Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Συγκριτική Μελέτη Γραφοθεωρητικών Αλγορίθμων Ημι-Επιτηρούμενης Μηχανικής Μάθησης σε Προβλήματα Ταξινόμησης με μεγάλη Ταξική Ανισορροπία**
Αγγλικός Τίτλος Διατριβής	**Comparative Study of Graph-Based Semi-Supervised Machine Learning Algorithms on Classification Problems with Extreme Class Imbalance**
Ονοματεπώνυμο Φοιτητή	**Αλεξανδροπούλου Χαρίκλεια**
Πατρώνυμο	**Κανέλλος**
Αριθμός Μητρώου	**ΜΠΣΠ/11006**
Επιβλέπων	**Τσιχριντζής Γεώργιος, Καθηγητής**

Πειραιάς

5/2013

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Τσιχριντζής Γεώργιος	Μαρία Βίρβου	Ευάγγελος Φούντας
Καηθηγητής	Καθηγήτρια	Καθηγητής

Περίληψη

Στη συγκεκριμένη διπλωματική εργασία γίνεται περιγραφή, ανάλυση και υλοποίηση τεσσάρων γραφοκεντρικών αλγορίθμων μερικώς επιτηρούμενης μηχανικής μάθησης. Αρχικά γίνεται μια προσπάθεια βιβλιογραφικής προσέγγισης του πεδίου της μηχανικής μάθησης με μερική επιτήρηση. Στη συνέχεια αλγόριθμοι αναλύονται θεωρητικά και εφαρμόζονται στην ταξινόμηση δεδομένων, που ανήκουν σε κλάσεις μεγάλης ταξικής ανισορροπίας. Εφαρμόζεται, επίσης, διορθωτική μέθοδος, η οποία χρησιμοποιεί την εκ των προτέρων γνώση των κλάσεων. Τέλος προτείνεται νέα διορθωτική μέθοδος, βασισμένη στις σωστές πληροφορίες που προκύπτουν από την προηγούμενη και οι οποίες αξιοποιούνται περαιτέρω.

Abstract

This thesis, entitled “Comparative Study of Graph-Based Semi-Supervised Machine Learning Algorithms on Classification Problems with Extreme Class Imbalance”, is a description, analysis and implementation of graph-based Semi-Supervised Learning algorithms. Initially we attempt to bibliographic approach the field of machine learning with partial supervision. Then, the algorithms are analyzed theoretically and applied to data classification of classes with large class imbalance. An applied correction method is also used. This method uses the prior knowledge of the classes. Finally, we propose a new correction method, based on the correct information which is obtained from the previous one and this information is being further developed.

Περιεχόμενα

Εισαγωγή

Η μηχανική μάθηση αποτελεί ένα πεδίο μελέτης αρκετά εκτενές και με μεγάλο ενδιαφέρον. Αφορά τόσο την βελτίωση των δυνατοτήτων των υπολογιστών σε συγκεκριμένους τομείς, όσο και την βελτίωση της σχέσης και της αλληλεπίδρασης μεταξύ ανθρώπου και υπολογιστή. Βρίσκει εφαρμογές στην εκπαίδευση, την οικονομία, την καθημερινότητα και τους προσωπικούς υπολογιστές και πολλές ακόμα που όλο και αυξάνονται. Η ανθρωπόμορφή συμπεριφορά των υπολογιστών αποτελεί κοινό στόχο και όνειρο, όχι μόνο των επιστημών, αλλά και όλων των ανθρώπων. Προκειμένου αυτό να επιτευχθεί χρειάζεται οι υπολογιστές να αποκτήσουν ικανότητες αντίστοιχες με αυτές της ανθρώπινης σκέψης, εν μέρει αρχικά, και αργότερα, γιατί όχι, να μπορούν να προσομοιάζουν την ανθρώπινη σκέψη σε όλες της τις εκφάνσεις.

Η συγκεκριμένη εργασία αφορά την μηχανική μάθηση με μερική επιτήρηση, που θεωρείται και ο πιο κοντινός τρόπος μηχανικής μάθησης με τον τρόπο που ο άνθρωπος μαθαίνει και αντιλαμβάνεται το περιβάλλον του. Σκοπός, προφανώς, δεν είναι η κατασκευή ανθρωπόμορφου υπολογιστή, κάτι τέτοιο απαιτεί αρκετά χρόνια συνδυαστικής έρευνας ακόμα, αλλά η μελέτη, ανάλυση και παρουσίαση αλγορίθμων που βασίζονται στην μηχανική μάθηση με ημι-επιτήρηση και συγκεκριμένα αλγορίθμων που χρησιμοποιούν γράφους δεδομένων για απόδοση ετικετών.

Γίνεται μια προσπάθεια περιληπτικής παρουσίασης του κλάδου, αναλύεται το θεωρητικό υπόβαθρο, παρουσιάζονται οι αλγόριθμοι και τα αποτελέσματά τους και γίνεται χρήση κάποιων μεθόδων για βελτίωση αυτών των αποτελεσμάτων. Οι κώδικες που χρησιμοποιήθηκαν βρίσκονται στο τέλος της εργασίας, ενώ τα βασικά συμπεράσματα από την εκτέλεσή τους παρουσιάζονται σε μορφή πινάκων και διαγραμμάτων. Η μάθηση με ημι-επιτήρηση στην εργασία αναφέρεται στην ημι-επιβλεπόμενη ταξινόμηση και ο στόχος είναι η σωστή ταξινόμηση δεδομένων όταν διαθέτουμε ένα διάνυσμα δεδομένων με ταξινομημένα και μη ταξινομημένα δεδομένα με περιορισμούς.

