχυσης είναι ότι βασίζεται στην ενσωμάτωση πάνω σε όλες τις διαδρομές του γραφήματος. Αυτό καθιστά την απόσταση διάχυσης πιο ισχυρή σε βραχυκύκλωμα από, π.χ., τη γεωδεσιακή απόσταση που χρησιμοποιείται στο Isomap. Στο πλαίσιο των χαρτών διάχυσης, πρώτα καταρτίζεται ένα γράφημα των δεδομένων. Τα βάρη των άκρων του γραφήματος υπολογίζονται χρησιμοποιώντας τη συνάρτηση Gaussian kernel, οδηγώντας σε έναν πίνακα **W** με καταχωρήσεις

$$w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \qquad (8)$$

όπου το σ δείχνει τη διακύμανση του Gaussian. Ακολούθως, η κανονικοποίηση του πίνακα *W* εκτελείται κατά τέτοιο τρόπο ώστε οι σειρές του να προσθέτουν μέχρι 1. Με αυτό τον τρόπο, σχηματίζεται ένας πίνακας $\mathbf{P}^{(1)}$ με καταχωρήσεις

$$p_{ij}^{(1)} = \frac{w_{ij}}{\sum_k w_{ik}} \qquad (9)$$

Δεδομένου ότι οι χάρτες διάχυσης προέρχονται από τη θεωρία των δυναμικών συστημάτων, ο προκύπτων πίνακας $\mathbf{P}^{(1)}$ θεωρείται πίνακας Markov που ορίζει τον πίνακα πιθανοτήτων μετάβασης προς τα μπροστά μιας δυναμικής διαδικασίας. Επομένως, ο πίνακας $\mathbf{P}^{(1)}$

αντιπροσωπεύει την πιθανότητα μετάβασης από ένα σημείο δεδομένων σε ένα άλλο σημείο δεδομένων σε ένα απλό χρονικό βήμα. Επομένως, ο πίνακας πιθανοτήτων πρόβλεψης για t χρονικά βήματα $\mathbf{P}^{(t)}$ δίνεται από το $(\mathbf{P}^{(1)})^t$. Χρησιμοποιώντας τις πιθανότητες τυχαίου μονοπατιού προς τα μπροστά $p_{ij}^{(t)}$, η απόσταση διάχυσης ορίζεται από το

$$D^{(t)}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_k \frac{\left(p_{ik}^{(t)} - p_{jk}^{(t)}\right)^2}{\psi(\mathbf{x}_k)^{(0)}}} \qquad (10)$$

Στην εξίσωση, $\psi(x_i)^{(0)}$ είναι ένας όρος που αποδίδει περισσότερο βάρος σε μέρη του γραφήματος με υψηλή πυκνότητα. Ορίζεται από το $\psi(\mathbf{x}_i)^{(0)} = \frac{m_i}{\sum_j m_j}$ όπου m_i είναι ο βαθμός του κόμβου $\mathbf{x}_i$ που ορίζεται από την εξίσωση, $m_i = \sum_j p_{ij}$.Από την εξίσωση (10) μπορεί να παρατηρηθεί ότι ζεύγη σημείων δεδομένων με μεγάλη πιθανότητα μετάβασης προς τα εμπρός έχουν μικρή απόσταση διάχυσης. Δεδομένου ότι η απόσταση διάχυσης βασίζεται στην ενσωμάτωση πάνω σε όλες τις διαδρομές του γραφήματος, είναι πιο ισχυρή σε βραχυκύκλωμα από τη γεωδεσιακή απόσταση που χρησιμοποιείται στο Isomap. Στην λίγων διαστάσεων αναπαράσταση των δεδομένων **Y**, οι χάρτες διάχυσης επιχειρούν να διατηρήσουν τις αποστάσεις διάχυσης. Χρησιμοποιώντας τη φασματική θεωρία για την τυχαία διαδρομή, έχει αποδειχθεί (βλ., Π.χ., (Lafon & Lee, 2006)) ότι η χαμηλής διάστασης αναπαράσταση **Y** που διατηρεί τις αποστάσεις διάχυσης $D^{(t)}(x_i, x_j)$ όσο το δυνατόν καλύτερα (υπό ένα κριτήριο τετραγωνικού σφάλματος) σχηματίζεται από τα *d* κυριότερα ιδιοδιανύσματα του προβλήματος ιδιοτιμών.

$$\mathbf{P}^{(t)}\mathbf{v} = \lambda\mathbf{v} \qquad (11)$$

Επειδή το γράφημα είναι πλήρως συνδεδεμένο, η μεγαλύτερη ιδιοτιμή είναι ασήμαντη (viz. $\lambda_1 = 1$), και το ιδιοδιάνυσμα $\mathbf{v}_1$ απορρίπτεται. Η λίγων διαστάσεων αναπαράσταση **Y** δίνεται από τα επόμενα *d* κύρια ιδιοδιανύσματα. Στην αναπαράσταση λίγων διαστάσεων, τα ιδιοδιανύσματα κανονικοποιούνται από τις αντίστοιχες ιδιοτιμές τους. Ως εκ τούτου, η αναπαράσταση δεδομένων λίγων διαστάσεων δίνεται από

$$Y = \{\lambda_2\mathbf{v}_2, \lambda_3\mathbf{v}_3, \ldots, \lambda_{d+1}\mathbf{v}_{d+1}\} \qquad (12)$$

Οι χάρτες διάχυσης έχουν εφαρμοστεί με επιτυχία, για παράδειγμα, στην αντιστοίχιση σχήματος (Rajpoot et al, 2007) και στην ανάλυση γονιδιακής έκφρασης (Xu et al, 2007).

3.5. LLE

Η τοπική γραμμική ενσωμάτωση (LLE) (Roweis & Saul, 2000) είναι μια τεχνική που είναι παρόμοια με το Isomap (και το MVU) στο ότι κατασκευάζει μια γραφική αναπαράσταση των σημείων δεδομένων. Σε αντίθεση με την Isomap, επιχειρεί να διατηρήσει αποκλειστικά τις τοπικές ιδιότητες των δεδομένων. Ως αποτέλεσμα, το LLE είναι λιγότερο ευαίσθητο σε βραχυκύκλωμα από το Isomap, επειδή επηρεάζεται μόνο ένας μικρός αριθμός τοπικών ιδιοτήτων σε περίπτωση βραχυκυκλώματος. Επιπλέον, η διατήρηση των τοπικών ιδιοτήτων επιτρέπει την επιτυχή ενσωμάτωση μη κυρτών πολλαπλών χωρίων. Στο LLE, οι τοπικές ιδιότητες δεδομένων των πολλαπλών χωρίων κατασκευάζονται με την εγγραφή των σημείων δεδομένων πολλών διαστάσεων ως γραμμικό συνδυασμό των πλησιέστερων γειτόνων τους. Στην λίγων διαστάσεων αναπαράσταση των δεδομένων, η LLE προσπαθεί να διατηρήσει τα βάρη ανασυγκρότησης στους γραμμικούς συνδυασμούς όσο το δυνατόν καλύτερα.

Το LLE περιγράφει τις τοπικές ιδιότητες των πολλαπλών χωρίων γύρω από ένα σημείο δεδομένων $\mathbf{x}_i$ γράφοντας το σημείο δεδομένων σαν ένα γραμμικό συνδυασμό $\mathbf{w}_i$ (τα αποκαλούμενα βάρη ανασυγκρότησης) των k πλησιέστερων γειτόνων $\mathbf{x}_{ij}$. Ως εκ τούτου, το LLE ταιριάζει ένα υπερεπίπεδο μέσω του σημείου δεδομένων $\mathbf{x}_i$ και των πλησιέστερων γειτόνων του, υποθέτοντας έτσι ότι το πολλαπλό χωρίο είναι τοπικά γραμμικό. Η παραδοχή τοπικής γραμμικότητας υποδηλώνει ότι τα βάρη ανασυγκρότησης $\mathbf{w}_i$ των σημείων δεδομένων $\mathbf{x}_i$ είναι αμετάβλητα για μεταφορά, περιστροφή και ανακατάταξη. Λόγω της μη μεταβλητότητας σε αυτούς τους μετασχηματισμούς, οποιαδήποτε γραμμική χαρτογράφηση του υπερεπιπέδου σε ένα χώρο λιγότερων διαστάσεων διατηρεί τα βάρη ανασυγκρότησης στο χώρο των λιγότερων διαστάσεων. Με άλλα λόγια, αν η παράσταση δεδομένων λίγων διαστάσεων διατηρεί την τοπική γεωμετρία του πολλαπλού χωρίου, τα βάρη ανακατασκευής $\mathbf{w}_i$ που ανασυγκροτούν το σημείο δεδομένων $\mathbf{x}_i$ από τους γείτονές του στην αναπαράσταση δεδομένων πολλών διαστάσεων ανασυνθέτουν επίσης το σημείο δεδομένων $\mathbf{y}_i$ από τους γείτονές του στις λίγες διαστάσεις. Κατά

συνέπεια, η εύρεση της *d*-διαστάσεων αντιπροσώπευσης δεδομένων **Y** ισοδυναμεί με ελαχιστοποίηση της συνάρτησης κόστους

$$\varphi(\mathbf{Y}) = \sum_i \left\| \mathbf{y}_i - \sum_{j=1}^{k} w_{ij}\mathbf{y}_{ij} \right\|^2 \text{ υπό την συνθήκη } \left\|\mathbf{y}^{(k)}\right\|^2 = 1 \ \forall k \qquad (13)$$

όπου $\mathbf{y}^{(k)}$ αντιπροσωπεύει την κ-οστή στήλη του πίνακα λύσης **Y**. Ο περιορισμός στη συνδιακύμανση των στηλών του **Y** απαιτείται για να αποκλειστεί η ασήμαντη λύση **Y = 0**. Οι Roweis και Saul (2000) έδειξαν[3] ότι οι συντεταγμένες των αναπαραστάσεων λίγων διαστάσεων $\mathbf{y}_i$ που μειώνουν την συνάρτηση κόστους βρίσκονται με τον υπολογισμό των ιδιοδιανυσμάτων που αντιστοιχούν στην μικρότερη μη μηδενική ιδιοτιμή του γινομένου $(\mathbf{I} - \mathbf{W})^T (\mathbf{I} - \mathbf{W})$, όπου **W** είναι ένας πολύ αραιός n × n πίνακας του οποίου οι καταχωρίσεις έχουν οριστεί σε 0 αν τα *i* και *j* δεν είναι συνδεδεμένα στο γράφημα της γειτονίασης και ίσα με το αντίστοιχο βάρος ανοικοδόμησης σε διαφορετική περίπτωση. Σε αυτόν τον τύπο, το **I** είναι μοναδιαίος πίνακας n × n.

Η δημοτικότητα του LLE έχει οδηγήσει στην πρόταση γραμμικών παραλλαγών του αλγορίθμου (He et al, 2005), και σε επιτυχημένες εφαρμογές, π.χ. στην υπέρ ανάλυση (Chang et al, 2004) και στον εντοπισμό της πηγής ήχου (Duraiswami & Raykar, 2005). Ωστόσο, υπάρχουν επίσης πειραματικές μελέτες που αναφέρουν χαμηλές επιδόσεις του LLE. Στην εργασία των Lim et al, (2003), αναφέρθηκε ότι η LLE αποτυγχάνει στην απεικόνιση ακόμη και απλών συνθετικών βιοϊατρικών συνόλων δεδομένων. Στην εργασία των Jenkins & Mataric (2002), αναφέρεται ότι το LLE εκτελεί χειρότερα από το Isomap στην παραγωγή ενεργειών αντιληπτικής κινητικότητας. Μια πιθανή εξήγηση έγκειται στις δυσκολίες που αντιμετωπίζει το LLE όταν αντιμετωπίζει πολλαπλά χωρία που περιέχουν οπές (Roweis & Saul, 2000). Επιπρόσθετα, το LLE τείνει να καταρρέει μεγάλα τμήματα των δεδομένων που είναι πολύ κοντά μεταξύ τους στο χώρο των

[3] $\varphi(\mathbf{Y}) = (\mathbf{Y} - \mathbf{WY})^2 = \mathbf{Y}^T(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})\mathbf{Y}$ είναι η συνάρτηση που πρέπει να ελαχιστοποιηθεί. Έτσι τα ιδιοδιανύσματα του $(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$ που αντιστοιχούν στις μικρότερες μη μηδενικές ιδιοτιμές σχηματίζουν τη λύση που μεγιστοποιεί το $\varphi(\mathbf{Y})$

λίγων διαστάσεων, επειδή ο περιορισμός της συνδιακύμανσης στη λύση είναι πολύ απλός (van der Maaten et al, 2008). Επίσης, ο περιορισμός της συνδιακύμανσης μπορεί να προκαλέσει ανεπιθύμητες αναπροσαρμογές των δεδομένων των πολλαπλών χωρίων στην ενσωμάτωση (Goldberg et al, 2008).

3.6. Laplacian Eigenmaps

Παρόμοια με το LLE, οι Laplacian Eigenmaps βρίσκουν μια λίγων διαστάσεων αναπαράσταση δεδομένων διατηρώντας τις τοπικές ιδιότητες των πολλαπλών χωρίων (Belkin & Niyogi, 2002). Στα Laplacian Eigenmaps, οι τοπικές ιδιότητες βασίζονται στις αποστάσεις ζευγών μεταξύ των κοντινών γειτόνων. Οι Laplacian Eigenmaps υπολογίζουν μια λίγων διαστάσεων αναπαράσταση των δεδομένων στην οποία ελαχιστοποιούνται οι αποστάσεις μεταξύ ενός σημείου δεδομένων και των k πλησιέστερων γειτόνων του. Αυτό γίνεται με σταθμισμένο τρόπο, δηλ. Η απόσταση στην αναπαράσταση δεδομένων λίγων διαστάσεων μεταξύ ενός σημείου δεδομένων και του πρώτου πλησιέστερου γείτονα συμβάλλει περισσότερο στην συνάρτηση κόστους από την απόσταση μεταξύ του σημείου δεδομένων και του δεύτερου πλησιέστερου γείτονά του. Χρησιμοποιώντας τη θεωρία των φασματικών γραφημάτων, η ελαχιστοποίηση της συνάρτησης κόστους ορίζεται ως ένα πρόβλημα ιδιοτιμών.

Ο αλγόριθμος Laplacian Eigenmaps κατασκευάζει πρώτα ένα γράφημα γειτονίασης *G* στο οποίο κάθε σημείο δεδομένων $\mathbf{x}_i$ συνδέεται με τους πλησιέστερους *k* γείτονές του. Για όλα τα σημεία $\mathbf{x}_i$ και $\mathbf{x}_j$ στο γράφημα *G* που συνδέονται με μια ακμή, υπολογίζεται το βάρος της ακμής με τη χρήση της συνάρτησης Gaussian kernel (βλ. Εξίσωση 8), οδηγώντας σε έναν αραιό πίνακα συγγενείας **W**. Στον υπολογισμό της αναπαράστασης των $\mathbf{y}_i$ λίγων διαστάσεων, η συνάρτηση κόστους που ελαχιστοποιείται δίνεται από

$$\varphi(\mathbf{Y}) = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{ij} \qquad (14)$$

Στη συνάρτηση κόστους, τα μεγάλα βάρη w_{ij} αντιστοιχούν σε μικρές αποστάσεις μεταξύ των σημείων δεδομένων πολλών διαστάσεων $\mathbf{x}_i$ και $\mathbf{x}_j$. Έτσι, η διαφορά μεταξύ των

αναπαραστάσεων τους με λίγες διαστάσεις $\mathbf{y}_i$ και $\mathbf{y}_j$ συμβάλλει σημαντικά στη συνάρτηση του κόστους. Κατά συνέπεια, τα κοντινά σημεία στον χώρο πολλών διαστάσεων τοποθετούνται όσο το δυνατόν πλησιέστερα στην αναπαράσταση λίγων διαστάσεων.

Ο υπολογισμός του πίνακα βαθμίδων **M** και του λαπλασιανού γραφήματος **L** του γραφήματος **W** επιτρέπει τη διατύπωση του προβλήματος ελαχιστοποίησης στην Εξίσωση 14 ως πρόβλημα ιδιοτιμών (Anderson & Morley, 1985). Ο πίνακας βαθμίδων **M** της **W** είναι ένας διαγώνιος πίνακας, του οποίου οι εγγραφές είναι τα αθροίσματα σειρών του **W** (δηλ.$m_{ii} = \sum_j w_{ij}$). Το λαπλασιανό γράφημα **L** υπολογίζεται από **L = M - W**. Μπορεί να φανεί ότι τα ακόλουθα είναι διαθέσιμα[4]

$$\varphi(\mathbf{Y}) = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{ij} = 2\mathbf{Y}^T\mathbf{L}\mathbf{Y} \qquad (15)$$

Συνεπώς, η ελαχιστοποίηση του φ (**Y**) είναι ανάλογη με την ελαχιστοποίηση του $\mathbf{Y}^T\mathbf{LY}$ που υπόκειται σε $\mathbf{Y}^T\mathbf{MY} = \mathbf{I_n}$, έναν περιορισμό συνδιακύμανσης που είναι παρόμοιος με αυτόν της LLE. Επομένως, η αναπαράσταση δεδομένων λίγων διαστάσεων **Y** μπορεί να βρεθεί με επίλυση του γενικευμένου προβλήματος ιδιοτιμών

$$\mathbf{Lv} = \lambda\mathbf{Mv} \qquad (16)$$

για τις *d* μικρότερες μη μηδενικές ιδιοτιμές. Τα *d* ιδιοδιανύσματα $\mathbf{v}_i$ που αντιστοιχούν στις μικρότερες μη-μηδενικές ιδιοτιμές επιτελούν τη γραφική απεικόνιση **Y** λίγων διαστάσεων. Οι Laplacian Eigenmaps πάσχουν από πολλές από τις ίδιες αδυναμίες με το LLE, όπως η παρουσία μιας ασήμαντης λύσης η οποία εμποδίζεται να επιλεγεί από έναν περιορισμό συνδιακύμανσης που μπορεί εύκολα να εξαπατηθεί. Παρά τις αδυναμίες αυτές, οι Laplacian Eigenmaps (και οι παραλλαγές τους) εφαρμόστηκαν με επιτυχία, π.χ. στην αναγνώριση προσώπων (He et al, 2005) και στην ανάλυση των δεδομένων fMRI (Brun et al, 2003). Επιπλέον, οι παραλλαγές των Laplacian Eigenmaps μπορούν να εφαρμοστούν σε επιβλεπόμενα ή ημι-εποπτευόμενα

[4] $\varphi(\mathbf{Y}) = \sum_{ij}\|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{ij} = \sum_{ij}\left(\|\mathbf{y}_i\|^2 + \|\mathbf{y}_j\|^2 - 2\mathbf{y}_i\mathbf{y}_j^T\right)w_{ij} = \sum_i\|\mathbf{y}_i\|^2 m_{ii} + \sum_j\|\mathbf{y}_j\|^2 m_{jj} - 2\sum_{ij}\mathbf{y}_i\mathbf{y}_j^T w_{ij} = 2\mathbf{Y}^T\mathbf{MY} - 2\mathbf{Y}^T\mathbf{WY} = 2\mathbf{Y}^T\mathbf{LY}$

μαθησιακά προβλήματα (Costa & Hero, 2005). Μια γραμμική παραλλαγή των Laplacian Eigenmaps παρουσιάζεται στην εργασία των He & Niyogi (2004). Στη φασματική συσσώρευση, η συσσώρευση πραγματοποιείται βάσει του προσήμου των συντεταγμένων που λαμβάνονται από τη μέθοδο Laplacian Eigenmaps (Ng et al, 2001).

3.7. Χαρτογράφηση Sammon

Προηγουμένως συζητήθηκε η κλασική κλιμάκωση, μια κυρτή τεχνική για την πολυδιάστατη κλιμάκωση και σημειώθηκε ότι η κύρια αδυναμία αυτής της τεχνικής είναι ότι επικεντρώνεται κυρίως στη διατήρηση μεγάλων αποστάσεων ανά ζεύγη και όχι στη συγκράτηση των μικρών αποστάσεων ανά ζεύγη, οι οποίες είναι πολύ πιο σημαντικές για τη γεωμετρία των δεδομένων. Έχουν προταθεί αρκετές πολυδιάστατες παραλλαγές κλιμάκωσης που στοχεύουν στην αντιμετώπιση αυτής της αδυναμίας (Agrafiotis, 2003). Σε αυτήν την υποενότητα, συζητείται μια τέτοια παραλλαγή MDS που ονομάζεται Sammon χαρτογράφηση (Sammon, 1969).

Η χαρτογράφηση Sammon προσαρμόζει τη συνάρτηση κόστους κλασικής κλιμάκωσης (βλ. Εξίσωση 2) σταθμίζοντας τη συνεισφορά κάθε ζεύγους (*i, j)* στη συνάρτηση κόστους από το αντίστροφο της απόστασης ανά ζεύγος d_{ij} στον πολλών διαστάσεων χώρο. Με αυτό τον τρόπο, η συνάρτηση κόστους αντιστοιχεί περίπου το ίδιο βάρος στο να διατηρεί κάθε μια από τις αποστάσεις ζευγών, διατηρώντας έτσι την τοπική δομή των δεδομένων καλύτερα από την κλασσική κλιμάκωση. Μαθηματικώς, η συνάρτηση κόστους Sammon δίνεται από

$$\varphi(\mathbf{Y}) = \frac{1}{\sum_{ij} d_{ij}} \sum_{i \neq j} \frac{\left(d_{ij} - \|\mathbf{y}_i - \mathbf{y}_j\|\right)^2}{d_{ij}} \qquad (20)$$

όπου το d_{ij} αντιπροσωπεύει την ανά ζεύγη ευκλείδια απόσταση μεταξύ των σημείων δεδομένων πολλών διαστάσεων $\mathbf{x}_i$ και $\mathbf{x}_j$ και η σταθερά μπροστά προστίθεται προκειμένου να απλοποιηθεί η κλίση της συνάρτησης κόστους. Η ελαχιστοποίηση της συνάρτησης κόστους Sammon εκτελείται γενικά χρησιμοποιώντας μια ψευδό-Νευτώνια μέθοδο (Cox & Cox, 1994). Η χαρτογράφηση Sammon χρησιμοποιείται κυρίως για σκοπούς απεικόνισης (Martin-Merino & Munoz, 2004).

Η κύρια αδυναμία της χαρτογράφησης Sammon είναι ότι αποδίδει ένα πολύ μεγαλύτερο βάρος στη διατήρηση μιας απόστασης, ας πούμε 10^{-5}, παρά στη διατήρηση μιας απόστασης, ας πούμε, 10^{-4}. Επιτυχημένες εφαρμογές χαρτογράφησης Sammon έχουν αναφερθεί, σε π.χ., γονιδιακά δεδομένα (Ekins et al, 2006) και γεωχωρικά δεδομένα (Takatsuka, 2001).

3.8. LLC

Ο Τοπικός Γραμμικός Συντονισμός (LLC) (Teh & Roweis, 2002) υπολογίζει έναν αριθμό τοπικά γραμμικών μοντέλων και στη συνέχεια εκτελεί μια συνολική ευθυγράμμιση των γραμμικών μοντέλων. Αυτή η διαδικασία αποτελείται από δύο στάδια: (1) υπολογισμό ενός μείγματος τοπικών γραμμικών μοντέλων πάνω στα δεδομένα με τη βοήθεια ενός αλγόριθμου EM και (2) ευθυγράμμιση των τοπικών γραμμικών μοντέλων ώστε να ληφθεί η αναπαράσταση δεδομένων λίγων διαστάσεων χρησιμοποιώντας μια παραλλαγή LLE.

Η LLC κατασκευάζει πρώτα ένα μείγμα *m* αναλυτών παράγοντα (MoFA)[5] χρησιμοποιώντας τον αλγόριθμο EM (Dempster et al, 1977). Εναλλακτικά, θα μπορούσε να χρησιμοποιηθεί ένα μείγμα πιθανολογικού μοντέλου PCA (MoPPCA) (Tipping & Bishop, 1999). Τα τοπικά γραμμικά μοντέλα στο μείγμα χρησιμοποιούνται για την κατασκευή των m αναπαραστάσεων $\mathbf{z}_{ij}$ και των αντίστοιχων βαρών τους r_{ij} (όπου j ϵ {1,...,m}) για κάθε σημείο δεδομένων $\mathbf{x}_{i.}$ Τα βάρη r_{ij} περιγράφουν σε ποιο βαθμό το σημείο δεδομένων $\mathbf{x}_i$ αντιστοιχεί στο μοντέλο j; ικανοποιούν το $\sum_j r_{ij} = 1$. Χρησιμοποιώντας τα τοπικά μοντέλα και τα αντίστοιχα βάρη, υπολογίζονται οι αντιπροσωπεύσεις δεδομένων που υπολογίζονται με βάση τα βάρη $\mathbf{u}_{ij} = r_{ij}\mathbf{z}_{ij}$. Οι κατανομές δεδομένων $\mathbf{u}_{ij}$ που σταθμίστηκαν με τα βάρη αποθηκεύονται σε έναν $n \times mD$ πίνακα μπλοκ **U**. Η ευθυγράμμιση των τοπικών μοντέλων γίνεται με βάση το **U** και έναν πίνακα **M** που δίνεται από το $\mathbf{M} = (\mathbf{I}_n - \mathbf{W})^T (\mathbf{I}_n - \mathbf{W})$. Εδώ, ο πίνακας **W** περιέχει τα βάρη ανακατασκευής που υπολογίζονται από το LLE (βλέπε 3.2.1), και το $\mathbf{I}_n$ δηλώνει τον n × n μοναδιαίο πίνακα. Η LLC ευθυγραμμίζει τα τοπικά μοντέλα με την επίλυση του γενικευμένου προβλήματος ιδιοτιμών.

$$\mathbf{Av} = \lambda\mathbf{Bv} \qquad (21)$$

[5] Παρατηρήστε ότι το μίγμα των αναλυτών παραγόντων (και το μίγμα των πιθανοτικών PCA μοντέλων) είναι ένα μίγμα μοντέλου Γκαουσιανών με περιορισμό στην συνδιακύμανση των Γκαουσιανών

για τις *d* μικρότερες μη μηδενικές ιδιοτιμές[6]. Στην εξίσωση, το **A** είναι το inproduct των $\mathbf{M}^T$ **U** και το **B** είναι το inproduct του **U**. Τα *d* ιδιοδιανύσματα $\mathbf{v}_i$ σχηματίζουν έναν πίνακα **L** ο οποίος ορίζει μια γραμμική χαρτογράφηση από την αναπαράσταση δεδομένων **U** με βάση τα βάρη στην υποκείμενη αναπαράσταση δεδομένων **Y** λίγων διαστάσεων. Η αναπαράσταση των δεδομένων λίγων διαστάσεων επιτυγχάνεται με τον υπολογισμό του **Y** = **UL**.

Η κύρια αδυναμία της LLC είναι ότι η τοποθέτηση του μείγματος των αναλυτών παραγόντων είναι ευαίσθητη στην παρουσία τοπικών μέγιστων στη συνάρτηση (log-likelihood) log-πιθανότητας. Η LLC έχει εφαρμοστεί με επιτυχία σε εικόνες προσώπων ενός ατόμου με μεταβλητή στάση και έκφραση και σε χειρόγραφα ψηφία (Teh & Roweis, 2002).

3.9. Χαρτογράφηση πολλαπλών χωρίων

Παρόμοια με την LLC, η χαρτογράφηση κατασκευών πολλαπλών χωρίων δημιουργεί μια λίγων διαστάσεων αναπαράσταση δεδομένων, ευθυγραμμίζοντας ένα μοντέλο MoFA ή MoPPCA (Brand, 2002). Σε αντίθεση με την LLC, η χαρτογράφηση πολλαπλών χωρίων δεν ελαχιστοποιεί μια συνάρτηση κόστους που αντιστοιχεί σε μια άλλη τεχνική μείωσης των διαστάσεων (όπως η συνάρτηση κόστους LLE). Η χαρτογράφηση πολλαπλών χωρίων ελαχιστοποιεί μια κυρτή συνάρτηση κόστους που μετρά το μέγεθος της διαφοράς μεταξύ των γραμμικών μοντέλων στις καθολικές συντεταγμένες των σημείων δεδομένων. Η ελαχιστοποίηση αυτής της συνάρτησης κόστους μπορεί να εκτελεστεί με την επίλυση ενός γενικευμένου προβλήματος ιδιοτιμών.

Η χαρτογράφηση πολλαπλών χωρίων εκτελεί πρώτα τον αλγόριθμο EM για την εκμάθηση ενός μείγματος αναλυτών παραγόντων, προκειμένου να λάβει *m* αναπαραστάσεις δεδομένων λίγων διαστάσεων $\mathbf{z}_{ij}$ και αντίστοιχα βάρη r_{ij} (όπου j ϵ {1,...,m}) για όλα τα σημεία δεδομένων $\mathbf{x}_i$. Η χαρτογράφηση πολλαπλής βρίσκει μια γραμμική χαρτογράφηση **M** από τις αναπαραστάσεις δεδομένων $\mathbf{z}_{ij}$ προς τις καθολικές συντεταγμένες $\mathbf{y}_i$ που ελαχιστοποιεί τη συνάρτηση κόστους

$$\varphi(\mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}\|\mathbf{y}_i - \mathbf{y}_{ij}\|^2 \qquad (22)$$

[6] Η εξαγωγή αυτού του προβλήματος ιδιοτιμών μπορεί να βρεθει στην εργασία των Teh & Roweis (2002)

Όπου $\mathbf{y}_i = \sum_{k=1}^{m} r_{ik}\mathbf{y}_{ik}$ $και$ $\mathbf{y}_{ij} = \mathbf{z}_{ij}\mathbf{M}$.Η διαίσθηση πίσω από τη συνάρτηση κόστους είναι ότι κάθε φορά που υπάρχουν δύο γραμμικά μοντέλα στα οποία ένα σημείο δεδομένων έχει μεγάλη ευθύνη, αυτά τα γραμμικά μοντέλα πρέπει να συμφωνήσουν στην τελική συντεταγμένη του σημείου δεδομένων. Η συνάρτηση κόστους μπορεί να ξαναγραφεί στη μορφή:

$$\varphi(\mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} r_{ij}r_{ik}\left\|\mathbf{y}_{ij} - \mathbf{y}_{ik}\right\|^2 \qquad (23)$$

που επιτρέπει τη συνάρτηση κόστους να ξαναγραφεί με τη μορφή ενός πηλίκου Rayleigh To πηλίκο Rayleigh μπορεί να κατασκευαστεί από τον ορισμό ενός μπλοκ-διαγώνιου πίνακα D με *m* μπλοκ από :

$$\mathbf{D} = \begin{pmatrix} \mathbf{D_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{D_m} \end{pmatrix} \qquad (24)$$

όπου $\mathbf{D}_j$ είναι το άθροισμα των ζυγισμένων συνδιακυμάνσεων των παραστάσεων δεδομένων $\mathbf{z}_{ij}$. Ως εκ τούτου, $\mathbf{D}_j$ δίνεται από :

$$\mathbf{D}_j = \sum_{i=1}^{n} r_{ij} cov([\mathbf{Z}_j \;\; \mathbf{1}]) \qquad (25)$$

Στην Εξίσωση 25, η 1-στήλη προστίθεται στην αναπαράσταση δεδομένων $\mathbf{Z}_j$ προκειμένου να διευκολυνθούν οι μεταφράσεις στην κατασκευή του $\mathbf{y}_i$ από τις αναπαραστάσεις δεδομένων $\mathbf{z}_{ij}$. Χρησιμοποιώντας τον ορισμό του πίνακα **D** και του διαγωνίου πίνακα $n \times mD$ **U** με καταχωρήσεις $\mathbf{u}_{ij} = r_{ij}\,[\mathbf{z}_{ij}\;\mathbf{1}]$, η συνάρτηση κόστους χαρτογράφησης πολλαπλών χωρίων μπορεί να ξαναγραφεί ως

$$\varphi(\mathbf{Y}) = \mathbf{L}^T(\mathbf{D} - \mathbf{U}^T\mathbf{U})\mathbf{L} \qquad (26)$$

όπου το **L** αντιπροσωπεύει τη γραμμική χαρτογράφηση επί του πίνακα **Z** που μπορεί να χρησιμοποιηθεί για τον υπολογισμό της τελικής αναπαράστασης δεδομένων λίγων διαστάσεων

Y. Η γραμμική χαρτογράφηση **L** μπορεί έτσι να υπολογιστεί με επίλυση του γενικευμένου προβλήματος ιδιοτιμών.

$$(\mathbf{D} - \mathbf{U}^T\mathbf{U})\mathbf{v} = \lambda\mathbf{U}^T\mathbf{U}\mathbf{v} \qquad (27)$$

για τις *d* μικρότερες μη μηδενικές ιδιοτιμές. Τα *d* ιδιοδιανύσματα $\mathbf{v}_i$ σχηματίζουν τις στήλες του γραμμικού συνδυασμού **L** από [**U 1**] έως **Y**.

4. Η τεχνική t-SNE

Παρουσιάζεται μια νέα τεχνική που ονομάζεται "t-SNE", η οποία απεικονίζει τα δεδομένα πολλών διαστάσεων δίδοντας σε κάθε σημείο δεδομένων μια θέση σε ένα χάρτη δύο ή τριών διαστάσεων. Η τεχνική είναι μια παραλλαγή της Στοχαστικής Ενσωμάτωσης Γειτόνων (Hinton & Roweis, 2002) που είναι πολύ πιο εύκολη στη βελτιστοποίηση και παράγει σημαντικά βελτιωμένες οπτικές αναπαραστάσεις μειώνοντας την τάση να συγκεντρώνονται σημεία στο κέντρο του χάρτη. Το t-SNE είναι καλύτερο από τις υπάρχουσες τεχνικές στη δημιουργία ενός ενιαίου χάρτη που αποκαλύπτει τη δομή σε πολλές διαφορετικές κλίμακες. Αυτό είναι ιδιαίτερα σημαντικό για τα δεδομένα πολλών διαστάσεων που βρίσκονται σε πολλά διαφορετικά, αλλά σχετικά, λίγων διαστάσεων πολλαπλά χωρία, όπως εικόνες αντικειμένων από πολλαπλές κλάσεις που φαίνονται από πολλαπλές οπτικές γωνίες

4.1. Στοχαστική Ενσωμάτωση Γειτόνων

Η Στοχαστική Ενσωμάτωση Γειτόνων (SNE) ξεκινά με τη μετατροπή των πολλών διαστάσεων Ευκλείδειων αποστάσεων μεταξύ σημείων δεδομένων σε υπό όρους πιθανότητες που αντιπροσωπεύουν ομοιότητες[7]. Η ομοιότητα του σημείου δεδομένων x_j με το σημείο δεδομένων x_i είναι η πιθανότητα υπό όρους, $p_{j|i}$, ότι το x_i θα επιλέξει το x_j ως γείτονα του εάν οι γείτονες συλλέγονταν ανάλογα με την πυκνότητα πιθανότητας τους υπό μία Gaussian κατανομή με κέντρο το x_i.