1. Τι είναι μηχανική μάθηση

Μηχανική μάθηση ονομάζεται ο κλάδος της τεχνητής νοημοσύνης που ασχολείται με τη μελέτη ή την κατασκευή συστημάτων που έχουν την ικανότητα να «μαθαίνουν» από ένα πλήθος δεδομένων. Ένας ορισμός για την μηχανική μάθηση δόθηκε από τον Arthur Samuel το 1959 ως «ένα πεδίο μελέτης που δίνει στους ηλεκτρονικούς υπολογιστές την δυνατότητα να μαθαίνουν χωρίς να έχουν ρητά προγραμματιστεί». Το 1997 ο T. M Mitchell διατύπωσε έναν ορισμό για την μηχανική μάθηση ως εξής: «Ένα πρόγραμμα στον ηλεκτρονικό υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E συναρτήσει κάποιων ομάδων λειτουργιών T και μιας μετρικής για την απόδοση P, εάν η απόδοση σε ζητήματα διαδικασιών επί των ομάδων λειτουργιών T, σύμφωνα με την P, βελτιώνεται σύμφωνα την εμπειρία E».

Άλλοι ορισμοί στην βιβλιογραφία ορίζουν την μηχανική μάθηση από την οπτική γωνία του ελέγχου υποθέσεων, ως δηλαδή στατιστικό πρόβλημα, όπου αναζητάτε η καλύτερη προσέγγιση των δεδομένων ή η υπόθεση που αρμόζει καλύτερα στα δεδομένα. Ουσιαστικά ο κλάδος της μηχανικής μάθηση ασχολείται με όλες εκείνες τις διαδικασίες που αποδίδουν στους ηλεκτρονικούς υπολογιστές, δυνατότητες λίγο πιο σύνθετες από την διεκπεραίωση απλών, προδιαγεγραμμένων και προγραμματισμένων διαδικασιών. Προφανώς κάθε πρόβλημα μηχανικής μάθησης εστιάζει, προς το παρών, σε μία ή σε περιορισμένο αριθμό δυνατοτήτων[1] κάθε φορά.

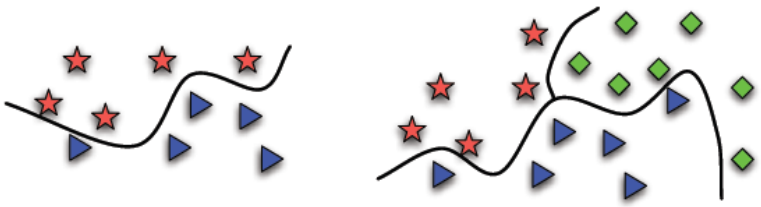
Εφαρμογές της μηχανικής μάθησης είναι η δυνατότητα ενός συστήματος να κατηγοριοποιεί δεδομένα, να αναγνωρίζει χαρακτήρες ή πρόσωπα, να αναγνωρίζει ή να προσομοιώνει την ανθρώπινη φωνή, να αναλύει και να βγάζει συμπεράσματα από μια εικόνα. Επίσης, αντικείμενο μελέτης της μηχανικής μάθησης αποτελούν η ανίχνευση εισβολών σε δίκτυα ή η επιβεβαίωση απάτης, η ανάπτυξη παιχνιδιών, οι ιατρικές εφαρμογές, ο επαναπροσδιορισμός της διεπαφής χρήστη σύμφωνα με τις προτιμήσεις του χρήστη, οι μηχανές αναζήτησης, τα recommended systems, τα συστήματα εξαγωγής πληροφορίας και άλλες εφαρμογές. Είναι προφανές ότι τα δεδομένα που μπορεί να επεξεργάζεται ένα σύστημα μηχανικής μάθησης μπορεί να είναι πίνακες, λίστες, εικόνες, χαρακτήρες, σύνθεση των προηγούμενων καθώς και

[1] Ως δυνατότητες θα μπορούσαν να εννοηθούν οι τεχνικές που προσομοιάζουν ανθρώπινη συμπεριφορά ή συμπεριφορές ζωντανών οργανισμών, όσο το δυνατόν καλύτερα.

άλλες μορφές, ανάλογα με την περίπτωση. Το μεγάλο πλήθος δεδομένων που είναι πλέον διαθέσιμο αλλά και οι όλο και αυξημένες απαιτήσεις μας από τους ηλεκτρονικούς υπολογιστές, καθιστούν, την μηχανική μάθηση αναπόσπαστο κομμάτι της επιστήμης των υπολογιστών.