Για τα κοντινά σημεία δεδομένων, το $p_{j|i}$ είναι σχετικά υψηλό, ενώ για τα ευρέως διαχωρισμένα σημεία δεδομένων, το $p_{j|i}$ θα είναι σχεδόν απειροελάχιστο (για λογικές τιμές της διακύμανσης της Gaussian, σ_i). Μαθηματικά, η υπό όρους πιθανότητα $p_{j|i}$ δίνεται από

$$p_{f|i} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2)} \qquad (1)$$

[7] Η SNE μπορεί επίσης να εφαρμοστεί σε ομάδες δεδομένων που αποτελούνται από ομοιότητες ζευγών μεταξύ αντικειμένων περισσότερο από αντιπροσωπεύσεις διανυσμάτων πολλών διαστάσεων για κάθε αντικείμενο, αν αυτές οι ομοιότητες μπορούν να ερμηνευθούν σαν υπό συνθήκη πιθανότητες. Για παράδειγμα, τα δεδομένα που συνδέονται με ανθρώπινη λέξη αποτελούνται από την πιθανότητα της παραγωγής κάθε πιθανής λέξης ως απάντηση σε δεδομένη λέξη, σαν παράδειγμα για το οποίο είναι ήδη στην μορφή που απαιτείται από την SNE

όπου το σ_i είναι η διακύμανση της Gaussian κατανομής που επικεντρώνεται στο σημείο δεδομένων x_i. Η μέθοδος για τον προσδιορισμό της τιμής του σ_i παρουσιάζεται αργότερα. Επειδή υπάρχει ενδιαφέρον μόνο για να μοντελοποιηθούν οι ομοιότητες ζευγών, η τιμή του $p_{i|i}$ ρυθμίστηκε στο μηδέν. Για τα λίγων διαστάσεων ομόλογα y_i και y_j των πολλών διαστάσεων σημείων δεδομένων x_i και x_j, είναι δυνατόν να υπολογιστεί μια παρόμοια υπό όρους πιθανότητα, η οποία υποδηλώνεται με $q_{j|i}$. Ορίζεται[8] η διακύμανση της Gaussian κατανομής που χρησιμοποιείται στον υπολογισμό των υπό συνθήκη πιθανοτήτων $q_{j|i}$ σε $\frac{1}{\sqrt{2}}$ Ως εκ τούτου, μοντελοποιήθηκε η ομοιότητα του σημείου χάρτη y_j με το σημείο χάρτη y_i από :

$$q_{f|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

Και πάλι, αφού ενδιαφέρει μόνο να μοντελοποιηθούν ομοιότητες ζευγών, ορίστηκε ότι $q_{i|i} = 0$. Εάν τα σημεία χάρτη y_i και y_j σωστά μοντελοποιούν την ομοιότητα μεταξύ των πολλών διαστάσεων σημείων δεδομένων x_i και x_j, οι πιθανότητες υπό όρους p_{jji} και q_{jji} θα είναι ίσες. Με την παρατήρηση αυτή, ο SNE επιδιώκει να βρει μια αναπαράσταση δεδομένων λίγων διαστάσεων που ελαχιστοποιεί την αναντιστοιχία μεταξύ $p_{j|i}$ και $q_{j|i}$. Ένα φυσικό μέτρο της πιστότητας με την οποία το $q_{j|i}$ μοντελοποιεί το $p_{j|i}$ είναι η απόκλιση Kullback - Leibler (η οποία στην περίπτωση αυτή είναι ίση με την διασταυρούμενη εντροπία μέχρι μιας προστιθέμενης σταθεράς). Ο SNE ελαχιστοποιεί το άθροισμα των αποκλίσεων Kullback - Leibler σε όλα τα σημεία δεδομένων χρησιμοποιώντας μια μέθοδο μείωσης κλίσης. Η συνάρτηση κόστους C δίνεται από

$$C = \sum_i KL(P_i \backslash\backslash Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}} \qquad (2)$$

όπου το P_i αντιπροσωπεύει την κατανομή πιθανότητας υπό όρους έναντι όλων των άλλων σημείων δεδομένων δεδομένου του σημείου δεδομένων x_i, και το Q_i αντιπροσωπεύει την

[8] Το να τεθεί η διακύμανση στις λίγων διαστάσεων Gaussians σε άλλη τιμή έχει ως αποτέλεσμα μόνο την έκδοση του τελικού χάρτη σε άλλη κλίμακα. Σημειώστε ότι χρησιμοποιώντας την ίδια διακύμανσης σε κάθε σημείο δεδομένων στο χάρτη λίγων διαστάσεων, χάνεται η ιδιότητα ότι τα δεδομένα αποτελούν τέλειο μοντέλο του σημείου αν το ενσωματώσουμε σε έναν χώρο της ίδιας διάστασης, γιατί σε έναν πολλών διαστάσεων χώρο χρησιμοποιείται διαφορετική διακύμανση σ_i σε κάθε Gaussian

κατανομή πιθανοτήτων υπό όρους σε όλα τα άλλα σημεία χάρτη με δεδομένο το σημείο χάρτη y_i. Επειδή η απόκλιση Kullback - Leibler δεν είναι συμμετρική, οι διαφορετικοί τύποι σφαλμάτων στις αποστάσεις των ζευγών στο χάρτη λίγων διαστάσεων δεν σταθμίζονται εξίσου. Συγκεκριμένα, υπάρχει ένα μεγάλο κόστος για τη χρήση ευρέως διαχωρισμένων σημείων χάρτη για την αντιπροσώπευση κοντινών σημείων δεδομένων (π.χ. για τη χρήση ενός μικρού $q_{j|i}$ για να αναπαραστήσει ένα μεγάλο $p_{j|i}$), αλλά υπάρχει μόνο ένα μικρό κόστος για τη χρήση κοντινών σημείων χάρτη για την αντιπροσώπευση ευρέως διαχωρισμένων σημείων δεδομένων. Αυτό το μικρό κόστος προέρχεται από τη σπατάλη κάποιας μάζας πιθανότητας στις σχετικές κατανομές *Q*. Με άλλα λόγια, η συνάρτηση κόστους SNE επικεντρώνεται στη διατήρηση της τοπικής δομής των δεδομένων στο χάρτη (για λογικές τιμές της διακύμανσης της Gaussian κατανομής στον χώρο πολλών διαστάσεων, σ_i).

Η παραμένουσα παράμετρος που πρέπει να επιλεγεί είναι η διακύμανση σ_i της Gaussian κατανομής που είναι κεντραρισμένη σε κάθε πολλών διαστάσεων σημείο δεδομένων, x_i. Δεν είναι πιθανό ότι υπάρχει μια μοναδική τιμή του σ_i που είναι βέλτιστη για όλα τα σημεία δεδομένων του συνόλου δεδομένων επειδή η πυκνότητα των δεδομένων είναι πιθανό να ποικίλει. Σε πυκνές περιοχές, μια μικρότερη τιμή του σ_i είναι συνήθως πιο κατάλληλη από ότι σε πιο αραιές περιοχές. Οποιαδήποτε συγκεκριμένη τιμή του σ_i προκαλεί μια κατανομή πιθανότητας, P_i, πάνω από όλα τα άλλα σημεία δεδομένων. Αυτή η κατανομή έχει μια εντροπία η οποία αυξάνεται καθώς αυξάνεται η σ_i. Ο SNE εκτελεί μια δυαδική αναζήτηση για την τιμή του σ_i που παράγει μια P_i με μια σταθερή πολυπλοκότητα που καθορίζεται από το χρήστη[9]. Η πολυπλοκότητα ορίζεται ως

$$Perp(P_i) = 2^{H(P_i)}$$

όπου $H(P_i)$ είναι η εντροπία Shannon της P_i και μετριέται σε bits

$$H(P_i) = -\sum_j p_{j|i} log_2 p_{j|i}$$

Η πολυπλοκότητα μπορεί να ερμηνευθεί ως μια ομαλή μέτρηση του αποτελεσματικού αριθμού γειτόνων. Η απόδοση του SNE είναι αρκετά συμπαγής σε αλλαγές στην πολυπλοκότητα, και οι τυπικές τιμές είναι μεταξύ 5 και 50.

[9] Παρατηρήστε ότι η πολυπλοκότητα αυξάνεται μονοτονικά με τη διακύμανση σ_i

Η ελαχιστοποίηση της συνάρτησης κόστους στην εξίσωση 2 εκτελείται χρησιμοποιώντας μια μέθοδο ελάττωσης κλίσης. Η κλίση έχει αρκετά απλή εξίσωση

$$\frac{\delta C}{\delta y_i} = 2\sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Φυσικά, η κλίση μπορεί να ερμηνευτεί ως η προκύπτουσα κινούσα δύναμη που δημιουργείται από ένα σύνολο ελαστικών συνδέσμων μεταξύ του σημείου χάρτη y_i και όλων των άλλων σημείων χάρτη y_j. Όλα οι ελαστικοί σύνδεσμοι ασκούν επιρροή κατά μήκος της κατεύθυνσης (y_i-y_j).

Ο ελαστικός σύνδεσμος μεταξύ των y_i και y_j απωθεί ή προσελκύει τα σημεία χάρτη ανάλογα με το αν η απόσταση μεταξύ των δύο στο χάρτη είναι πολύ μικρή ή υπερβολικά μεγάλη ώστε να αντιπροσωπεύει τις ομοιότητες μεταξύ των δύο σημείων δεδομένων πολλών διαστάσεων. Η κινούσα δύναμη που ασκείται από τον ελαστικό σύνδεσμο μεταξύ των y_i και y_j είναι ανάλογη προς το μήκος του και επίσης ανάλογη της ακαμψίας του, η οποία είναι η αναντιστοιχία ($p_{j|i}$ -$q_{j|i}$ + $p_{i|j}$ -$q_{i|j}$) μεταξύ των ομοιοτήτων των ζευγών των σημείων δεδομένων και των σημείων του χάρτη.

Η μείωση της κλίσης αρχίζει με τη δειγματοληψία σημείων χάρτη τυχαία από μία ισοτροπική Gaussian κατανομή με μικρή διακύμανση που επικεντρώνεται γύρω από την προέλευση. Προκειμένου να επιταχυνθεί η βελτιστοποίηση και να αποφευχθούν τα φτωχά τοπικά ελάχιστα, προστίθεται ένας σχετικά μεγάλος όρος ορμής στην κλίση. Με άλλα λόγια, η τρέχουσα κλίση προστίθεται σε ένα εκθετικά φθίνων άθροισμα προηγούμενων κλίσεων προκειμένου να προσδιοριστούν οι αλλαγές στις συντεταγμένες των σημείων χάρτη σε κάθε επανάληψη της αναζήτησης κλίσης. Μαθηματικά, η ενημέρωση της κλίσης με έναν όρο ορμής δίνεται από το

$$Y^{(t)} = Y^{(t-1)} + n\frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$$

όπου $Y^{(t)}$ υποδεικνύει τη λύση στην επανάληψη *t*, το η δείχνει τον ρυθμό εκμάθησης, και το α(t) αντιπροσωπεύει την ορμή στην επανάληψη *t*.

Επιπλέον, στα πρώιμα στάδια της βελτιστοποίησης, ο Gaussian θόρυβος προστίθεται στα σημεία χάρτη μετά από κάθε επανάληψη. Η σταδιακή μείωση της διακύμανσης αυτού του θορύβου

εκτελεί έναν τύπο προσομοίωσης ανόπτησης που βοηθά τη βελτιστοποίηση να ξεφύγει από τα κακά τοπικά ελάχιστα στη συνάρτηση κόστους. Εάν η διακύμανση του θορύβου αλλάζει πολύ αργά στο κρίσιμο σημείο κατά το οποίο αρχίζει να διαμορφώνεται η καθολική δομή του χάρτη, ο SNE τείνει να βρει χάρτες με καλύτερη καθολική οργάνωση. Δυστυχώς, αυτό απαιτεί λογικές επιλογές της αρχικής ποσότητας του Gaussian θορύβου και του ρυθμού με τον οποίο ελαττώνεται.

Επιπλέον, αυτές οι επιλογές αλληλεπιδρούν με την ποσότητα ορμής και το μέγεθος βήματος που χρησιμοποιούνται στην μείωση της κλίσης. Επομένως, είναι σύνηθες η εκτέλεση της βελτιστοποίησης αρκετές φορές σε ένα σύνολο δεδομένων για να βρεθούν οι κατάλληλες τιμές για τις παραμέτρους. Από την άποψη αυτή, ο SNE είναι κατώτερος από τις μεθόδους που επιτρέπουν την κυρτή βελτιστοποίηση και θα ήταν χρήσιμο να βρεθεί μια μέθοδος βελτιστοποίησης που να δίνει καλά αποτελέσματα χωρίς να απαιτείται ο επιπλέον χρόνος υπολογισμού.

4.2. t-Κατανεμημένη Στοχαστική Ενσωμάτωση γειτόνων

Προηγούμενα συζητήθηκε η SNE όπως παρουσιάστηκε από τους Hinton και Roweis (2002). Παρόλο που η SNE κατασκευάζει εύλογα καλές οπτικές αναπαραστάσεις, παρεμποδίζεται από μια συνάρτηση κόστους που είναι δύσκολο να βελτιστοποιηθεί και από ένα πρόβλημα που ονομάζουμε "πρόβλημα συγκέντρωσης". Σε αυτό το σημείο παρουσιάζεται μια νέα τεχνική που ονομάζεται " t-Κατανεμημένη Στοχαστική Ενσωμάτωση γειτόνων " ή "t-SNE" που στοχεύει στην αποφυγή αυτών των προβλημάτων. Η συνάρτηση κόστους που χρησιμοποιείται από την t-SNE διαφέρει από εκείνη που χρησιμοποιείται από την SNE με δύο τρόπους: (1) χρησιμοποιεί μία συμμετρική έκδοση της συνάρτησης κόστους SNE με απλούστερες κλίσεις που παρουσιάστηκε συνοπτικά από τους Cook et al (2007) και (2) χρησιμοποιεί μια κατανομή Student-t αντί της Gaussian κατανομής για να υπολογίσει την ομοιότητα μεταξύ δύο σημείων στο χώρο των λίγων διαστάσεων. Η t-SNE χρησιμοποιεί μια κατανομή με υψηλή συγκέντρωση άκρων στο χώρο των λίγων διαστάσεων για να επιλύσει τόσο το πρόβλημα συγκέντρωσης όσο και τα προβλήματα βελτιστοποίησης της SNE.

4.3. Συμμετρική SNE

Ως εναλλακτική λύση για την ελαχιστοποίηση του αθροίσματος των αποκλίσεων Kullback - Leibler μεταξύ των υπό συνθήκη πιθανοτήτων $p_{j|i}$ και $q_{j|i}$, είναι επίσης δυνατό να ελαχιστοποιηθεί μια απόκλιση Kullback - Leibler μεταξύ μιας κοινής κατανομής πιθανοτήτων, *P*, στον χώρο πολλών διαστάσεων και μιας κοινής κατανομής πιθανοτήτων, *Q*, στον χώρο λίγων διαστάσεων:

$$C = KL(P\backslash\backslash Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{if}}$$

όπου και πάλι, θέτουμε p_{ii} και q_{ii} ίσα με μηδέν. Αναφέρεται αυτός ο τύπος SNE ως συμμετρική SNE, διότι αυτή έχει την ιδιότητα $p_{ij} = p_{ji}$ και $q_{ij} = q_{ji}$ για κάθε i, j. Στη συμμετρική SNE, οι ομοιότητες ανά ζεύγη στον χάρτη λίγων διαστάσεων q_{ij} δίδονται από

$$q_{ij} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq l} \exp(-\|y_k - y_l\|^2)}$$

(3)

Ο προφανής τρόπος για τον ορισμό των ομοιοτήτων των ζευγών στον χώρο πολλών διαστάσεων p_{ij} είναι

$$p_{ij} = \frac{\exp(-\left\|x_i - x_j\right\|^2/2\sigma^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2/2\sigma^2)}$$

αλλά αυτό προκαλεί προβλήματα όταν ένα πολλών διαστάσεων σημείο δεδομένων x_i είναι εξωτερικό (δηλαδή, όλες οι αποστάσεις ανά ζεύγος $||x_i - x_j||^2$ είναι μεγάλες για το x_i). Για ένα τέτοιο σημείο, οι τιμές του p_{ij} είναι εξαιρετικά μικρές για όλα τα j, οπότε η θέση του σημείου y_i στον χάρτη λίγων διαστάσεων έχει πολύ μικρή επίδραση στη συνάρτηση του κόστους.

Ως αποτέλεσμα, η θέση του σημείου του χάρτη δεν καθορίζεται καλά από τις θέσεις των άλλων σημείων του χάρτη. Ξεπερνιέται αυτό το πρόβλημα καθορίζοντας τις δεσμευμένες πιθανότητες p_{ij} στον πολλών διαστάσεων χώρο να είναι οι συμμετρικές υπό συνθήκη πιθανότητες, δηλαδή, ορίζουμε $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$. Αυτό εξασφαλίζει ότι $\sum_j p_{ij} > \frac{1}{2n}$ για όλα τα σημεία δεδομένων x_i, ως

αποτέλεσμα του οποίου κάθε σημείο δεδομένων x_i συμβάλλει σημαντικά στη συνάρτηση κόστους. Στο λίγων διαστάσεων χώρο, η συμμετρική SNE χρησιμοποιεί απλώς την Εξίσωση 3.

Το κύριο πλεονέκτημα της συμμετρικής εκδοχής της SNE είναι η απλούστερη μορφή της κλίσης της, η οποία είναι ταχύτερη στον υπολογισμό. Η κλίση της συμμετρικής SNE είναι αρκετά παρόμοια με αυτή της ασύμμετρης SNE και δίνεται από

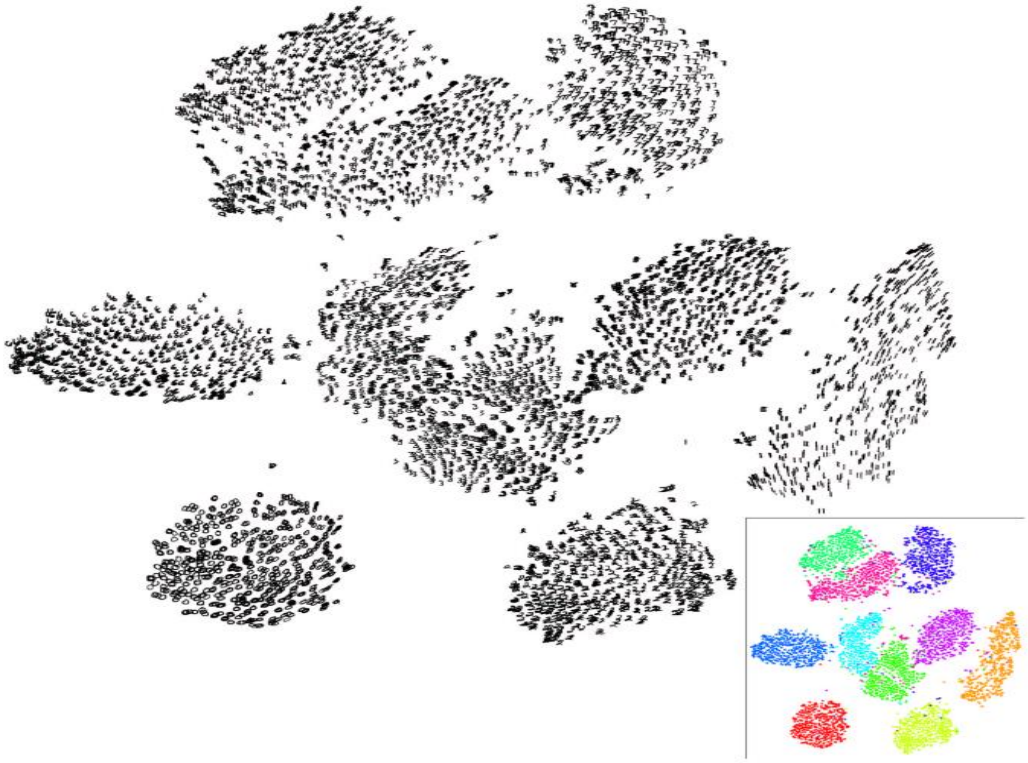
$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)$$

Σε προκαταρκτικά πειράματα, παρατηρήθηκε ότι η συμμετρική SNE φαίνεται να παράγει χάρτες που είναι εξίσου καλοί με αυτούς της ασύμμετρης SNE και μερικές φορές ακόμα καλύτεροι.

4.4. Σύγκριση με συναφείς τεχνικές

Η κλασική κλιμάκωση (Torgerson, 1952), η οποία σχετίζεται στενά με την PCA (Mardia et al, 1979, Williams, 2002), βρίσκει ένα γραμμικό μετασχηματισμό των δεδομένων που ελαχιστοποιεί το άθροισμα των τετραγωνικών σφαλμάτων μεταξύ των αποστάσεων ζευγών πολλών διαστάσεων και των αντίστοιχων τους λίγων διαστάσεων. Μια γραμμική μέθοδος όπως η κλασσική κλιμάκωση δεν είναι καλή στη μοντελοποίηση των καμπύλων πολλαπλών χωρίων και εστιάζει περισσότερο στη διατήρηση των αποστάσεων μεταξύ των ευρέως διαχωρισμένων σημείων δεδομένων παρά στη διατήρηση των αποστάσεων μεταξύ των κοντινών σημείων δεδομένων. Μια σημαντική προσέγγιση που επιχειρεί να αντιμετωπίσει τα προβλήματα της κλασσικής κλιμάκωσης είναι η χαρτογράφηση Sammon (Sammon, 1969) η οποία μεταβάλλει τη συνάρτηση κόστους της κλασσικής κλιμάκωσης διαιρώντας το τετραγωνισμένο σφάλμα στην αναπαράσταση κάθε ζεύγους ευκλείδειας απόστασης με την αρχική ευκλείδεια απόσταση στον πολλών διαστάσεων χώρο. Η προκύπτουσα συνάρτηση κόστους δίνεται από

$$C = \frac{1}{\sum_{ij} \|x_i - x_j\|} \sum_{i \neq j} \frac{(\|x_i - x_j\| - \|y_i - y_j\|)^2}{\|x_i - x_j\|}$$



Σχήμα 1: Οπτικοποίηση 6000 ψηφίων από το σύνολο δεδομένων MNIST που παράγεται από την έκδοση τυχαίας βάδισης του t-SNE (που χρησιμοποιεί όλες τις 60.000 ψηφιακές εικόνες).

όπου προστίθεται η σταθερά έξω από το άθροισμα προκειμένου να απλοποιηθεί η εξαγωγή της κλίσης. Η κύρια αδυναμία της συνάρτησης κόστους Sammon είναι ότι η σημασία της διατήρησης μικρών αποστάσεων ζευγών στον χάρτη εξαρτάται σε μεγάλο βαθμό από τις μικρές διαφορές σε αυτές τις αποστάσεις ζευγών. Συγκεκριμένα, ένα μικρό σφάλμα στο μοντέλο δύο σημείων πολλών διαστάσεων που είναι πολύ κοντά μεταξύ τους οδηγεί σε μεγάλη συμβολή στη συνάρτηση του κόστους. Δεδομένου ότι όλες οι μικρές αποστάσεις ζευγών συνιστούν την τοπική δομή των δεδομένων, φαίνεται πιο ενδεδειγμένο να επιδιώκεται να αποδίδεται περίπου ίση σημασία σε όλες τις μικρές αποστάσεις ζευγών.

Σε αντίθεση με τη χαρτογράφηση Sammon, η Gaussian kernel που χρησιμοποιείται από την t-SNE στο χώρο πολλών διαστάσεων ορίζει ένα μαλακό όριο μεταξύ της τοπικής και της καθολικής δομής των δεδομένων και των ζευγών των σημείων δεδομένων που είναι κοντά

μεταξύ τους σε σχέση με την τυπική απόκλιση της Gaussian κατανομής, και η σημασία της μοντελοποίησης αυτών των διαχωρισμών τους είναι σχεδόν ανεξάρτητη από τα μεγέθη αυτών των διαχωρισμών.

Επιπλέον, η t-SNE καθορίζει το μέγεθος της τοπικής γειτνίασης για κάθε σημείο δεδομένων χωριστά με βάση την τοπική πυκνότητα των δεδομένων.

Η ισχυρή απόδοση της t-SNE σε σύγκριση με την Isomap εξηγείται εν μέρει από την ευαισθησία του Isomap σε «βραχυκύκλωμα». Επίσης, η Isomap επικεντρώνεται κυρίως στη μοντελοποίηση μεγάλων γεωδαισιακών αποστάσεων αντί μικρών.

Η ισχυρή απόδοση της t-SNE σε σύγκριση με τη LLE οφείλεται κυρίως σε μια βασική αδυναμία της LLE: το μόνο πράγμα που εμποδίζει όλα τα σημεία δεδομένων να καταρρεύσουν σε ένα μόνο σημείο είναι ένας περιορισμός στη συνδιακύμανση της αναπαράστασης λίγων διαστάσεων. Στην πράξη, αυτός ο περιορισμός ικανοποιείται συχνά τοποθετώντας τα περισσότερα σημεία του χάρτη κοντά στο κέντρο του χάρτη και χρησιμοποιώντας μερικά ευρέως διασκορπισμένα σημεία για να δημιουργηθεί μεγάλη συνδιακύμανση. Για γραφήματα γειτνίασης που είναι σχεδόν αποσυνδεδεμένα, ο περιορισμός συνδιακύμανσης μπορεί επίσης να ικανοποιηθεί από έναν "καμπύλο" χάρτη στον οποίο υπάρχουν μερικά ευρέως διαχωρισμένα, καταρρέοντα υποσύνολα που αντιστοιχούν στα σχεδόν αποσυνδεδεμένα στοιχεία.

Επιπλέον, οι τεχνικές που βασίζονται σε γραφήματα γειτνίασης (όπως η Isomap και η LLE) δεν είναι ικανές να απεικονίζουν δεδομένα που αποτελούνται από δύο ή περισσότερα διαχωρισμένα υποσύνολα πολλαπλού χωρίου, επειδή αυτά τα δεδομένα δεν δημιουργούν ένα σχετικό γράφημα γειτνίασης. Είναι δυνατόν να δημιουργηθεί ένας ξεχωριστός χάρτης για κάθε συνδεδεμένο στοιχείο, αλλά αυτό χάνει πληροφορίες για τις σχετικές ομοιότητες των ξεχωριστών στοιχείων.

Όπως στην Isomap και στην LLE, η έκδοση τυχαίας διαδρομής της t-SNE χρησιμοποιεί γραφήματα γειτνίασης, αλλά δεν υποφέρει από προβλήματα βραχυκυκλώματος, επειδή οι ομοιότητες ζευγών μεταξύ των πολλών διαστάσεων σημείων δεδομένων υπολογίζονται με την ολοκλήρωση πάνω σε όλες τις διαδρομές μέσω του γραφήματος γειτνίασης.

Λόγω της ερμηνείας που βασίζεται στη διάχυση των υπό συνθήκη πιθανοτήτων που υπογραμμίζουν την εκδοχή τυχαίας διαδρομής της t-SNE, είναι χρήσιμο να συγκριθεί η t-SNE

με τους χάρτες διάχυσης. Οι χάρτες διάχυσης ορίζουν μια "απόσταση διάχυσης" στα πολλών διαστάσεων σημεία δεδομένων που δίνεται από

$$D^{(t)}(x_i, x_j) = \sqrt{\sum_k \frac{\left(p_{ik}^{(t)} - p_{jk}^{(t)}\right)^2}{\psi(x_k)^{(0)}}}$$

όπου η $p_{ij}^{(t)}$ αντιπροσωπεύει την πιθανότητα ενός σωματιδίου που ταξιδεύει από x_i σε x_j σε t χρονικά βήματα μέσω ενός γραφήματος πάνω στα δεδομένα με Gaussian πιθανότητες εκπομπής. Ο όρος $\psi(x_k)^{(0)}$ είναι ένα μέτρο για την τοπική πυκνότητα των σημείων και εξυπηρετεί έναν παρόμοιο σκοπό με την σταθερή πολυπλοκότητα της Gaussian kernel που χρησιμοποιείται στην SNE. Ο χάρτης διάχυσης σχηματίζεται από τα κύρια μη ασήμαντα ιδιοδιανύσματα του πίνακα Markov των τυχαίων διαδρομών μήκους *t*. Μπορεί να αποδειχθεί ότι όταν χρησιμοποιούνται όλα τα μη ασήμαντα (n-1) ιδιοδιανύσματα, οι Ευκλείδειες αποστάσεις στο χάρτη διάχυσης είναι ίσες με τις αποστάσεις διάχυσης στην αναπαράσταση δεδομένων πολλών διαστάσεων (Lafon & Lee, 2006). Μαθηματικά, οι χάρτες διάχυσης ελαχιστοποιούν το

$$C = \sum_i \sum_j \left(D^{(t)}(x_i, x_j) - \|y_i - y_j\|\right)^2$$

Ως αποτέλεσμα, οι χάρτες διάχυσης είναι επιρρεπείς στα ίδια προβλήματα με την κλασική κλιμάκωση: δίνουν μεγαλύτερη σημασία στη μοντελοποίηση των μεγάλων αποστάσεων διάχυσης των ζευγών από τις μικρές και ως εκ τούτου δεν είναι καλοί στη διατήρηση της τοπικής δομής των δεδομένων. Επιπλέον, σε αντίθεση με την έκδοση τυχαίας διαδρομής της t-SNE, οι χάρτες διάχυσης δεν έχουν φυσικό τρόπο επιλογής του μήκους *t* των τυχαίων διαδρομών.

Στο συμπληρωματικό υλικό παρουσιάζονται αποτελέσματα που αποκαλύπτουν ότι η t-SNE ξεπερνά τη CCA (Demartines & Herault, 1997), MVU (Weinberger et al., 2004) και Laplacian Eigenmaps (Belkin & Niyogi, 2002). Για τη CCA και τη στενά συσχετισμένη CDA (Lee et al, 2000), αυτά τα αποτελέσματα μπορούν να εξηγηθούν εν μέρει από το σκληρό όριο λ που οι τεχνικές αυτές ορίζουν μεταξύ της τοπικής και της καθολικής δομής, σε αντίθεση με το μαλακό όριο της t-SNE. Επιπλέον, εντός της περιοχής λ, η CCA πάσχει από την ίδια αδυναμία με τη

χαρτογράφηση Sammon: αποδίδει εξαιρετικά μεγάλη σημασία στη μοντελοποίηση της απόστασης μεταξύ δύο σημείων δεδομένων που είναι πολύ κοντά.

Όπως και η t-SNE, η MVU (Weinberger et al, 2004) προσπαθεί να μοντελοποιήσει όλους τους μικρούς διαχωρισμούς αλλά η MVU επιμένει να τις μοντελοποιήσει τέλεια (δηλαδή τους αντιμετωπίζει ως περιορισμούς) και ένας μοναδικός εσφαλμένος περιορισμός μπορεί να επηρεάσει σοβαρά την απόδοση της MVU. Αυτό μπορεί να συμβεί όταν υπάρχει βραχυκύκλωμα μεταξύ δύο τμημάτων ενός καμπύλου πολλαπλού χωρίου που βρίσκονται πολύ μακριά στις εγγενείς συντεταγμένες πολλαπλού χωρίου. Επίσης, η MVU δεν επιχειρεί να μοντελοποιήσει δομή μεγαλύτερης εμβέλειας: Απλά τραβάει τα σημεία του χάρτη όσο το δυνατόν πιο μακριά, με την επιφύλαξη των σκληρών περιορισμών και, σε αντίθεση με την t-SNE, δεν μπορεί να αναμένεται να παράγει λογική δομή μεγάλης κλίμακας στο χάρτη.

Για την Laplacian Eigenmaps, τα φτωχά αποτελέσματα σε σχέση με την t-SNE μπορούν να εξηγηθούν από το γεγονός ότι η Laplacian Eigenmaps έχει τον ίδιο περιορισμό συνδιακύμανσης με τον LLE και είναι εύκολο να εξαπατηθεί αυτός ο περιορισμός.

4.5. Αδυναμίες

Παρότι έχουμε δείξει ότι η t-SNE συγκρίνεται ευνοϊκά με άλλες τεχνικές απεικόνισης δεδομένων, η tSNE έχει τρεις πιθανές αδυναμίες: (1) είναι ασαφής ο τρόπος με τον οποίο η t-SNE ασκεί τα καθήκοντα μείωσης των διαστάσεων, (2) η σχετική τοπική φύση της t-SNE την καθιστά ευαίσθητη στην επίδραση της εγγενούς διαστάσεως των δεδομένων και (3) η t-SNE δεν είναι εγγυημένο ότι θα συγκλίνει σε ένα ολικό βέλτιστο της συνάρτηση κόστους. Παρακάτω, συζητούνται λεπτομερέστερα οι τρεις αδυναμίες.

1) Μείωση της διάστασης για άλλους σκοπούς. Δεν είναι προφανές πως η t-SNE θα επιτελέσει το γενικότερο έργο της μείωσης των διαστάσεων (δηλ. Όταν η διάσταση των δεδομένων δεν μειώνεται σε δύο ή τρεις, αλλά σε $d > 3$ διαστάσεις). Για την απλοποίηση των ζητημάτων αξιολόγησης, εξετάζεται μόνο η χρήση της t-SNE για την οπτικοποίηση δεδομένων. Η συμπεριφορά της t-SNE κατά τη μείωση των δεδομένων σε δύο ή τρεις διαστάσεις δεν μπορεί εύκολα να επεκταθεί σε $d > 3$ διαστάσεις λόγω των πυκνών άκρων της κατανομής Student-t. Σε

χώρους μεγάλης διαστάσεως, τα πυκνά άκρα περιλαμβάνουν ένα σχετικά μεγάλο μέρος της μάζας πιθανότητας κάτω από την κατανομή Student-t, το οποίο μπορεί να οδηγήσει σε *d* διαστάσεων αναπαραστάσεις που δεν διατηρούν καλά την τοπική δομή των δεδομένων. Ως εκ τούτου, για εργασίες στις οποίες η διάσταση των δεδομένων πρέπει να μειωθεί σε μια διάσταση μεγαλύτερη από τρεις, οι κατανομές Student t με περισσότερους από έναν βαθμό ελευθερίας[10] είναι πιθανό να είναι πιο κατάλληλες.

2) Η επίδραση της εγγενούς διάστασης. Η t-SNE μειώνει τη διάσταση των δεδομένων βασιζόμενη κυρίως στις τοπικές ιδιότητες των δεδομένων, γεγονός που καθιστά την t-SNE ευαίσθητη στην επίδραση της εγγενούς διάστασης των δεδομένων (Bengio, 2007). Σε σύνολα δεδομένων με υψηλή εγγενή διάσταση και για ένα υποκείμενο πολλαπλό χωρίο που ποικίλλει, μπορεί να παραβιαστεί η τοπική παραδοχή γραμμικότητας στο πολλαπλό χωρίο που κάνει έμμεσα η t-SNE (χρησιμοποιώντας ευκλείδειες αποστάσεις μεταξύ κοντινών γειτόνων). Ως αποτέλεσμα, η t-SNE μπορεί να είναι λιγότερο επιτυχημένη εάν εφαρμοστεί σε σύνολα δεδομένων με πολύ μεγάλη εγγενή διάσταση (για παράδειγμα, μια πρόσφατη μελέτη των Meytlis και Sirovich (2007) εκτιμά ότι ο χώρος των εικόνων των προσώπων θα αποτελείται από περίπου 100 διαστάσεις). Οι μέθοδοι που μαθαίνουν από τα πολλαπλά χωρία όπως η Isomap και η LLE υποφέρουν από ακριβώς τα ίδια προβλήματα (βλ. Π.χ. Bengio, 2007, van der Maaten et al, 2008). Ένας πιθανός τρόπος για την (μερική) αντιμετώπιση αυτού του προβλήματος είναι η εκτέλεση της t-SNE σε μια αναπαράσταση δεδομένων που λαμβάνεται από ένα μοντέλο που αντιπροσωπεύει αποτελεσματικά την ποικιλομορφία δεδομένων των πολλαπλών χωρίων σε μια σειρά μη γραμμικών στρωμάτων, όπως ένας αυτόματος κωδικοποιητής (Hinton & Salakhutdinov, 2006). Τέτοιες αρχιτεκτονικές βαθιάς στρώσης μπορούν να αντιπροσωπεύουν πολύπλοκες μη γραμμικές συναρτήσεις με πολύ απλούστερο τρόπο και ως εκ τούτου απαιτούν λιγότερα σημεία δεδομένων για να μάθουν μια κατάλληλη λύση (όπως φαίνεται για μια εργασία ισοτιμίας *d*-bits από τον Bengio (2007)). Η εκτέλεση της t-SNE σε αναπαράσταση δεδομένων που για παράδειγμα παράγεται από έναν αυτόματο κωδικοποιητή, είναι πιθανό να βελτιώσει την ποιότητα των κατασκευασμένων απεικονίσεων, διότι οι αυτόματοι κωδικοποιητές μπορούν να αναγνωρίσουν πολύ πολλαπλά χωρία πολλών περιπτώσεων καλύτερα από μια τοπική μέθοδο

[10] Η αύξηση των βαθμών ελευθερίας μιας κατανομής Student-t κάνει τα άκρα της κατανομής πιο λεπτά. Με άπειρους βαθμούς ελευθερίας η κατανομή Student-t είναι ίδια με την Gaussian κατανομή.

όπως η t-SNE. Ωστόσο, θα πρέπει να σημειωθεί ότι είναι εξ ορισμού αδύνατο να εκπροσωπείται πλήρως η δομή των εγγενώς πολλών διαστάσεων δεδομένων σε δύο ή τρεις διαστάσεις.

3) Η μη κυρτότητα της συνάρτησης κόστους t-SNE. Μια ωραία ιδιότητα των περισσότερων σύγχρονων τεχνικών μείωσης των διαστάσεων (όπως η κλασική κλιμάκωση, η Isomap, η LLE και οι χάρτες διάχυσης) είναι η κυρτότητα των συναρτήσεων του κόστους. Μια σημαντική αδυναμία της t-SNE είναι ότι η συνάρτηση του κόστους δεν είναι κυρτή, με αποτέλεσμα να πρέπει να επιλεγούν διάφορες παράμετροι βελτιστοποίησης. Οι κατασκευασμένες λύσεις εξαρτώνται από αυτές τις επιλογές παραμέτρων βελτιστοποίησης και μπορεί να είναι διαφορετικές κάθε φορά που η t-SNE εκτελείται από μια αρχική τυχαία διαμόρφωση σημείων χάρτη. Έχει αποδειχθεί ότι η ίδια επιλογή των παραμέτρων βελτιστοποίησης μπορεί να χρησιμοποιηθεί για μια ποικιλία διαφορετικών εργασιών απεικόνισης και έχει διαπιστωθεί ότι η ποιότητα του βέλτιστου δεν ποικίλλει πολύ από εκτέλεση σε εκτέλεση.

Επομένως, πιστεύεται ότι η αδυναμία της μεθόδου βελτιστοποίησης είναι ένας ανεπαρκής λόγος για να απορριφθεί η t-SNE υπέρ μεθόδων που οδηγούν σε κυρτά προβλήματα βελτιστοποίησης αλλά προκαλούν αξιοσημείωτα χειρότερες οπτικές αναπαραστάσεις. Το τοπικό βέλτιστο μιας συνάρτησης κόστους που συλλαμβάνει με ακρίβεια αυτό που θέλουμε σε μια απεικόνιση είναι συχνά προτιμότερο από το ολικό βέλτιστο της συνάρτησης κόστους που δεν καταγράφει σημαντικές πτυχές αυτού που θέλουμε. Επιπλέον, η κυρτότητα των συναρτήσεων κόστους μπορεί να είναι παραπλανητική, επειδή η βελτιστοποίηση τους είναι συχνά μη υπολογιστική για μεγάλα σετ δεδομένων πραγματικού κόσμου, προκαλώντας τη χρήση τεχνικών προσέγγισης (de Silva & Tenenbaum, 2003, Weinberger et al., 2007). Ακόμη και για την LLE και την Laplacian Eigenmaps, η βελτιστοποίηση γίνεται χρησιμοποιώντας επαναληπτικές μεθόδους Arnoldi (Arnoldi, 1951) ή Jacobi-Davidson (Fokkema et al., 1999), οι οποίες μπορεί να αποτύχουν να βρουν το ολικό βέλτιστο λόγω προβλημάτων σύγκλισης.