Η υλοποίηση όλων αυτών των εφαρμογών γίνεται σε συνεργασία με σχετικούς κλάδους για καλύτερα αποτελέσματα. Τα βήματα, γενικά, για την υλοποίηση ενός συστήματος που χρησιμοποιεί την μηχανική μάθηση περιλαμβάνουν την θεωρητική μελέτη και προσδιορισμό του προβλήματος, την αναζήτηση πλήρους θεωρητικά ορισμένων λύσεων, την ανάπτυξη αλγορίθμων και την ενσωμάτωσή τους σε ολοκληρωμένες εφαρμογές. Χαρακτηριστικό της μηχανικής μάθησης είναι ότι επιδιώκει λύσεις ανεξάρτητες των δεδομένων, ώστε να είναι επεκτάσιμες και αξιόπιστες, δηλαδή δεν προσπαθεί να λύσει ένα μόνο πρόβλημα για ένα συγκεκριμένο σύνολο δεδομένων, αλλά ενδιαφέρεται για λύσεις που εφαρμόζονται σε διαφόρων ειδών δεδομένα, χωρίς απαραίτητα το σύστημα να γνωρίζει τι είδους δεδομένα μελετά. Τα δεδομένα χωρίζονται σε γνωστά και άγνωστα δεδομένα, τα γνωστά χρησιμοποιούνται για την εκπαίδευση του συστήματος, το σύστημα μετά την εκπαίδευση προβλέπει το είδος των άγνωστων δεδομένων και βαθμολογείται η επίδοσή του, βάση κάποιων μετρικών. Στην περίπτωση που δεν υπάρχουν γνωστά δεδομένα, αλλά μόνο άγνωστα, εξάγονται αποτελέσματα, τα οποία βασίζονται σε γενικεύσεις των άγνωστων, δηλαδή σε επαγωγικά συμπεράσματα. Τα αποτελέσματα που προκύπτουν αναφέρονται σε μια γενίκευση και ένα συμπέρασμα που εξάχθηκε από το σύνολο των δεδομένων που εισήχθηκαν στο σύστημα. Το συμπέρασμα είναι τοπικό και συγκεκριμένο, δεν αποτελεί μια γενικευμένη πραγματικότητα και δεν αντιπροσωπεύει απόλυτα τον πραγματικό κόσμο. Επίσης, όπως και σε κάθε στατιστικό πρόβλημα, για βελτίωση των αποτελεσμάτων, καλό είναι να χρησιμοποιούνται αντιπροσωπευτικά δείγματα δεδομένων στις εκπαιδεύσεις συστημάτων.

Τα είδη των προβλημάτων που καλείται να λύσει η μηχανική μάθηση ποικίλουν τόσο όσο οι εφαρμογές και το είδος των δεδομένων που είναι δυνατόν αυτές να επεξεργαστούν. Το πιο συνηθισμένο πρόβλημα είναι η δυαδική ταξινόμηση η μελέτη της οποίας έχει οδηγήσει σε πληθώρα εφαρμογών, αλγορίθμων και θεωρητικών συμπερασμάτων. Στη δυαδική ταξινόμηση το σύστημα καλείται να αποφασίσει σε πια από τις δύο κλάσεις ανήκει ένα στοιχείου του συνόλου δεδομένων προς επεξεργασία και να αποδώσει στο στοιχείο την κατάλληλη τιμή ανάλογα με την κλάση. Ο τρόπος απόδοσης της τιμής μπορεί να προκύπτει με χρήση μεταβιβαστικής ή επαγωγικής μεθόδου, να αποδίδεται κατά την είσοδο του δεδομένου στο σύστημα ή μετά την ολοκλήρωση της απόδοσης ετικέτας σε όλα τα δεδομένα ή να γίνεται βάση ελλιπών στοιχείων. Άλλο πρόβλημα είναι αυτό της πολλαπλής ταξινόμηση, το οποίο



Εικόνα 1.1. : Αριστερά δυαδική ταξινόμηση, Δεξιά ταξινόμηση με τρείς κλάσεις.

απαιτεί μεγάλη προσοχή στην εφαρμογή, ειδικά όταν εφαρμόζεται σε ιατρικά δεδομένα. Επιλύονται, επίσης, προβλήματα οπισθοδρόμησης (π.χ. εκτίμηση μετοχής την επόμενη μέρα), προσέγγιση της δομής (π.χ. στην ταξινόμηση ιστοσελίδων), προσέγγιση κατανομής, πρόβλεψης συμπεριφοράς ενός στοιχείου, δεδομένου ενός προτύπου, ή πρόβλεψη ασυνήθιστης συμπεριφοράς, δεδομένου ενός προτύπου και αρκετά άλλα είδη προβλημάτων, που λόγω των πολλών, διαφορετικών εφαρμογών είναι αδύνατον να καταγραφούν πλήρως και να ταξινομηθούν.

1.1. Τι είναι Μηχανική Μάθηση με Ημιεπιτήρηση

Η μηχανική μάθηση, παραδοσιακά, χωρίζεται σε δύο υποκατηγορίες, την μάθηση με επιτήρηση και την μάθηση χωρίς επιτήρηση.