5. Εφαρμογή μεθόδων PCA και T-SNE

5.1 PCA

Η μέθοδος Principal Component Analysis, όπως προαναφέραμε, αποτελεί μία τεχνική για την μείωση των διαστάσεων ενός συνόλου δεδομένων σε λιγότερες διαστάσεις. Διατηρώντας όσο το δυνατόν περισσότερες πληροφορίες.

Πρόκειται για έναν ορθογώνιο γραμμικό μετασχηματισμό που αντιστοιχεί τα δεδομένα σε έναν νέο χώρο διαστάσεων έτσι ώστε οι συνιστώσες του νέου χώρου να περιέχουν σε μεγάλο βαθμό την διακύμανση των δεδομένων στον αρχικό, μεγαλυτέρων διαστάσεων χώρο.

Για να υπολογιστεί η PCA αρχικά υπολογίζουμε τον πίνακα συνδιακύμανσης των δεδομένων και στη συνέχεια βρίσκουμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα.

Τα ιδιοδιανύσματα με τις μεγαλύτερες ιδιοτιμές μας παρέχουν τις κατευθύνσεις προς τις οποίες μεταβάλλονται κυρίως τα δεδομένα μας (και άρα την μεγαλύτερη πληροφορία).

Επομένως, το ιδιοδιάνυσμα με την μεγαλύτερη ιδιοτιμή αποκαλείται **πρώτη κύρια συνιστώσα**, το ιδιοδιάνυσμα με την δεύτερη μεγαλύτερη ιδιοτιμή αποκαλείται **δεύτερη κύρια συνιστώσα**, κ.ο.κ.

Τέλος, ο τελικός πίνακας που μετασχηματίζει τα δεδομένα από τον αρχικό χώρο στον τελικό χώρο K διαστάσεων βρίσκεται κατατάσσοντας με φθίνοντα ιδιοτιμή τα ιδιοδιανύσματα του πίνακα και παίρνοντας τα K πρώτα ιδιοδιανύσματα, καθώς όσο μεγαλύτερη ιδιοτιμή έχει ένα ιδιοδιάνυσμα, τόσο μεγαλύτερη πληροφορία μας παρέχει

Επομένως, ο αλγόριθμος για την PCA έχει την ακόλουθη μορφή:

1. Αφαίρεση της μέσης τιμής κάθε διάστασης από τον εαυτό της.
2. Υπολογισμός της μήτρας συνδιακύμανσης Σ:
3. Υπολογισμός των ιδιοδιανυσμάτων της μήτρας Σ, y_i και των αντίστοιχων ιδιοτιμών λ_i.
4. Φθίνουσα ταξινόμηση των ιδιοδιανυσμάτων με βάση την τιμή της ιδιοτιμής τους.

5. Δημιουργία της μήτρας $T_{K\times M}$ που αποτελείται από τα πρώτα K ιδιοδιανύσματα (Κάθε γραμμή της μήτρας είναι και ένα ιδιοδιάνυσμα, με φθίνουσα σειρά σημαντικότητας από πάνω προς τα κάτω).
6. Υπολογισμός των δεδομένων στον νέο χώρο:

$$Y = TX^T$$

όπου $X_{N\times M}$ είναι τα αρχικά N διανύσματα διάστασης M και $Y_{K\times N}$ τα τελικά δεδομένα.

5.2 T-SNE

Σε αντίθεση με την PCA που είναι μια ντετερμινιστική διαδικασία μείωσης της διαστατικότητας των δεδομένων η t-SNE (t-Stochastic Neighbor Embedding) είναι στοχαστική διαδικασία, δηλαδή κάθε φορά που εφαρμόζεται στα ίδια δεδομένα μας δίνει διαφορετικό αποτέλεσμα.

Αν υποθέσουμε ένα σύνολο δεδομένων από N διανύσματα x διάστασης M το καθένα, η t-SNE υπολογίζει υπό-συνθήκη πιθανότητες $p_{j|i}$ μεταξύ του ζεύγους διανυσμάτων $x_i - x_j$, που αναπαριστούν ομοιότητα μεταξύ διανυσμάτων:

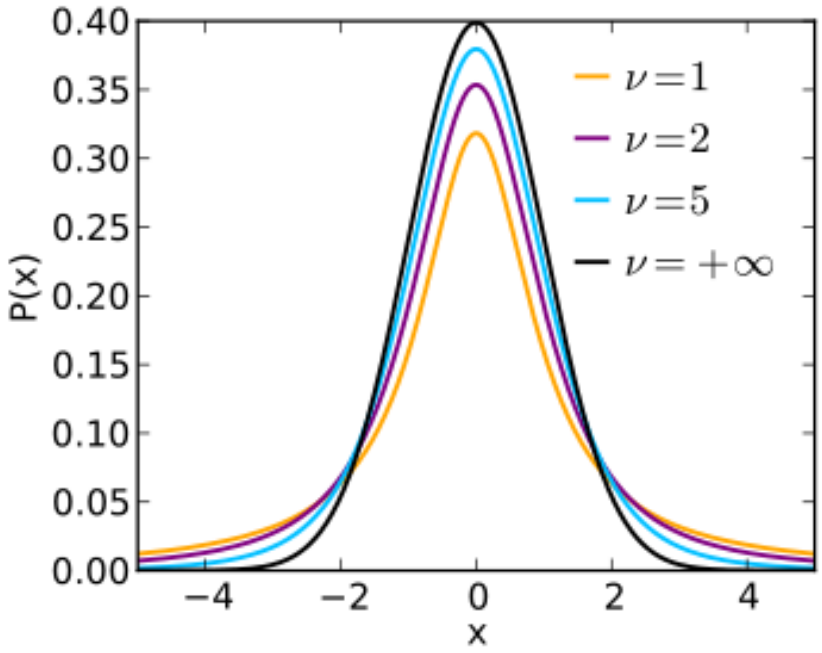
$$p_{j|i} = \frac{\exp(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma_i^2)}{\sum_{k\neq i} \exp(-\left|\left|x_i - x_k\right|\right|^2 / 2\sigma_i^2)}$$

όπου σ_i είναι η διακύμανση μιας Γκαουσιανής συνάρτησης που είναι κεντραρισμένη στο διάνυσμα x_i.

Έστω τώρα ότι η αναπαράσταση στο χώρο μικρότερων διαστάσεων των διανυσμάτων x_i, x_j είναι y_i, y_j. Στο νέο χώρο, η ομοιότητα μεταξύ των διανυσμάτων δίνεται από:

$$q_{j|i} = \frac{\left(1 + \left|\left|y_i - y_j\right|\right|^2\right)^{-1}}{\sum_{k \neq i}\left(1 + \left|\left|y_i - y_k\right|\right|^2\right)^{-1}}$$

δηλαδή στο νέο χώρο, χρησιμοποιείται κατανομή Student t με έναν βαθμό ελευθερίας.



Σχήμα 2: Κατανομή Student – t για διαφορετικούς βαθμούς ελευθερίας *ν*.

Για να βρεθούν λοιπόν οι χαμηλής διαστατικότητας αναπαραστάσεις των δεδομένων, ελαχιστοποιείται το ακόλουθο κόστος:

$$C = KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

που αποτελεί την σχετική εντροπία (που επίσης ονομάζεται και Kullback-Leibler απόκλιση) μεταξύ των κατανομών πιθανότητας P και Q. Η ελαχιστοποίηση του κόστους αυτού γίνεται με βάση μεθόδους αριθμητικής ανάλυσης, δηλαδή gradient descent.

5.3 Διαφορές PCA και T-SNE

1. Όπως προαναφέραμε, σε αντίθεση με την PCA, η T-SNE αποτελεί στοχαστική διαδικασία στην οποία η λύση με gradient descent αρχικοποιείται τυχαία κάθε φορά. Επιπλέον, η συνάρτηση κόστους είναι μη κυρτή (non convex) επομένως σταματάει σε τοπικά – αντί για κάποιο ολικό – ελάχιστα. Γι' αυτό και η οπτικοποίηση θα είναι διαφορετική για κάθε εφαρμογή του αλγορίθμου.
2. Σε αντίθεση με την T-SNE η PCA μας παρέχει τη μήτρα T με την οποία μπορούμε να μετασχηματίσουμε νέα δεδομένα.
3. Η T-SNE βρίσκει μη γραμμική απεικόνιση σε αντίθεση με την PCA που βρίσκει γραμμική απεικόνιση.

6. Εφαρμογή Μεθόδων

6.1 Πολυδιάστατα Δεδομένα Εισόδου

Θα εφαρμόσουμε τις ανωτέρω δύο μεθόδους στην βάση CMCC-Corpus η οποία αποτελείται από 757 αρχεία κειμένου. Τα αρχεία αυτά ανήκουν σε 21 συγγραφείς οι οποίοι αναπτύσσουν 6 διαφορετικά θέματα συζήτησης μέσω 6 διαφορετικών ειδών κειμένου.

Τα 6 διαφορετικά είδη (genre) κειμένου διακρίνονται σε:

Blogs

Chat

Discussion

Emails

Essays

Inteviews

Ενώ τα 6 διαφορετικά θέματα (topic) ανάπτυξης σε:

Church

Privacy rights

Gay Marriage

Marijuana Legalization

War in Iraq

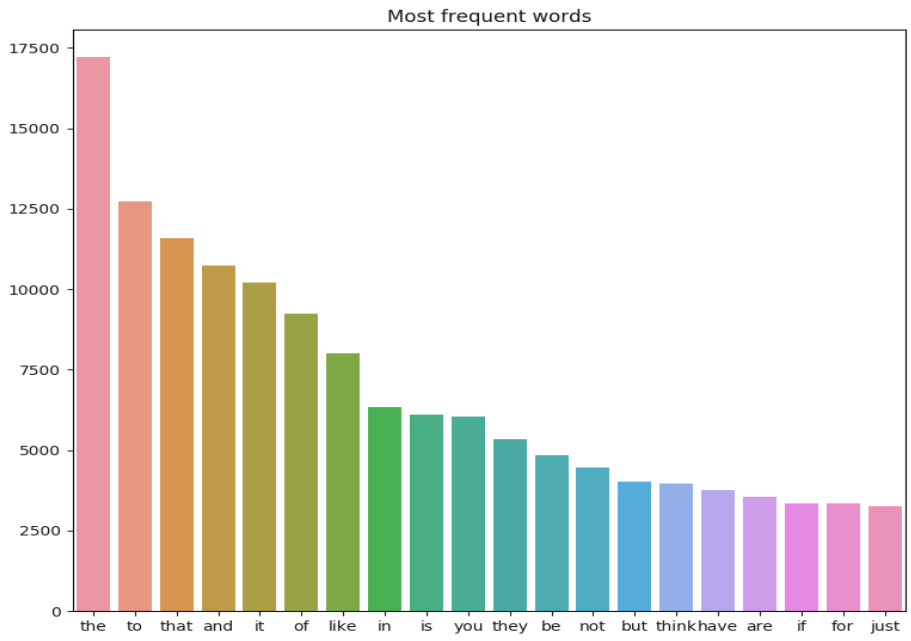
Gender Discrimination

Στις επόμενες δύο υποενότητες θα παρουσιάσουμε πειραματικά αποτελέσματα που προκύπτουν από την οπτικοποίηση του CMCC-Corpus, για τις μεθόδους PCA και t-SNE αντίστοιχα. Ενώ μελετάμε την επίδραση των διαφορετικού είδους χαρακτηριστικών που χρησιμοποιούμε για είσοδο στον εκάστοτε αλγόριθμο.

Τα χαρακτηριστικά (features) για τα οποία θα κάνουμε τους διαφορετικούς ελέγχους είναι:

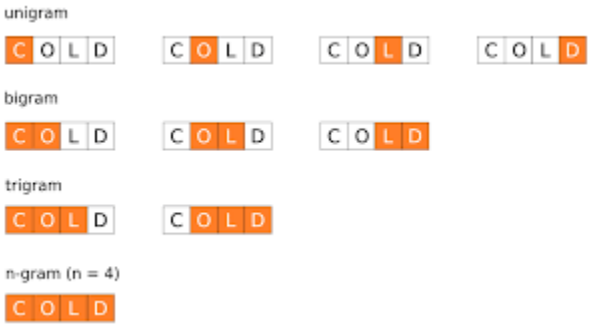
A) "words" : εάν αναλογιστούμε το κείμενο ως ένα πίνακα μια σειράς και $j_{i\ ,\ i=1,2,3..}$ στηλών, με την κάθε λέξη των κειμένων ως μια μοναδική τιμή της j_i και την τιμή της εκάστοτε $(1,j_i)$ ως τον αριθμό εμφάνισης, δηλαδή πόσες φορές εμφανίστηκε η λέξη j_i στο σύνολο των 757 κειμένων, μπορούμε να ταξινομήσουμε τον πίνακα με αριθμητική σειρά. Έτσι όταν θελήσουμε να μελετήσουμε την οπτικοποίηση των κειμένων σύμφωνα με τις 50 πιο συχνές λέξεις αναφερόμαστε στις 50 λέξεις με το μεγαλύτερο αριθμό εμφάνισης μέσα στα κείμενα της βάσης μας.

Στην παρακάτω σχήμα απεικονίζεται ένα διάγραμμα με τις 20 περισσότερο εμφανιζόμενες λέξεις μέσα στη βάση δεδομένων μας και τον αριθμό των εμφανίσεων τους.



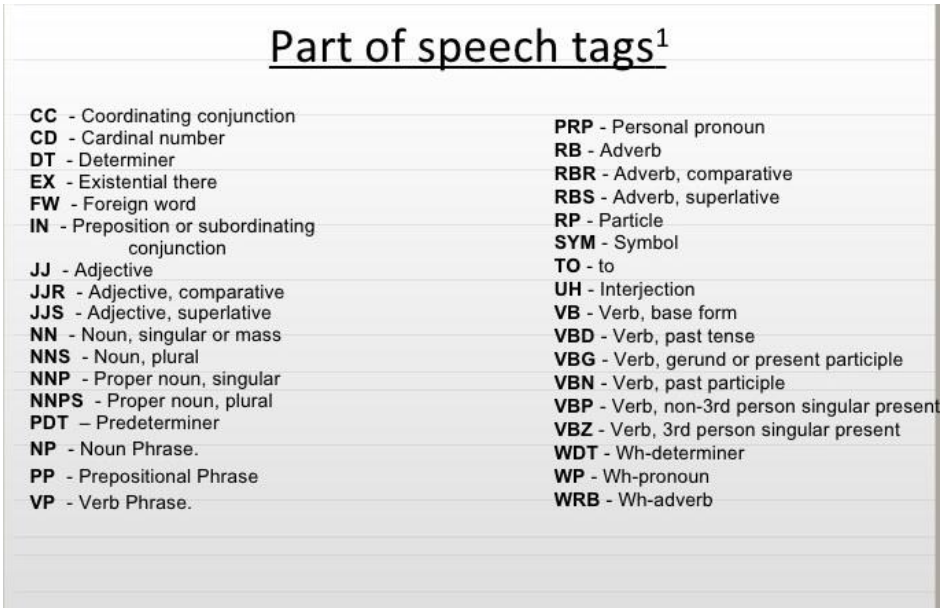
Σχήμα 3: Περισσότερο εμφανιζόμενες λέξεις στην βάση CMCC-Corpus.

Β) “ character n-grams ” : εδώ γίνεται αντίστοιχη καταμέτρηση των συνεχόμενων χαρακτήρων. Δηλαδή αντί για λέξεις κάνουμε καταμέτρηση συνεχόμενων χαρακτήρων. Τα ν-γράμματα χαρακτήρων είναι συνεχόμενες ακολουθίες n-στοιχείων σε μια πρόταση, όπου n πάντα θετικός ακέραιος αριθμός.



Σχήμα 4: Παραδείγματα ν-γράμματα χαρακτήρων για n=[1,2,3] (Besbes, 2018)

Γ) “Part of Speech (POS) n-grams” : Είναι ο συντακτικός χαρακτηρισμός των λέξεων μέσα σε ένα κείμενο. Στα παραδείγματα μας θα ασχοληθούμε με POS 2-grams. Με ένα συνδυασμό δηλαδή 2-gram τα οποία έχουν σημανθεί με POS χαρακτηρισμούς όπως στο Σχήμα 4.



Σχήμα 5: Part of Speech tags. (Dhiraj, 2017)

Unigrams	Token	*Although*	*I*	*Accept*	*The*	*authority*	*of*	*earlier*	*decisions*
	Part of speech	IN	PRP	VBP	DT	NN	IN	JJR	NNS
Bigrams	Token	*Although I*	*I accept*	*accept the*	*the authority*	*authority of*	*of earlier*	*earlier decisions*	
	Part of speech	IN PRP	PRP VBP	VBP DT	DT NN	NN IN	IN JJR	JJR NNS	
Trigrams	Token	*Although I accept*	*I accept the*	*accept the authority*	*the authority of*	*authority of earlier*	*of earlier decisions*		
	Part of speech	IN PRP VBP	PRP VBP DT	VBP DT NN	DT NN IN	NN IN JJR	IN JJR NNS		

Σχήμα 6: n-gram marked POS. (Russell, 2017)

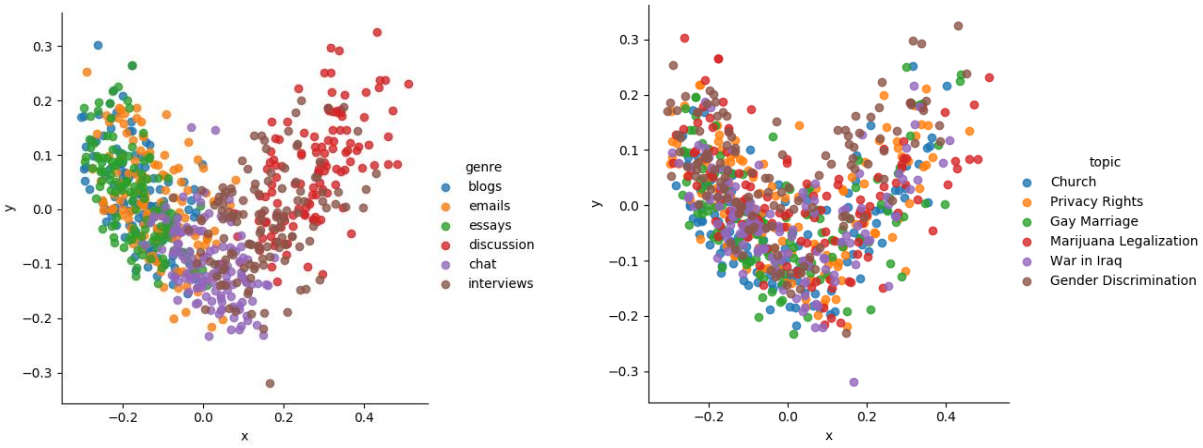
6.2 Εφαρμογή Μεθόδου PCA

Στην ενότητα αυτή θα παρουσιάσουμε τα πειραματικά αποτελέσματα που προκύπτουν από την οπτικοποίηση του CMCC-Corpus, για τη μέθοδο PCA. Επιπλέον, θα μελετήσουμε την επίδραση των διαφορετικού είδους χαρακτηριστικών που χρησιμοποιούμε για είσοδο στον εκάστοτε αλγόριθμο.

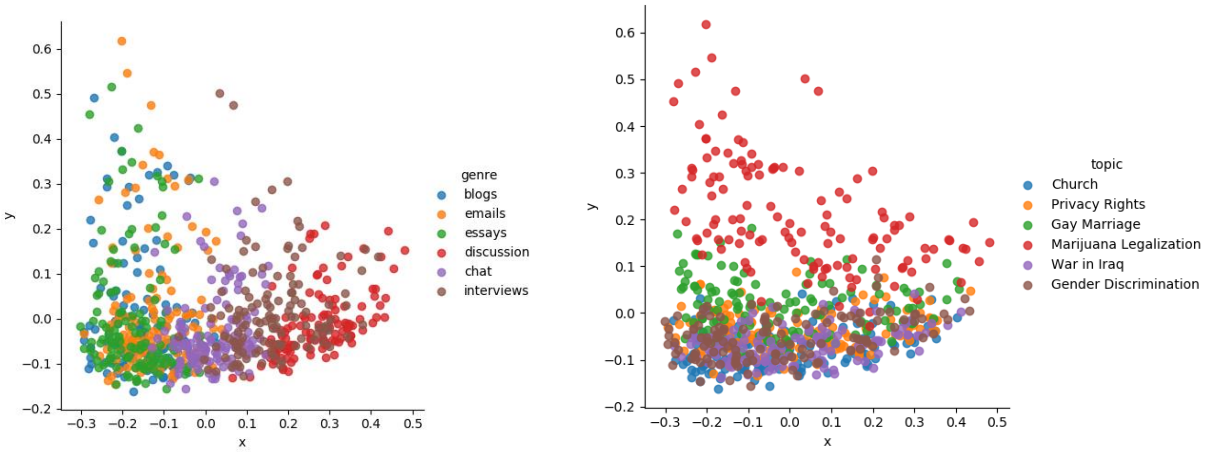
Πείραμα 1

Στο πρώτο πείραμα θα μελετήσουμε την αναπαράσταση των κειμένων με βάση ν-γράμματα χαρακτήρων σύμφωνα με το είδος αλλά και του θέμα του κειμένου.

Στις επόμενες αναπαραστάσεις γίνεται ομαδοποίηση των κειμένων σύμφωνα με το είδος και το θέμα, βάση τα τριακόσια πιο συχνά 3-γράμματα χαρακτήρων μέσα στα κείμενα και βάση των τριών χιλιάδων πιο συχνών 3-γράμματα χαρακτήρων.



Σχήματα 7: Οπτικοποίηση του CMCC με χρήση των 300 πιο συχνών character 3-grams

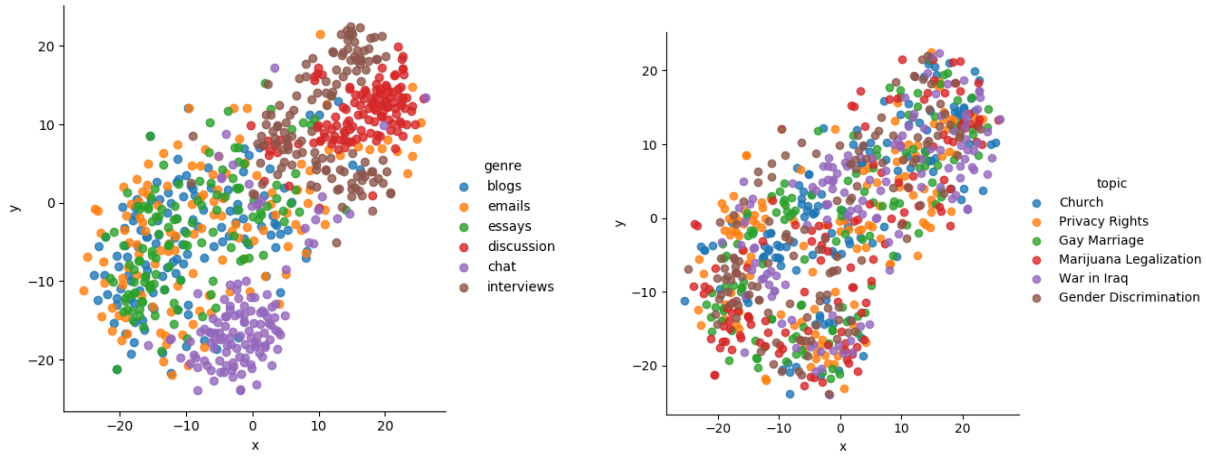


Σχήματα 8: Οπτικοποίηση του CMCC με χρήση των 3000 πιο συχνών character 3-grams

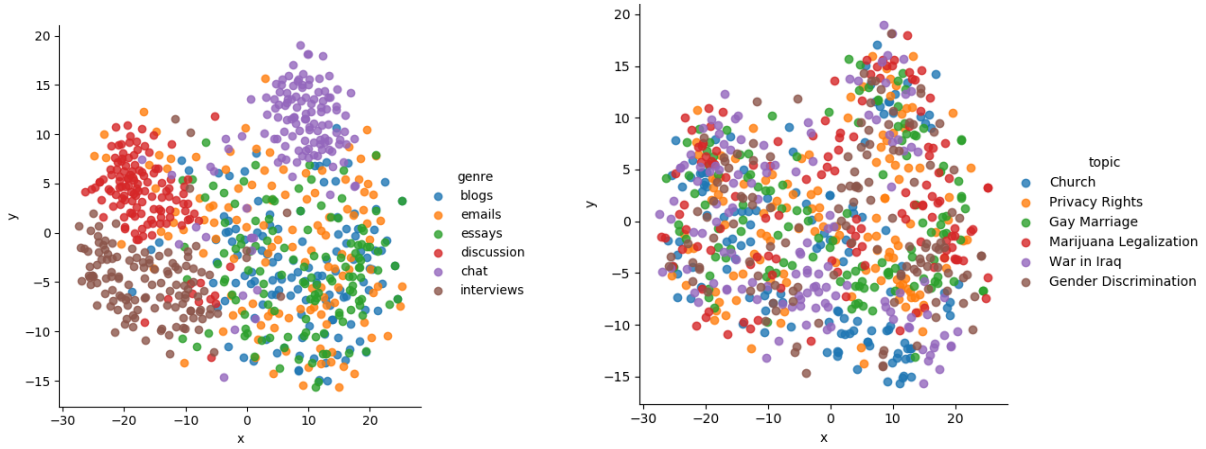
Αυτό το οποίο παρατηρείται και στα τέσσερα παραδείγματα είναι ότι για κανένα δε δημιουργήθηκαν ευδιάκριτες ομάδες, είτε αναφερόμαστε σε ένα μικρότερο δείγμα όπως αυτό των 300 πιο συχνών character 3-grams, είτε σε ένα μεγαλύτερο δείγμα όπως των δεκαπλάσιων πιο συχνών character 3-grams στη βάση δεδομένων, γεγονός το οποίο δε μας οδηγεί σε κάποιο συμπέρασμα.

Πείραμα 2

Στο δεύτερο πείραμα θα ασχοληθούμε με την οπτικοποίηση των κειμένων βάση των πιο συχνών 2-γραμμάτων μερών του λόγου (POS 2-grams) μέσα στο κείμενο.



Σχήματα 9: Οπτικοποίηση του CMCC με χρήση των 100 πιο συχνών POS 2-grams



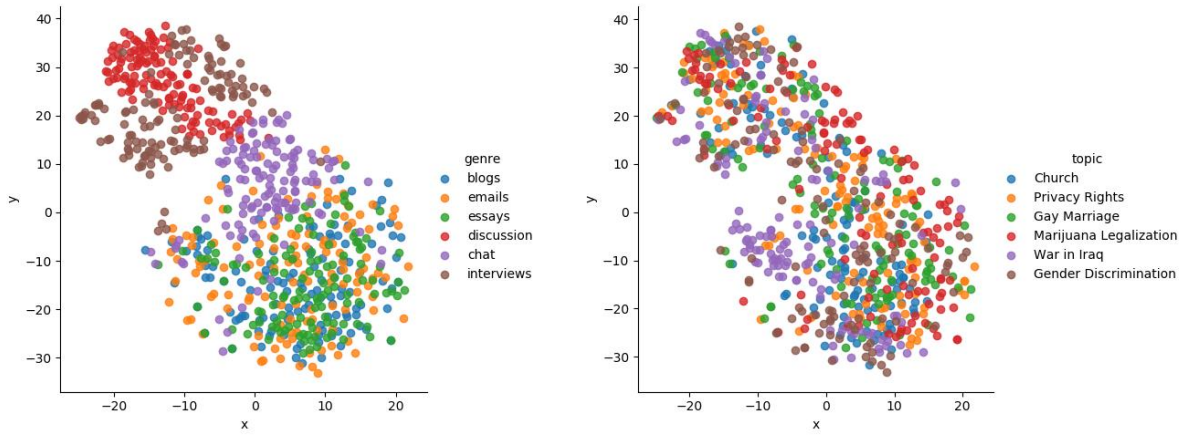
Σχήματα 10: Οπτικοποίηση του CMCC με χρήση των 500 πιο συχνών POS 2-grams

Παρατηρούμε ότι βάση των POS 2-grams δε μπορούμε να βγάλουμε συμπέρασμα για το θέμα των κειμένων καθώς για κανένα από τα δύο δείγματα δε δημιουργήθηκαν clusters. Αναφορικά με το είδος των κειμένων ωστόσο μπορούμε να καταλήξουμε σε κάποια επισφαλή συμπεράσματα για τα chat, τις συζητήσεις και ίσως τις συνεντεύξεις. Ειδικά όταν αναφερόμαστε στα 500 πιο συχνά POS 2-grams των κειμένων μας καθώς εκεί παρατηρούμε ότι είναι σχετικά

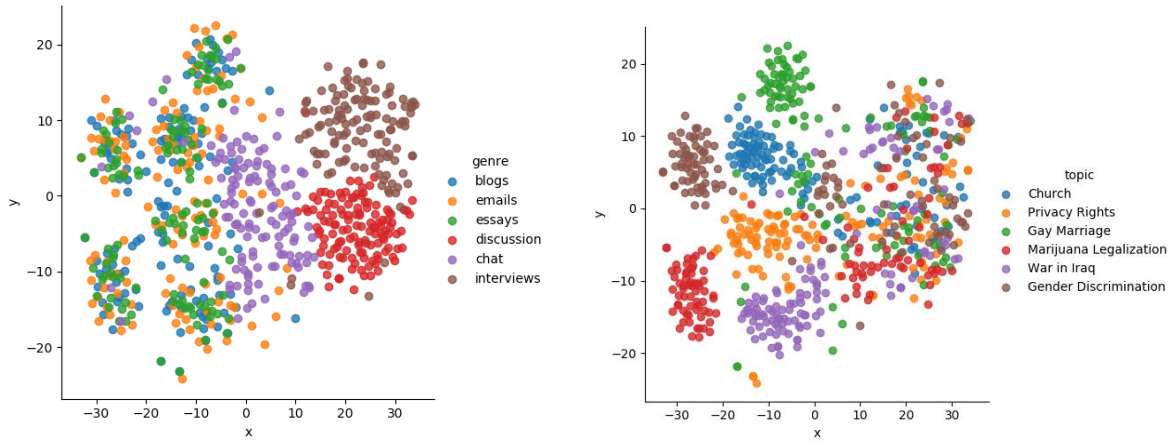
καλύτερη η οπτικοποίηση σε σχέση με τα 100 πιο συχνά POS 2-grams. Σαφέστατα όμως δεν αποτελεί καμία από τις τέσσερις μια σωστή, ακέραια αξιοποιήσιμη οπτικοποίηση δεδομένων.

Πείραμα 3

Στο τρίτο πείραμα μελετάμε την οπτικοποίηση των κειμένων βάση των πιο συχνών λέξεων που έχουν παρατηρηθεί μέσα στα κείμενα. Στα δύο πρώτα παραδείγματα λαμβάνουμε ως κριτήριο τις 50 πιο συχνές λέξεις μέσα στα κείμενα της βάσης, ενώ στη συνέχεια τις 500 πιο συχνές.



Σχήματα 11: Οπτικοποίηση του CMCC με χρήση των 50 πιο συχνών λέξεων.



Σχήματα 12: Οπτικοποίηση του CMCC με χρήση των 500 πιο συχνών λέξεων.

Στο εν λόγω παράδειγμα παρατηρούμε για πρώτη φορά ευδιάκριτα cluster εφαρμόζοντας τη μέθοδο PCA. Βλέπουμε ότι υπάρχει μια υποτυπώδης διάκριση των ειδών των κειμένων βάση των πενήντα (50) πιο συχνών λέξεων όσον αφορά τις συζητήσεις, το τσατ και τις συνεντεύξης. Σαφέστατα, η πιο αξιοποιήσιμη οπτικοποίηση είναι εκείνη των θεμάτων βάση των 500 πιο συχνών λέξεων . Εδώ έχουμε ανεξάρτητα συμπαγή cluster για όλα τα είδη των θεμάτων των κειμένων, γεγονός το οποίο δεν επαναλαμβάνεται με τα είδη των κειμένων.

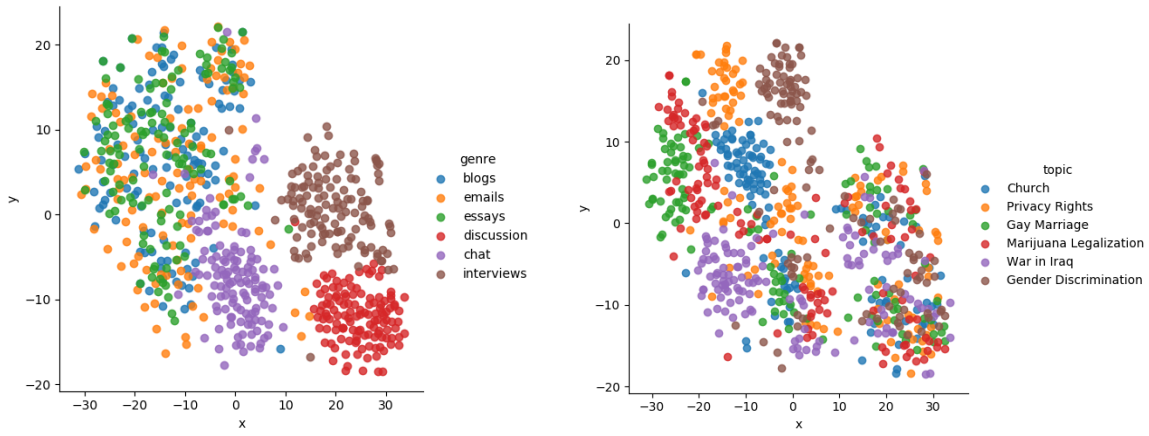
6.3 Εφαρμογή Μεθόδου T-SNE

Στην ενότητα αυτή θα παρουσιάσουμε, παρομοίως, τα πειραματικά αποτελέσματα που προκύπτουν από την οπτικοποίηση του CMCC-Corpus, για τη μέθοδο t-SNE. Επιπλέον, θα μελετήσουμε την επίδραση των διαφορετικού είδους χαρακτηριστικών που χρησιμοποιούμε για είσοδο στον εκάστοτε αλγόριθμο.

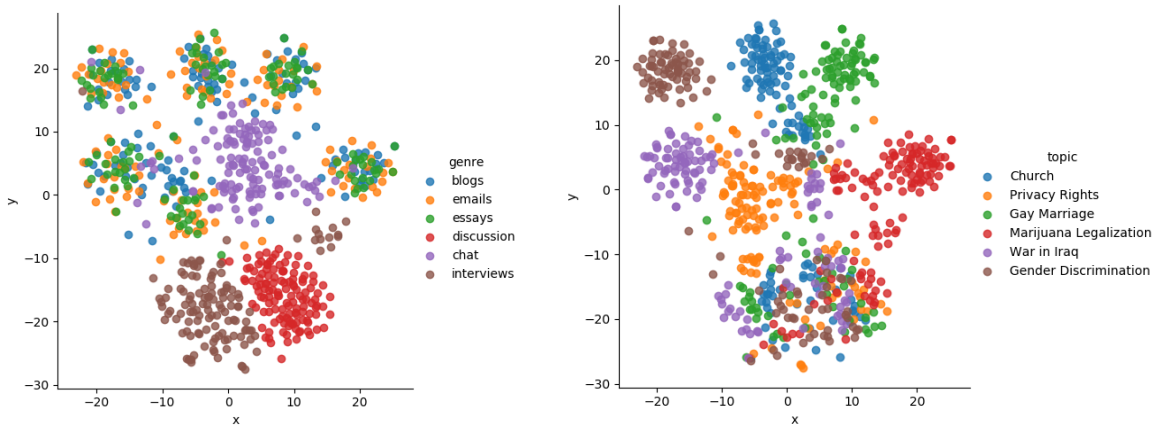
Πείραμα 1

Αρχικά, πειραματιστήκαμε με την ομαδοποίηση των κειμένων μελετώντας τα ν-γράμματα χαρακτήρων, εν προκειμένω τα 3-γράμματα χαρακτήρων .

Στα δύο πρώτα παραδείγματα η οπτικοποίηση γίνεται σύμφωνα με τα τριακόσια (300) πιο συχνά character 3-grams. Παρατηρούμε ότι η πληροφορία χάνεται και τα clusters είναι μπερδεμένα μεταξύ τους, τόσο για το είδος των κειμένων όσο και για το θέμα. Η μοναδική ίσως πληροφορία είναι αναφορικά με το είδος του κειμένου όσον αφορά τις συνεντεύξεις, τις συζητήσεις και σε μικρότερο βαθμό το chat καθώς εκεί είναι σχετικά πιο συμπαγή τα cluster.