- Στην μηχανική μάθηση *με επιτήρηση*, το μαθητευόμενο σύστημα έχει ως στόχο να παρατηρήσει ένα σύνολο δεδομένων εκπαίδευσης, αποτελούμενα από ζεύγη (χαρακτηριστικό, ετικέτα) = (x_i, y_i), και να μάθει να εκτιμά την τιμή της ετικέτας για κάθε χαρακτηριστικό του οποίου η ετικέτα δεν είναι γνωστή. Τα y_i ονομάζονται ετικέτες ή στόχοι των χαρακτηριστικών x_i. Εάν $y \in R$, δηλαδή εάν το σύνολο τιμών των ετικετών είναι συνεχές, το πρόβλημα καλείται πρόβλημα οπισθοδρόμησης, ενώ εάν οι τιμές του διανύσματος y είναι διακριτές το πρόβλημα ονομάζεται πρόβλημα ταξινόμησης. Οι αλγόριθμοι μάθησης με επιτήρηση ανήκουν σε δύο μεγάλες οικογένειες, τους Γεννητικούς και τους αλγορίθμους Διάκρισης.
- Στην μάθηση χωρίς επιτήρηση, το εκπαιδευόμενο σύστημα παρατηρεί ένα διάνυσμα από μη χαρακτηρισμένα δεδομένα, τα οποία αντιπροσωπεύονται από τα χαρακτηριστικά τους X := (x1, x2, …, xn), και εξάγει ένα συμπέρασμα, τα οποίο προκύπτει από τη γενίκευση των δεδομένων που δέχεται ως είσοδο, καθώς και τον προσδιορισμό δομών με ενδιαφέρων μέσα στο X. Συνήθως, θεωρείται ότι οι τιμές είναι τυχαίες και ανεξάρτητα επιλεγμένες από μια κατανομή του X. Μερικά τυπικά προβλήματα που λύνονται με μάθηση χωρίς επιτήρηση είναι η εκτίμηση της πυκνότητας από την οποία είναι πιθανόν να έχουν προέλθει τα στοιχεία του X, η ιεραρχική ομαδοποίηση, η ανίχνευση ακραίων τιμών.

Η *μάθηση με μερική επιτήρηση ή μάθηση με ήμι – επιτήρηση ή ημι – επιβλεπόμενη μάθηση, (Semi-Suprvised learning (ssl))* αποτελεί μια ενδιάμεση, κατηγορία μεταξύ των δύο παραπάνω, αν και αρκετές φορές θεωρείται υποκατηγορία της επιτηρούμενης μάθησης. Στο μαθητευόμενο σύστημα δίνονται πληροφορίες για τις ετικέτες των χαρακτηριστικών, όχι όμως τόσο επαρκείς όσο στη μάθηση με επιτήρηση. Το πιο συνηθισμένο μοντέλο μάθησης με ήμι – επιτήρηση είναι το επόμενο: Δίνεται στο σύστημα ένα διάνυσμα με ταξινομημένα (labeled) δεδομένα $Xl := (x1, \ldots, xl)$, ένα διάνυσμα με τις ετικέτες τους $Yl : = (y1, \ldots, yl)$ και ένα διάνυσμα με μη ταξινομημένα (unlabeled) δεδομένα $Xu := (xl+1, \ldots, xl+u)$, όπου το σύστημα καλείται να προσεγγίσει την ετικέτα των μη ταξινομημένων δεδομένων. Εάν το σύνολο των δεδομένων είναι $n \in N$ τότε ισχύει $l + u = n$ με $l, u \in N$ και το ζητούμενο διάνυσμα είναι το $Yu : = (yl+1, \ldots, yl+u)$.

Μια άλλη προσέγγιση της μάθησης με μερική επιτήρηση είναι να θεωρηθεί μάθηση με περιορισμούς. Η μάθηση με περιορισμούς μπορεί να είναι είτε επιτηρούμενη μάθηση, όπου είναι γνωστές κάποιες επιπλέον πληροφορίες από την κατανομή των δεδομένων X, είτε μη επιτηρούμενη μάθηση, όπου είναι γνωστοί κάποιοι κανόνες σχετικά με τα δεδομένα, όπως για παράδειγμα ότι «τα συγκεκριμένα σημεία ανήκουν (ή δεν ανήκουν) στην ίδια κλάση» (Abu – Mostafa, 1995). Η πρώτη κατηγορία απαιτεί ο αριθμός των κλάσεων που κατανέμονται τα δεδομένα να είναι γνωστός εκ των προτέρων, ενώ η δεύτερη όχι.

Παραδοσιακά οι ταξινομητές χρησιμοποιούν για εκπαίδευση δεδομένα ήδη ταξινομημένα, της μορφής (χαρακτηριστικό, ετικέτα) = $(x_i. y_i)$. Οι ετικέτες, δηλαδή η κλάση των δεδομένων εκπαίδευσης, είναι ήδη γνωστές. Όμως, η συλλογή ταξινομημένων δεδομένων είναι χρονοβόρα σε αντίθεση με τη συλλογή δεδομένων που απλά σχετίζονται με το πρόβλημα, αλλά δεν είναι γνωστή η κλάση στην οποία ανήκουν. Από την άλλη πλευρά, δεν έχουν μελετηθεί εκτενώς οι τρόποι που μπορεί κανείς να εξάγει αποτελέσματα από τα τελευταία. Οι αλγόριθμοι μάθησης με ήμι – επιτήρηση χρησιμοποιούν μεγάλο όγκο μη ταξινομημένων δεδομένων, αλλά και ταξινομημένων, για την κατασκευή καλύτερων ταξινομητών. Ο χρόνος συλλογής των δεδομένων, και κυρίως ο χρόνος που απαιτείται για να προ-ταξινομηθούν τα δεδομένα, είναι σαφέστατα μικρότερος, ενώ έχει επίσης αποδειχτεί ότι οι ταξινομητές αυτού του είδους είναι αρκετά ακριβείς. Έτσι, σε πειραματικό και θεωρητικό επίπεδο η μάθηση με ήμι – επιτήρηση αποτελεί έναν τομέα με ιδιαίτερο ενδιαφέρον.

Η κατασκευή τέτοιου είδους ταξινομητών απαιτεί ιδιαίτερη προσοχή, καθώς η έλλειψη ετικετών μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα εάν ο σχεδιασμός της μεθόδου ταξινόμησης είναι ελλιπής και ανακριβής. Ο χρόνος που κερδίζεται από την αποφυγή της προ - ταξινόμησης των δεδομένων εκπαίδευσης, πρέπει, σύμφωνα με τη βιβλιογραφία, εν μέρει, να καταναλωθεί για την ακριβέστερη

και καταλληλότερη σχεδίαση του ταξινομητή. Πολλοί ερευνητές έχουν παρατηρήσει ότι η χρήση μη ταξινομημένων δεδομένων μπορεί να οδηγήσει σε αρκετά μη ακριβή αποτελέσματα, εάν δεν χρησιμοποιηθεί η σωστή μέθοδος. Για παράδειγμα, οι (Elworthy, 1994) και (Cozman et al., 2003) αναφέρθηκαν σε εκπαίδευση Hidden Markov μοντέλων με μη ταξινομημένα δεδομένα όπου παρουσιάστηκε σημαντική μείωση της ακρίβειας αποτελεσμάτων. Επίσης, αρκετές μέθοδοι μάθησης με ήμι – επιτήρηση σε προβλήματα δυαδικής ταξινόμησης, με γκαουσιανές πυκνότητες πιθανότητας κλάσεων οι οποίες επικαλύπτονται σε μεγάλο βαθμό, παρουσιάζουν χαμηλή ακρίβεια. Τέτοιες μέθοδοι είναι οι transductive support vector machines (TSVMs), οι information regularization, οι γκαουσιανές διαδικασίες χωρίς μοντέλο θορύβου, οι μέθοδοι βασισμένες στην κατασκευή γραφήματος δεδομένων και άλλες. Το συγκεκριμένο πρόβλημα μπορεί να λυθεί, όμως, με επιτυχία από μοντέλα *EM (Expectation Maximization, ελλ. Αναμενόμενης Μεγιστοποίησης)*. Η επιλογή της καταλληλότερης μεθόδου για το εκάστοτε πρόβλημα αποτελεί ένα ανοιχτό ζήτημα.

Η συγκεκριμένη εργασία επικεντρώνεται στην χρήση αλγορίθμων μηχανικής μάθησης με μερική επιτήρηση σε προβλήματα ταξινόμησης σε κλάσεις. Οι κλάσεις των δεδομένων είναι δύο η αρνητική και η θετική κλάση. Οι κλάσεις παρουσιάζουν μεγάλη ταξική ανισσοροπία, δηλαδή ή θετική κλάση υπερβαίνει κατά πολύ σε σύνολο δεδομένων την αρνητική κλάση.[2][1]

1.2. Ιστορική Αναδρομή

Πιθανότατα, η πρώτη προσπάθεια χρήσης μη ταξινομημένων δεδομένων σε πρόβλημα ταξινόμησης είναι η αυτοεκπαίδευση (self-training). Πρόκειται για έναν αλγόριθμό που χρησιμοποιεί, επαναληπτικά, μάθηση με επιτήρηση. Αρχικά εκπαιδεύεται με ταξινομημένα δεδομένα. Σε κάθε βήμα ένα μέρος των μη ταξινομημένων δεδομένων, ταξινομείται σύμφωνα με την τρέχουσα συνάρτηση απόφασης. Μετά, η επιτηρούμενη μέθοδος χρησιμοποιείται ξανά για εκπαίδευση, λαμβάνοντας υπόψη και τις δικές της προβλέψεις ως επιπλέον ταξινομημένα δεδομένα. Η συγκεκριμένη ιδέα διατυπώθηκε από τους Scudder (1965), Fralick (1967), Agrawala (1970) και άλλους. Η μέθοδος στηριζόταν αποκλειστικά στην εσωτερική μέθοδο μάθησης με επιτήρηση που περιελάβανε.

Ο Vapnik (Vapnik and Chervonenkis, 1974, Vapnik and Sterin, 1977) διατύπωσε την ιδέα του μεταγωγικού συμπεράσματος ή μεταγωγής, η οποία θεωρείται ο πρόγονος της μάθησης με ημι – επιτήρηση. Σε αντίθεση με τις μεθόδους επαγωγικού συμπεράσματος, δεν προέκυπτε κάποιο γενικό συμπέρασμα, αλλά μόνο η

ταξινόμηση των μη ταξινομημένων δεδομένων. Νωρίτερα, οι Hartley and Rao (1968), είχαν ήδη εισάγει την ιδέα του μεταγωγικού συμπεράσματος.

Τη δεκαετία του 1970 με την θεώρηση του προβλήματος της εκτίμησης του κανόνα γραμμικότητας της διακρίνουσας Fisher με μη ταξινομημένα δεδομένα, δόθηκε ώθηση στην μάθηση με ημι – επιτήρηση (Hosmer 1973, McLachlan 1977, O' Neill 1978, McLachlan και Ganesalingam 1982). Πιο συγκεκριμένα, η πυκνότητα πιθανότητας κάθε κλάσης θεωρείται, και ρυθμίζεται να είναι γκαουσιανή καμπύλη με ίδια μήτρα συνδιακύμανσης με τις υπόλοιπες κλάσεις του συστήματος. Η μεγιστοποίηση της πυκνότητας πιθανότητας του μοντέλου επιτυγχάνεται με χρήση τόσο των επισημασμένων, όσο και των μη επισημασμένων δεδομένων και τη βοήθεια ένας επαναληπτικού αλγόριθμου, όπως ενός EM αλγόριθμου (Dempster et al, 1977). Μια άλλη παρόμοια προσέγγιση και μελέτη, ήταν η χρήση, αντί γκαουσιανών κατανομών, πολυωνυμικών κατανομών, που προκύπτουν από δεδομένα με ή χωρίς ετικέτες (Cooper and Freeman, 1970). Περαιτέρω γενικεύσεις της μεθόδου αυτής δημοσιεύτηκαν από διάφορους ερευνητές (Shahshahani και Landgrebe, 1994), (Miller και Uyar (1997).

Οι Ratsaby και Venkatesh (1995) διατύπωσαν κα μελέτησαν άλλη μία μέθοδο μάθησης με ημι – επιτήρηση με χρήση δύο γκαουσιανών κατανομών ενώ το 1995 οι Castelli και Cover έδειξαν ότι, στην περίπτωση ενός αναγνωρίσιμου μίγματος, με έναν άπειρο αριθμό σημείων μη ταξινομημένων, η πιθανότητα σφάλματος παρουσιάζει εκθετική σύγκλιση σύμφωνα με τον κίνδυνο Bayes. Αναγνωρίσιμες σημαίνει ότι, δοσμένης μιας πυκνότητας πιθανότητας P(x), το ανάπτυγμα $\sum_y P(y)\, P(x|y)$ είναι μοναδικό. Σχετική είναι και η περαιτέρω ανάλυση των Castelli και Cover (1996), όπου είναι γνωστές οι δεσμευμένες κατανομές των κλάσεων, αλλά δεν υπάρχει εκ των προτέρων γνώση για τις κλάσεις.

Το ενδιαφέρον για την μάθηση με ημι-επιτήρηση αυξήθηκε κατά τη δεκαετία του 1990, λόγω των εφαρμογών που έβρισκε σε προβλήματα φυσικής γλώσσας και ταξινόμησης κειμένου (Yarowsky,1995, Nigam, 1998, Blum και Mitchell, 1998 Collins και Singer, 1999, Joachims, 1999). Τη συγκεκριμένη δεκαετία μάλιστα χρησιμοποιήθηκε για πρώτη φορά ο όρος «semi – supervised» (*ελλ. ημι-επιτήρηση*) σε πρόβλημα ταξινόμησης που επιλύονται με χρήση δεδομένων ταξινομημένων και μη ταξινομημένων (Merz (1992)). Παρότι ο όρος είχε χρησιμοποιηθεί ξανά, ήταν η πρώτη φορά που η έννοιά του ήταν ίδια με την σημερινή και με αυτήν με την οποία χρησιμοποιείται στην παρούσα εργασία.

Τέλος, τα τελευταία χρόνια, χάρη στον όγκο και την ποικιλία των δεδομένων που είναι διαθέσιμα ή εύκολα στην πρόσβαση, όπως τα περιεχόμενα των ιστοσελίδων, οι σειρές των πρωτεϊνών ή οι εικόνες, η μελέτη των μεθόδων μάθησης με ημι-επιτήρηση έχει γίνει, αρκετά, δημοφιλής. [2]

1.3. Υποθέσεις - Παραδοχές

Οι μέθοδοι μάθησης με μερική επιτήρηση, όπως, άλλωστε και κάθε άλλη υπολογιστική μέθοδος, προκειμένου να έχουν επιθυμητά αποτελέσματα, και να οριστούν πλήρως, θα πρέπει να γίνουν κάποιες παραδοχές. Οι παραδοχές που λαμβάνονται υπόψη αφορούν, κυρίως, τον γενικό τρόπο απόδοσης ετικέτας σε ένα σύνολο δεδομένων. Καθορίζουν τι θεωρείται ομοιότητα, πότε ένα δεδομένα μπορεί να θεωρηθεί όμοιο με ένα άλλο, αλλά και κάποιες άλλες παραμέτρους που χαρακτηρίζουν τις μεθόδους μάθησης με μερική επιτήρηση.

- *Παραδοχή ομαλότητας*: Εάν δύο σημεία x1, x2 σε μια περιοχή υψηλής έντασης είναι κοντά, τότε θα έχουν αντίστοιχα κοντινή τιμή και στις εκτιμώμενες ετικέτες y1, y2. Για παράδειγμα αν δύο σημεία ανήκουν στην ίδια ομάδα ή κατηγορία, το ίδιο θα ισχύει, πιθανότητα, και για τις τιμές που θα τους αποδοθούν.
- *Παραδοχή ομαδοποίησης (cluster)*: Εάν τα δεδομένα ανήκουν στον ίδιο cluster, πιθανότατα ανήκουν στην ίδια κλάση. Αυτό δε σημαίνει ότι κάθε cluster αποτελείται από μόνο μία κλάση, αλλά ότι, συνήθως, δεν παρατηρούνται στοιχεία από δύο διαφορετικές κλάσεις στο ίδιο cluster. Ουσιαστικά αποτελεί μια ιδιαίτερη περίπτωση της προηγούμενης παραδοχής, και, με απλά λόγια, θεωρείται ότι είναι λιγότερο πιθανό (όχι όμως και απίθανο) να βρεθούν ίδια δεδομένα σε περιοχές χαμηλής πυκνότητας πιθανότητας
- Manifold (πολλαπλής) παραδοχή: Τα δεδομένα υψηλών διαστάσεων μπορούν να απεικονιστούν σε χαμηλής διάστασης manifold. Ένα πολύ γνωστό πρόβλημα πολλών στατιστικών μεθόδων και των αλγορίθμων μηχανικής μάθησης είναι η λεγόμενη κατάρα των διαστάσεων (curse of dimensionality). Η αύξηση του όγκου των δεδομένων, αυξάνει συνήθως και την πολλαπλή τους. Εάν τα δεδομένα αναχθούν πάνω σε manifold χαμηλών διαστάσεων, τότε ο αλγόριθμος εκμάθησης μπορεί να δραστηριοποιείται κυρίως στον χώρο της αντίστοιχης διάστασης, αποφεύγοντας έτσι την κατάρα της μεγάλης διάστασης.
- *Αρχή Vapnik*: «Όταν επιδιώκεται η λύση ενός συγκεκριμένο προβλήματος, δεν χρειάζεται πρώτα να λυθεί ένα δυσκολότερου από αυτό». Η συγκεκριμένη αρχή εννοεί ότι, για την εκτίμηση των ετικετών των unlabeled δεδομένων, δεν είναι απαραίτητη η εκτίμηση πρώτα της κατανομής από την οποία προέρχονται τα δεδομένα. [1][2]

1.4. Κατηγοριοποίηση

Στη βιβλιογραφία οι οικογένειες αλγορίθμων μάθησης με μερική επιτήρηση που συναντώνται συχνότερα είναι οι παρακάτω:

- EM, mixture model, οι οποίες έχουν ως βασική ιδέα το γεννητικό μοντέλο που αναλύεται στη συνέχεια.
- Tranductive Support Vector Machines, όπου ασχολούνται με τις περιοχές χαμηλής πυκνότητας των κλάσεων.
- Co-training, χρήση διαφορετικών οπτικών γωνιών ταυτόχρονα για εκπαίδευση (Προτάθηκαν πρώτη φορά από τους Blum και Mitchell το 1998).
- Γραφικές μέθοδοι, όπου γίνεται απόδοση ετικετών σύμφωνα με την ομαλότητα των ετικετών σε γράφημα δεδομένων.

Η κατηγοριοποίηση όμως των αλγορίθμων δεν γίνεται σύμφωνα με το όνομά τους, αλλά με βάση τη μέθοδο αντιμετώπισης των προβλημάτων μάθησης. Η ραγδαία εξέλιξη του κλάδου, τα τελευταία χρόνια, καθιστά δύσκολη την απόλυτη κατηγοριοποίηση και καταγραφή όλων των αλγορίθμων που χρησιμοποιούν τέτοιου είδους μάθηση.

Ο στόχος της ημι-επιτηρούμενης μάθησης είναι η ελαχιστοποίηση του σφάλματος ταξινόμησης και η εύρεση του P(x) των δεδομένων. Σύμφωνα με αυτό μπορεί να θεωρηθεί ως υποκατηγορία της επιτηρούμενης μάθησης, αν και τα τελευταία χρόνια αποτελεί ξεκάθαρα ένα ξεχωριστό είδος μηχανικής μάθησης. Η βασική διαφορά με την επιτηρούμενη μάθηση είναι ότι μαζί με το σύνολο των ταξινομημένων δεδομένων Dl = $\{(x_i, y_i)| \, i = 1, \dots, n\}$, τα οποία έχουν επιλεγεί ανεξάρτητα και ταυτόσημα από μια συγκεκριμένη P(x,y), διαθέτουμε και ένα σύνολο μη ταξινομημένων (unlabeled) δεδομένων Du από το όριο της P(x). Θεωρούνται τα διανύσματα $X_l = (x1, \dots, xn)$, $Y_l = (y1, \dots, yn)$ και $X_u = (x_{n+1}, \dots, x_{n+m})$, $Y_u = (y_{n+1}, \dots, y_{n+m})$, δηλαδή τα γνωστά και άγνωστα δεδομένα και ετικέτες αντίστοιχα. Για το Y_u είναι γνωστές και κάποιες πληροφορίες σχετικά με την αβεβαιότητά του.

Υπάρχουν δύο βασικοί τρόποι αντιμετώπισης προβλημάτων ημι-επιτηρούμενης μάθησης. Ο ένας είναι να αντιμετωπιστεί σαν επιτηρούμενο πρόβλημα, αγνοώντας το διάνυσμα των unlabeled δεδομένων Y_u αρχικά, και ο άλλος είναι να θεωρηθούν τα Y_u ως μια λανθάνουσα κλάση σε μια γενική εκτίμηση του P(x) των δεδομένων, η οποία έγινε με χρήση ημι-επιτηρούμενης μεθόδου και ύστερα έγινε συσχέτιση της λανθάνουσας κλάση με τις γνωστές κλάσεις των labeled δεδομένων. Η χρήση και των δύο τρόπων αντιμετώπισης ταυτόχρονα θεωρείται ότι δίνει αρκετά καλά αποτελέσματα. Το πρόβλημα της ssl μάθησης είναι περισσότερο πρακτικό και

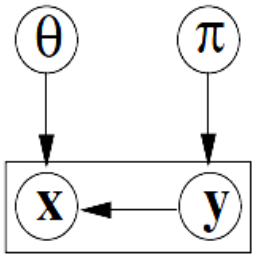
λιγότερο θεωρητικό και κάθε τρόπος επίλυσης θα πρέπει να λαμβάνει υπόψη τις λεπτομέρειες τους προβλήματος, όπως και στη μπεϋζιανή μάθηση.

Οι ssl μέθοδοι μπορούν να διαχωριστούν σε γεννητικούς και διαγνωστικούς αλγορίθμους, όπως και οι επιτηρούμενοι αλγόριθμοι. Ο διαχωρισμός είναι ίδιος, δεν είναι, όμως, απόλυτα όμοιος ως προς το τι θεωρείται γεννητικό και τι διαγνωστικό μοντέλο στις ssl μεθόδους. Στόχος των μεθόδων είναι ο υπολογισμός της P(y|x). Παραδοσιακά, τα γεννητικά μοντέλα το επιτυγχάνουν με προσδιορισμό της από κοινού πιθανότητας P(y,x) και μετά ταίριασμα αυτής στα δεδομένα με χρήση χαρακτηριστικών από την πραγματική από κοινού κατανομή. Μια εκτίμηση του P(x) μπορεί, πάντα, να προκύψει από περιθωριοποίηση της από κοινού εκτίμησης. Τα διαγνωστικά μοντέλα, από την άλλη μεριά, εστιάζουν στην εκτίμηση της δεσμευμένης πιθανότητας P(y|x) μόνο, και δεν προκύπτει κάποια εκτίμηση της P(x). Το τελευταίο είναι απαραίτητο για μια ssl μέθοδο, προκειμένου να εκτιμηθεί το διάνυσμα των unlabeled δεδομένων, για αυτό και πάντα γίνεται. Αναλόγως με το πότε και πως γίνεται η εκτίμηση της P(x) μπορούν, τελικά, να ταξινομηθούν στις δύο κατηγορίες οι ssl αλγόριθμοι. Η μπεϋζιανή στατιστική είναι η πλέον κατάλληλη για την περιγραφή και τον πλήρη ορισμό των δύο τρόπων αντιμετώπισης.

Η γραφική μέθοδος απεικόνισης των μοντέλων που φαίνεται στις εικόνες είναι ένας τρόπος συμβολισμού που χρησιμοποιείται συχνά στη στατιστική και την μηχανική μάθηση. Ονομάζονται *κατευθυνόμενα γραφικά μοντέλα ή διαγράμματα ανεξαρτησίας*. Οι κόμβοι των διαγραμμάτων αντιπροσωπεύουν τυχαίες μεταβλητές. Οι γονείς ενός κόμβου i είναι οι κόμβοι j για τους οποίους η ακμή j→i υπάρχει. Είναι δυνατόν να εκτιμηθεί και να χρησιμοποιεί ως δείγμα η τιμή ενός κόμβου, αν η είναι γνωστές οι τιμές όλων των γονέων του. Η δεσμευμένη δειγματοληψία από μια κατανομή ως προς συγκεκριμένες μεταβλητές μπορεί να παρασταθεί με τέτοια γραφήματα. Η δειγματοληψία ξεκινά από κόμβους χωρίς γονείς και συνεχίζεται σύμφωνα με την κατεύθυνση και φορά που υποδεικνύουν οι ακμές. Τα ορθογώνια κουτιά ή πιάτα *(plates)* υποδηλώνουν ένα σύνολο κόμβων. Έτσι, μπορεί να πραγματοποιηθεί δειγματοληψία στα πιάτα, με τρόπο επαναλαμβανόμενο και ανεξάρτητο, δεδομένου ότι η κατατομή των κόμβων που ανήκουν σε ένα πιάτο εξαρτάται από τους γονείς κόμβους, οι οποίοι είναι γονείς κάθε μέλους του πιάτου. Στην