Σχήματα 13: Οπτικοποίηση του CMCC με χρήση των 300 πιο συχνών character 3-grams

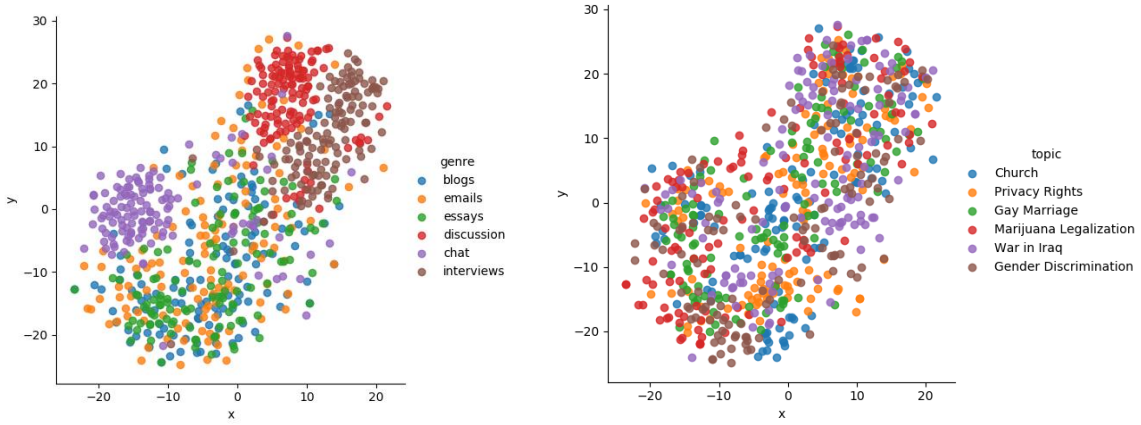


Σχήματα 14: Οπτικοποίηση του CMCC με χρήση των 3000 πιο συχνών character 3-grams

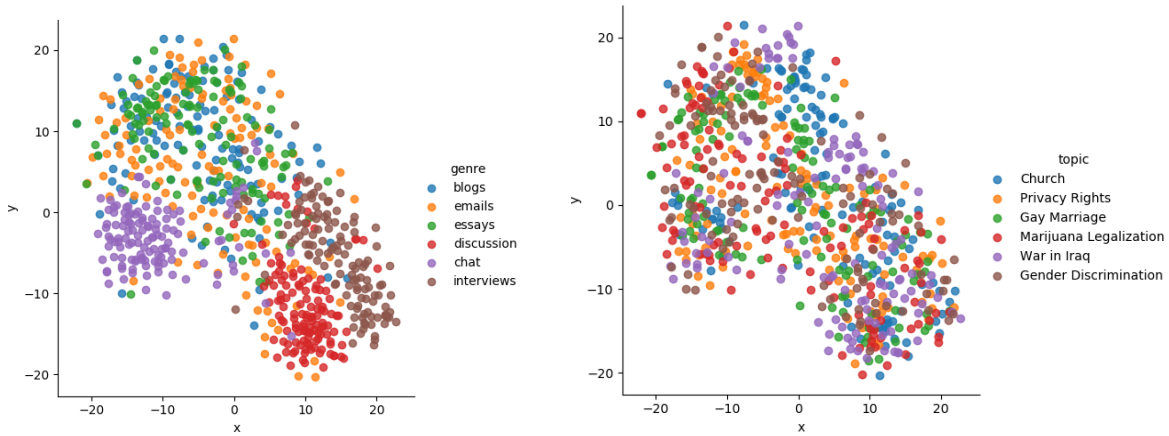
Στα επόμενα δύο παραδείγματα η οπτικοποίηση γίνεται βάση των τριών χιλιάδων (3000) πιο συχνών character 3-grams. Εδώ φαίνεται ότι τα κείμενα μπορούν να διακριθούν καλύτερα ως προς το θέμα σε σχέση με το είδος του κειμένου. Ως προς το είδος μπορούμε να διακρίνουμε τις συζητήσεις και τις συνεντεύξεις, ενώ ως προς το θέμα έχουμε πιο ευκρινή αποτελέσματα.

Πείραμα 2

Στο δεύτερο πείραμα θα ασχοληθούμε με την οπτικοποίηση των κειμένων βάση των πιο συχνών ν-γραμμάτων μερών του λόγου (POS 2-grams) μέσα στο κείμενο.



Σχήματα 15: Οπτικοποίηση του CMCC με χρήση των 100 πιο συχνών POS 2-grams



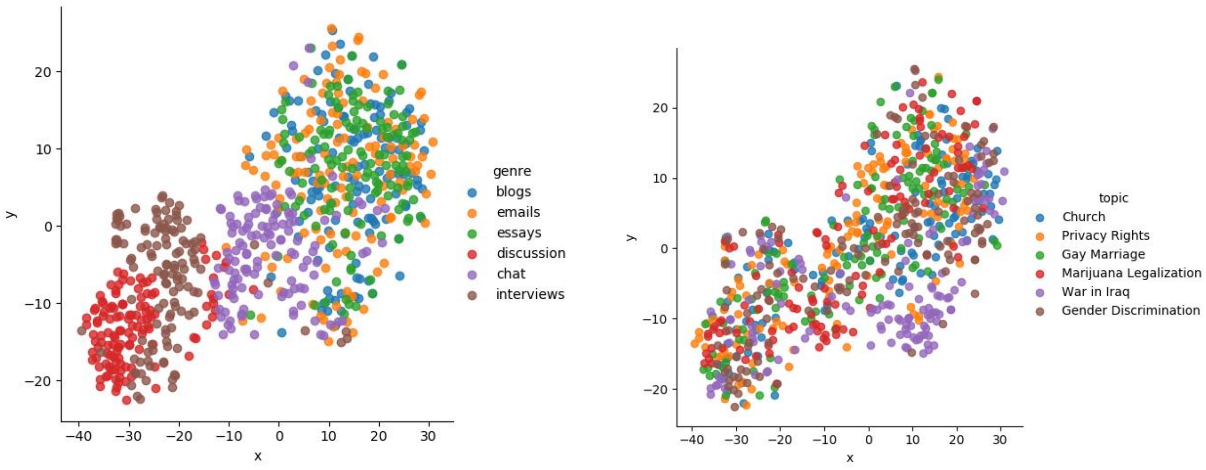
Σχήματα 16: Οπτικοποίηση του CMCC με χρήση των 500 πιο συχνών POS 2-grams

Τα δύο πρώτα παραδείγματα βασίζονται στα εκατό (100) πιο συχνά POS 2-grams ενός κειμένου, ενώ στη συνέχεια γίνεται βάση των πεντακοσίων (500) πιο συχνών POS 2-grams.

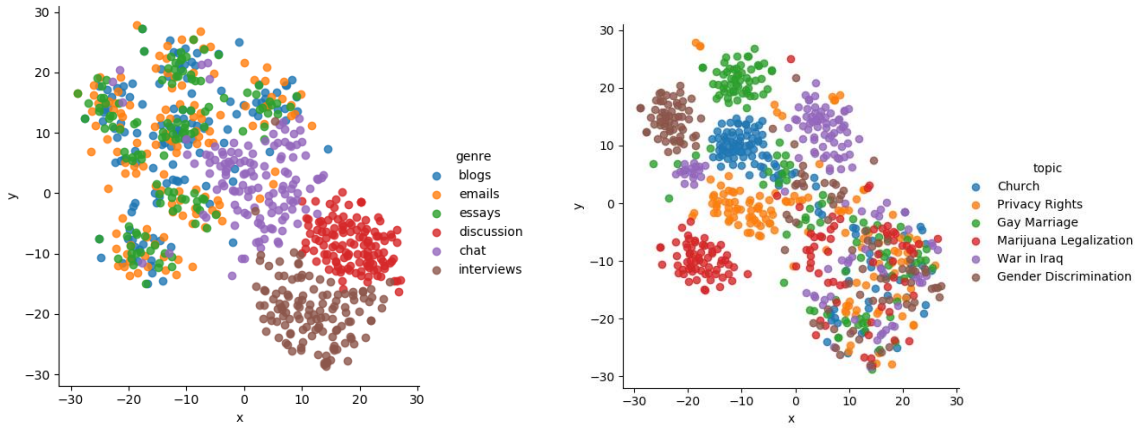
Παρατηρούμε ότι δεν είναι δυνατό να διακριθεί το θέμα των κειμένων ωστόσο είναι εφικτό να διακριθούν τα κείμενα ως προς κάποια συγκεκριμένα είδη καθώς παρατηρούνται και συμπαγή cluster. Πιο αναλυτικά, όταν οπτικοποιώ βάση των 100 πιο συχνών POS 2-grams μπορώ να διακρίνω το είδος του κειμένου για τα τσατ και τις συζητήσεις. Αντίστοιχα όταν βάση τα 500 πιο συχνά POS 2-grams μπορώ και πάλι να διακρίνω κάποια είδη κειμένων όπως οι συζητήσεις, τα τσατ και με λιγότερη ακρίβεια τις εκθέσεις και τις συνεντεύξεις.

Πείραμα 3

Στο τρίτο πείραμα εστιάζουμε στην οπτικοποίηση των κειμένων βάση των πιο συχνών λέξεων μέσα στο κείμενο.



Σχήματα 17: Οπτικοποίηση του CMCC με χρήση των 50 πιο συχνών λέξεων.



Σχήματα 18: Οπτικοποίηση του CMCC με χρήση των 500 πιο συχνών λέξεων.

Τα δύο πρώτα παραδείγματα βασίζονται στις πενήντα (50) πιο συχνές λέξεις ενός κειμένου, που συνήθως είναι άρθρα, προθέσεις κτλ., τα λεγόμενα function words, ενώ στη συνέχεια γίνεται βάση των πεντακοσίων (500) πιο συχνών λέξεων που εκτός από τα παραπάνω περιλαμβάνουν και αρκετά ουσιαστικά/ επίθετα/ ρήματα που εξαρτώνται από το θέμα. Παρατηρούμε ότι δεν είναι δυνατό να διακριθεί το θέμα των κειμένων ωστόσο είναι εφικτό να διακριθούν τα κείμενα ως προς κάποια συγκεκριμένα είδη καθώς παρατηρούνται πιο συμπαγή cluster. Πιο αναλυτικά, όταν οπτικοποιώ βάση των 50 πιο συχνών λέξεων μπορώ να διακρίνω το είδος του κειμένου για τα τσατ, τις συζητήσεις και τις συνεντεύξεις. Αντίστοιχα όταν οι συχνότητα εμφάνισης των λέξεων είναι 500 μπορώ και πάλι να διακρίνω κάποια είδη κειμένων όπως οι συζητήσεις, τα τσατ ενώ με λιγότερη ακρίβεια τις εκθέσεις και τις συνεντεύξεις. ‘Ομως η καλύτερη οπτικοποίηση βάση των 500 πιο συχνών λέξεων είναι για τα θέματα των κειμένων καθώς έχουμε ανεξάρτητα και συμπαγή cluster για όλα τα θέματα.

Εν κατακλείδι αξίζει να σημειωθεί ότι σύμφωνα με την ευκρίνεια των οπτικοποιήσεων , τα πιο ασφαλή συμπεράσματα τα βγάζουμε για το θέμα των κειμένων χρησιμοποιώντας είτε τις 500 πιο συχνές λέξεις των κειμένων είτε τα 3000 πιο συχνά 3-γράμματα χαρακτήρων.

7. Συμπεράσματα

Παρατηρώντας την εφαρμογή των μεθόδων PCA και t-SNE στην ίδια βάση δεδομένων CMCC-corpus και τα ίδια χαρακτηριστικά εισόδου, παρατηρούμε ότι λαμβάνουμε διαφορετικά αποτελέσματα. Στα character 3-grams παραδείγματα η PCA δε μας οδήγησε σε κανένα συμπέρασμα ούτε για το είδος , ούτε για το θέμα των κειμένων είτε στα 300 είτε στα 3000 πιο συχνά character 3-grams καθώς τα cluster ήταν μπερδεμένα και η πληροφορία χάνονταν . Η t-SNE ωστόσο μπορεί να υπολείπονταν σε πληροφορία στο μικρό δείγμα αλλά μας έδωσε ξεκάθαρα αποτελέσματα για το θέμα των κειμένων στο μεγαλύτερο δείγμα των character 3-grams. Τα παραδείγματα των POS 2-grams βρίσκουν σύμφωνες και τις δύο μεθόδους καθώς καμία δε μας έδωσε κάποιο σαφές ευδιάκριτο αποτέλεσμα σχετικά με είδος ή το θέμα των κειμένων ανεξαρτήτως της συχνότητας εμφάνισης των POS 2-grams. Τέλος η μοναδική συμφωνία μεταξύ των δύο μεθόδων είναι η αναγνώριση του θέματος των κειμένων όταν έχουμε ως είσοδο τις 500 πιο συχνές λέξεις, γεγονός το οποίο δεν επαληθεύεται στο μικρότερο δείγμα των 50 πιο συχνών λέξεων.

Εκ των παραδειγμάτων καταλήγουμε ότι το μεγαλύτερο δείγμα δε μας δίνει απαραίτητα και περισσότερη πληροφορία αλλά αλλάζει σαφέστατά η ποιότητα της πληροφορίας ανάλογα με τα χαρακτηριστικά εισόδου που χρησιμοποιούμε. Μια πρόταση για μελλοντική χρήση θα ήταν να χρησιμοποιηθούν πιο αντιπροσωπευτικά χαρακτηριστικά εισόδου για τα κείμενα εφαρμόζοντας τα σε συνδυασμό των δύο μεθόδων. Δηλαδή να εφαρμοστεί η t-SNE εφόσον έχει γίνει ήδη μείωση της διαστατικότητας με τη μέθοδο PCA.

8. Παράρτημα Α

```
import os
import sys
import sklearn.feature_extraction
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import sklearn.manifold
import sklearn.decomposition
import pandas as pd
from sklearn.preprocessing import MaxAbsScaler
import codecs
import nltk
import argparse

def name_to_topic_string(name):
        if name[-6] == "C":
                return "Church"
        elif name[-6] == "G":
                return "Gay Marriage"
        elif name[-6] == "I":
                return "War in Iraq"
        elif name[-6] == "P":
                return "Privacy Rights"
        elif name[-6] == "M":
                return "Marijuana Legalization"
```

```
        elif name[-6] == "S":
                return "Gender Discrimination"
        else:
                return "War in Iraq" #S1D113 file

def dir_to_subject_string(dirpath):
        return dirpath.split("/")[-1]

def dir_to_genre_string(dirpath):
        if "blogs" in dirpath:
                return "blogs"
        elif "chat" in dirpath:
                return "chat"
        elif "discussion" in dirpath:
                return "discussion"
        elif "emails" in dirpath:
                return "emails"
        elif "essays" in dirpath:
                return "essays"
        elif "interviews" in dirpath:
                return "interviews"
        else:
                raise


def prepare_dataset(corpus_path):
        documents_list = []
```

```
    labels_list_string = []
    genres_list_string = []
    topics_list_string = []
    for dirpath, subdirs, files in os.walk(corpus_path):
        for x in files:
            if x.endswith(".txt"):
                documents_list.append(os.path.join(dirpath, x))
                labels_list_string.append(dir_to_subject_string(dirpath))
                genres_list_string.append(dir_to_genre_string(dirpath))
                topics_list_string.append(name_to_topic_string(x))

    return documents_list, labels_list_string, genres_list_string, topics_list_string


def plot_most_frequent(X, feature_names, k=20):
    sums = np.sum(X,axis=0)
    sort_idx = np.argsort(sums)
    top_k_indices = sort_idx[-k:]
    top_k_indices = top_k_indices[::-1]
    top_k = sums[top_k_indices]
    plt.figure(figsize = (9, 8))
    plt.title('Most frequent words')
    sns.set_color_codes("pastel")
    sns.barplot(x=feature_names[top_k_indices], y=top_k)
    plt.savefig("topk.png")

    print("%d words appear only one time" % np.where(sums==1)[0].shape[0])
```

```
def main(args):
    documents_list, labels_list_string, genres_list_string, topics_list_string = prepare_dataset(args.corpus_path)
    print documents_list

    # words
    if args.token_type == "word":
        vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(input='filename', decode_error='ignore',use_idf=False,max_features=args.dim)
    elif args.token_type == "ngram":
        vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(input='filename', decode_error='ignore', ngram_range=(args.n,args.n), use_idf=False, analyzer='char', lowercase=False, max_features=args.dim)
    elif args.token_type == "pos_with_word":
        vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(input='string', decode_error='ignore', use_idf=False, max_features=args.dim)
        pos_tagged_documents = []
        for document in documents_list:
            d = open(document, 'r')

            with codecs.open(document, 'r', encoding='utf-8', errors='ignore') as fdata:
                d = fdata.read()

            tokens = nltk.word_tokenize(d)
            pos_tagged_tokens = nltk.pos_tag(tokens)
```

```
                pos_tagged_document = [' '.join(x) for x in pos_tagged_tokens]
                pos_tagged_documents.append(" ".join(pos_tagged_document))

        documents_list = pos_tagged_documents
    elif args.token_type == "pos":
        vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(input='string', decode_error='ignore', ngram_range=(args.n,args.n), use_idf=False, max_features=args.dim)
        pos_tagged_documents = []
        for document in documents_list:
            d = open(document, 'r')

            with codecs.open(document, 'r', encoding='utf-8', errors='ignore') as fdata:
                d = fdata.read()

            tokens = nltk.word_tokenize(d)
            pos_tagged_tokens = nltk.pos_tag(tokens)

            # print(pos_tagged_tokens)

            pos_tagged_document = [x[1] for x in pos_tagged_tokens]
            # print(pos_tagged_document)
            pos_tagged_documents.append(" ".join(pos_tagged_document))
            # print(pos_tagged_documents)

        documents_list = pos_tagged_documents
```

```
        vectorizer.fit(documents_list)
        full_vocabulary = vectorizer.vocabulary_
#       full_vocabulary_frequencies = np.sum(vectorizer.transform(documents_list).toarray(), axis=0)
#       print(len(full_vocabulary))
#       reduced_vocabulary = np.array(list(full_vocabulary.keys()))[np.where(full_vocabulary_frequencies>=args.ft)[0]]
#       print(len(reduced_vocabulary))
        for z in full_vocabulary:
      print z
        print(len(full_vocabulary))
#       print(len(reduced_vocabulary))
#
#       if args.token_type == "word":
#           vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(input='filename', decode_error='ignore', vocabulary=reduced_vocabulary, use_idf=False)
#       elif args.token_type == "ngram":
#           vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(input='filename', decode_error='ignore', ngram_range=(args.n,args.n), analyzer='char', vocabulary=reduced_vocabulary, use_idf=False, lowercase=False)
#       elif args.token_type == "pos" or args.token_type == "pos_with_word":
#           vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(input='string', decode_error='ignore', use_idf=False)
#
#       vectorizer.fit(documents_list)
        X = vectorizer.transform(documents_list).toarray()


        # scale data
```

```
if args.method == "pca":
        X_pca = sklearn.decomposition.PCA(n_components=2).fit_transform(X)

        df = pd.DataFrame()
        df['x'] = X_pca[:, 0]
        df['y'] = X_pca[:, 1]
        df['genre'] = genres_list_string
        df['subject'] = labels_list_string
        df['topic'] =  topics_list_string

        sns.lmplot(x='x', y='y', fit_reg=False, data=df, hue='genre')
        # sns.lmplot(x='x', y='y', fit_reg=False, data=df, hue='subject')
        sns.lmplot(x='x', y='y', fit_reg=False, data=df, hue='topic')
        plt.show()

elif args.method == "tsne":
        # if X.shape[1]>50:
        #         X_pca = sklearn.decomposition.PCA(n_components=50).fit_transform(X)
        # else:
        #         X_pca = X

        X_embedded = sklearn.manifold.TSNE(n_components=2).fit_transform(X)

        df = pd.DataFrame()
        df['x'] = X_embedded[:, 0]
        df['y'] = X_embedded[:, 1]
        df['genre'] = genres_list_string
```

```
            df['subject'] = labels_list_string
            df['topic'] =  topics_list_string
            sns.lmplot(x='x', y='y', fit_reg=False, data=df, hue='genre')

            sns.lmplot(x='x', y='y', fit_reg=False, data=df, hue='topic')

            plt.show()
        else:
            print("Unsupported method.")


if __name__ == '__main__':
        parser = argparse.ArgumentParser()
        parser.add_argument('--corpus_path', type=str, default="CMCC-Corpus", help='Path to corpus')
        parser.add_argument('--token_type', type=str, default="word", help='token type to use: word, ngram, pos, pos_with_word')
        parser.add_argument('--add_pos', type=bool, default=True, help='whether to include pos tags or not')
        parser.add_argument('--n', type=int, default=3, help='n-gram order (default=3) - used only if ngram')
        parser.add_argument('--ft', type=int, default=5, help='frequency threshold (default=5)')
        parser.add_argument('--method', type=str, default='tsne', help='Visualization with tsne or pca (default=tsne)')
        parser.add_argument('--dim', type=int, default=100, help='the initial dimentionality')
        args = parser.parse_args()

        main(args)
```

Βιβλιογραφία

Abo-Sinna, M. A., Abo-Elnaga, Y. Y., & Mousa, A. A. (2014). An interactive dynamic approach based on hybrid swarm optimization for solving multiobjective programming problem with fuzzy parameters. *Applied Mathematical Modelling*, *38*(7-8), 2000-2014.

Agrafiotis, D. K. (2003). Stochastic proximity embedding. *Journal of computational chemistry*, *24*(10), 1215-1221.

Anderson Jr, W. N., & Morley, T. D. (1985). Eigenvalues of the Laplacian of a graph. *Linear and multilinear algebra*, *18*(2), 141-145.

Arnoldi, W. E. (1951). The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quarterly of applied mathematics*, *9*(1), 17-29.

Balasubramanian, M., & Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, *295*(5552), 7-7.

Battista, G. D. (1994). Annoted bibliography on graph drawing algorithms. *Comput. Geom., Theory Appl.*, *4*, 235-282.

Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems* (pp. 585-591).

Belkin, M., & Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds. *Machine learning*, *56*(1-3), 209-239.

Bellman, R. E. (2015). *Adaptive control processes: a guided tour* (Vol. 2045). Princeton university press.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, *2*(1), 1-127.

Besbes, A. (2018). Overview and benchmark of traditional and deep learning models in text classification. Available from : https://www.kdnuggets.com/

Biggs, N., Biggs, N. L., & Norman, B. (1993). *Algebraic graph theory* (Vol. 67). Cambridge university press.

Borchers, B., & Young, J. G. (2007). Implementation of a primal–dual method for sdp on a shared memory parallel architecture. *Computational Optimization and Applications*, *37*(3), 355-369.

Boyd, D., & Crawford, K. (2011, September). Six provocations for big data. In *A decade in internet time: Symposium on the dynamics of the internet and society* (Vol. 21). Oxford, UK: Oxford Internet Institute.

Brand, M. (2002). MERL," Charting a manifold". In *Neural Information Proceeding Systems: Natural and Synthetic* (pp. 9-14).

Breur, T. (2016). Statistical Power Analysis and the contemporary “crisis” in social sciences.

Brun, A., Park, H. J., Knutsson, H., & Westin, C. F. (2003, February). Coloring of DT-MRI fiber traces using Laplacian eigenmaps. In *International conference on computer aided systems theory* (pp. 518-529). Springer, Berlin, Heidelberg.

Carreira-Perpinán, M. A. (1997). A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, *9*, 1-69.

Chang, H., Yeung, D. Y., & Xiong, Y. (2004, June). Super-resolution through neighbor embedding. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* (Vol. 1, pp. I-I). IEEE.

Chatfield, C., & Collin, A. J. (1980). Introduction to multivariate analysis. Published in the USA by Chapman and Hall in Association with Methuen. *Inc*, *733*, 100-117.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, *68*(342), 361-368.

Choi, H., & Choi, S. (2007). Robust kernel isomap. *Pattern Recognition*, *40*(3), 853-862.

Cook, J., Sutskever, I., Mnih, A., & Hinton, G. (2007, March). Visualizing similarity data with a mixture of maps. In *Artificial Intelligence and Statistics* (pp. 67-74).

Costa, J. A., & Hero, A. O. (2005, March). Classification constrained dimensionality reduction. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* (Vol. 5, pp. v-1077). IEEE.

Couldry, N., & Turow, J. (2014). Advertising, big data and the clearance of the public realm: marketers' new approaches to the content subsidy. *International Journal of Communication*, *8*, 1710-1726.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, *65*(3), 122-135.

De Oliveira, M. F., & Levkowitz, H. (2003). From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, *9*(3), 378-394.

Dedić, N., & Stanier, C. (2016, November). Towards differentiating business intelligence, big data, data analytics and knowledge discovery. In *International Conference on Enterprise Resource Planning Systems* (pp. 114-122). Springer, Cham.

Demartines, P., & Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on neural networks*, *8*(1), 148-154.

DeMers, D., & Cottrell, G. W. (1993). Non-linear dimensionality reduction. In *Advances in neural information processing systems* (pp. 580-587).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22.

Dhiraj, E. (2017). Stanford NLP POS Tagger Example(Maven + Eclipse). Available from: https://www.devglan.com

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269-271.

Doyle, P. (1984). Random walks and electri networks. *Mathematical Assoc. of America*.

Ekins, S., Balakin, K. V., Savchuk, N., & Ivanenkov, Y. (2006). Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *Journal of medicinal chemistry*, *49*(17), 5059-5071.

Floyd, R. W. (1962). Algorithm 97: shortest path. *Communications of the ACM*, *5*(6), 345.

Fodor, I. K. (2002). *A survey of dimension reduction techniques* (No. UCRL-ID-148494). Lawrence Livermore National Lab., CA (US).

Fokkema, D. R., Sleijpen, G. L., & Van der Vorst, H. A. (1998). Jacobi--Davidson style QR and QZ algorithms for the reduction of matrix pencils. *SIAM journal on scientific computing*, *20*(1), 94-125.

Fokkema, D. R., Sleijpen, G. L., & Van der Vorst, H. A. (1998). Jacobi--Davidson style QR and QZ algorithms for the reduction of matrix pencils. *SIAM journal on scientific computing*, *20*(1), 94-125.

Ghahramani, Z., & Hinton, G. E. (1996). *The EM algorithm for mixtures of factor analyzers* (Vol. 60). Technical Report CRG-TR-96-1, University of Toronto.

Goes, P. B. (2014). Design science research in top information systems journals. *MIS Quarterly: Management Information Systems*, *38*(1), iii-viii.

Goldberg, Y., Zakai, A., Kushnir, D., & Ritov, Y. A. (2008). Manifold learning: The price of normalization. *Journal of Machine Learning Research*, *9*(Aug), 1909-1939.

Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11), 1768-1783.

Grimes, S. (2013). Big data: Avoid 'wanna V'confusion. *Information Week*, *7*(August).

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, *47*, 98-115.

He, X., & Niyogi, P. (2004). Locality preserving projections. In *Advances in neural information processing systems* (pp. 153-160).

Hilbert, M. (2013). Big data for development: From information-to knowledge societies. *Available at SSRN 2205145*.

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, *34*(1), 135-174.

Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in neural information processing systems* (pp. 857-864).

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504-507.

Hoffmann, H. (2007). Kernel PCA for novelty detection. *Pattern recognition*, *40*(3), 863-874.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), 417.

Huber, R., Ramoser, H., Mayer, K., Penz, H., & Rubik, M. (2005). Classification of coins using an eigenspace approach. *Pattern Recognition Letters*, *26*(1), 61-75.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, *14*(1), 55-63.

Huser, V., & Cimino, J. J. (2016). Impending challenges for the use of big data. *International Journal of Radiation Oncology• Biology• Physics*, *95*(3), 890-894.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural networks*, *1*(4), 295-307.

Jenkins, O. C., & Mataric, M. J. (2002). Deriving action and behavior primitives from human motion data. In *IEEE/RSJ international conference on intelligent robots and systems* (Vol. 3, pp. 2551-2556). IEEE.

Kakutani, S. (1945). Markoff process and the Dirichlet problem. *Proceedings of the Japan Academy*, *21*(3-10), 227-233.

Kambhatla, N., & Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural computation*, *9*(7), 1493-1516.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15-25.

Keim, D. A. (2000). Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on visualization and computer graphics*, *6*(1), 59-78.

Kim, K. I., Jung, K., & Kim, H. J. (2002). Face recognition using kernel principal component analysis. *IEEE signal processing letters*, *9*(2), 40-42.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 2053951716631130.

Kokiopoulou, E., & Saad, Y. (2007). Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(12), 2143-2156.

L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: A comparative review. Online Preprint, 2008.

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence*, *28*(9), 1393-1403.

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE transactions on pattern analysis and machine intelligence*, *28*(9), 1393-1403.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of machine learning research*, *6*(Nov), 1783-1816.

Lee, J. A., & Verleysen, M. (2005). Nonlinear dimensionality reduction of data manifolds with essential loops. *Neurocomputing*, *67*, 29-53.

Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Lee, J. A., Lendasse, A., Donckers, N., & Verleysen, M. (2000). A robust nonlinear projection method. In *European Symposium on Artificial Neural Networks (ESANN'2000)*.

Lee, J., Bagheri, B., & Kao, H. A. (2014, July). Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics. In *International proceeding of int conference on industrial informatics (INDIN)* (pp. 1-6).

Lee, J., Lapira, E., Bagheri, B., & Kao, H. A. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing letters*, *1*(1), 38-41.

Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical systems and signal processing*, *42*(1-2), 314-334.

Lim, I. S., de Heras Ciechomski, P., Sarni, S., & Thalmann, D. (2003, June). Planar arrangement of high-dimensional biomedical data sets by isomap coordinates. In *16th IEEE Symposium Computer-Based Medical Systems, 2003. Proceedings.* (pp. 50-55). IEEE.

Lima, A., Zen, H., Nankaku, Y., Miyajima, C., Tokuda, K., & Kitamura, T. (2004). On the use of kernel PCA for feature extraction in speech recognition. *IEICE TRANSACTIONS on Information and Systems*, *87*(12), 2802-2811.

Lohr, S. (2013). The origins of ‘Big Data’: An etymological detective story. *New York Times*, *1*.

Maaten, Laurens van der, and Geoffrey Hinton. (2008). "Visualizing data using t-SNE." *Journal of machine learning research* : 2579-2605.

Mardia, K. V., Kent, J. T., & Bibby, J. M. Multivariate analysis. 1979. *Probability and mathematical statistics. Academic Press Inc*.

Martin-Merino, M., & Munoz, A. (2004, August). A new Sammon algorithm for sparse data visualization. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* (Vol. 1, pp. 477-481). IEEE.

Meytlis, M., & Sirovich, L. (2007). On the dimensionality of face space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(7), 1262-1267.

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, *309*(13), 1351-1352.

Nadler, B., Lafon, S., Coifman, R. R., & Kevrekidis, I. G. (2006). Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, *21*(1), 113-127.

Nam, K., Je, H., & Choi, S. (2004, July). Fast stochastic neighbor embedding: a trust-region algorithm. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)* (Vol. 1, pp. 123-128). IEEE.

Nene, S. A., Nayar, S. K., & Murase, H. (1996). Columbia object image library (coil-20).

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems* (pp. 849-856).

Niskanen, M., & Silvén, O. (2003, May). Comparison of dimensionality reduction methods for wood surface inspection. In *Sixth international conference on quality control by artificial vision* (Vol. 5132, pp. 178-189). International Society for Optics and Photonics.

O'donoghue, J., & Herbert, J. (2012). Data management within mHealth environments: Patient sensors, mobile devices, and databases. *Journal of Data and Information Quality (JDIQ)*, *4*(1), 5.

Partridge, M., & Calvo, R. A. (1998). Fast dimensionality reduction and simple PCA. *Intelligent data analysis*, *2*(3), 203-214.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Platt, J. (2005, January). FastMap, MetricMap, and Landmark MDS are all Nystrom Algorithms. In *AISTATS*.

Posadas, A. M., Vidal, F., De Miguel, F., Alguacil, G., Pena, J., Ibanez, J. M., & Morales, J. (1993). Spatial-temporal analysis of a seismic series using the principal components method: The Antequera Series, Spain, 1989. *Journal of Geophysical Research: Solid Earth*, *98*(B2), 1923-1932.

Power, D. J. (2014). Using ‘Big Data’for analytics and decision support. *Journal of Decision Systems*, *23*(2), 222-228.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health information science and systems*, *2*(1), 3.

Rajpoot, N. M., Arif, M., & Bhalerao, A. H. (2007). Unsupervised learning of shape manifolds.

Raykar, R., & Duraiswami, V. C. (2005). The manifolds of spatial hearing. *Acoust. Speech Signal Process.(ICASSP)*, 285-288.

Raytchev, B., Yoda, I., & Sakaue, K. (2004, August). Head pose estimation by nonlinear manifold learning. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* (Vol. 4, pp. 462-466). IEEE.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703-705.

Rieder, G., & Simon, J. (2017). Big data: A new empiricism and its epistemic and socio-political consequences. In *Berechenbarkeit Der Welt?* (pp. 85-105). Springer VS, Wiesbaden.

Roweis, S. T. (1998). EM algorithms for PCA and SPCA. In *Advances in neural information processing systems* (pp. 626-632).

Roweis, S. T. (2002). Automatic alignment of hidden representations. In *Sixteenth Annual Conference on Neural Information Processing Systems, Vancouver, Canada* (Vol. 15, pp. 841-848).

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323-2326.

Russell, S. (2017). Addressing 'Loss of Identity' in the Joint Judgment: Searching for 'The Individual Judge' in the Joint Judgments of the Mason Court. *The University of New South Wales law journal 40(2):670-711*

Sabarmathi, G., & Chinnaiyan, R. (2017, June). Investigations on big data features research challenges and applications. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 782-786). IEEE.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, *100*(5), 401-409.

Saxena, A., Gupta, A., & Mukerjee, A. (2004, November). Non-linear dimensionality reduction by locally linear isomaps. In *International Conference on Neural Information Processing* (pp. 1038-1043). Springer, Berlin, Heidelberg.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, *10*(5), 1299-1319.

Sejdic, E., & Falk, T. H. (Eds.). (2018). *Signal Processing and Machine Learning for Biomedical Big Data*. CRC Press.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge

university press.
Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *Departmental Papers (CIS)*, 107.
Silva, V. D., & Tenenbaum, J. B. (2003). Global versus local methods in nonlinear dimensionality reduction. In *Advances in neural information processing systems* (pp. 721-728).
Smith, Lindsay I. (2002). *A tutorial on principal components analysis.*
Snijders, C., Matzat, U., & Reips, U. D. (2012). " Big Data": big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, *7*(1), 1-5.
Song, L., Gretton, A., Borgwardt, K., & Smola, A. J. (2008). Colored maximum variance unfolding. In *Advances in neural information processing systems* (pp. 1385-1392).
Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993, July). Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization* (Vol. 1905, pp. 861-871). International Society for Optics and Photonics.
Sun, J., Zhang, H., Fang, Y., & Wang, L. H. (2005, June). Formal semantics and verification for feature modeling. In *10th IEEE International Conference on Engineering of Complex Computer Systems (ICECCS'05)* (pp. 303-312). IEEE.
Szummer, M., & Jaakkola, T. (2002). Partially labeled classification with Markov random walks. In *Advances in neural information processing systems* (pp. 945-952).
Takatsuka, M. (2001, September). An application of the self-organizing map and interactive 3-D visualization to geospatial data. In *Proceedings of the 6th International Conference on GeoComputation* (pp. 24-26).
Tenenbaum, J. B. (1998). Mapping a manifold of perceptual observations. In *Advances in neural information processing systems* (pp. 682-688).
Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, *290*(5500), 2319-2323.
Tipping, M. E. (2001). Sparse kernel principal component analysis. In *Advances in neural information processing systems* (pp. 633-639).
Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural computation*, *11*(2), 443-482.
Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4), 401-419.
Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*(4), 401-419.
Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3), 306-307.
Turk, M. A., & Pentland, A. P. (1991, June). Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

(pp. 586-591). IEEE.

van der Maaten, L. J. P., Postma, E. O., & van den Herik, H. J. (2008). Dimensionality reduction: A comparative review. published online.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, *10*(66-71), 13.

Vayena, E., Salathé, M., Madoff, L. C., & Brownstein, J. S. (2015). Ethical challenges of big data in public health.

Viceconti, M., Hunter, P., & Hose, R. (2015). Big data, big knowledge: big data for personalized healthcare. *IEEE journal of biomedical and health informatics*, *19*(4), 1209-1215.

Wang, Y. (2017, September). The Challenges and Promises of Big Data—An Engineering Perspective. In *International Workshop of Advanced Manufacturing and Automation* (pp. 599-604). Springer, Singapore.

Weinberger, K. Q., Packer, B., & Saul, L. K. (2005, January). Nonlinear Dimensionality Reduction by Semidefinite Programming and Kernel Matrix Factorization. In *AISTATS*.

Weinberger, K. Q., Sha, F., & Saul, L. K. (2004, July). Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the twenty-first international conference on Machine learning* (p. 106). ACM.

Weinberger, K. Q., Sha, F., Zhu, Q., & Saul, L. K. (2007). Graph Laplacian regularization for large-scale semidefinite programming. In *Advances in neural information processing systems* (pp. 1489-1496).

Weiss, Y. (1999). Segmentation using eigenvectors: a unifying view. In *Proceedings of the seventh IEEE international conference on computer vision* (Vol. 2, pp. 975-982). IEEE.

Welling, M., Williams, C., & Agakov, F. V. (2004). Extreme components analysis. In *Advances in Neural Information Processing Systems* (pp. 137-144).

Williams, C. K. (2002). On a connection between kernel PCA and metric multidimensional scaling. *Machine Learning*, *46*(1-3), 11-19.

Xu, R., Damelin, S., & Wunsch, D. C. (2007, August). Applications of diffusion maps in gene expression data-based cancer diagnosis analysis. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 4613-4616). IEEE.

Yang, Z., Zhang, Z., Zhang, S., Wang, J., Lin, H., & Zeng, B. (2017, July). Patent abstract analysis on Chinese big data. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 2116-2122). IEEE.

Zhu, X., Ghahramani, Z., & Lafferty, J. D. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)* (pp. 912-919).

https://en.wikipedia.org/wiki/Principal_component_analysis

https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence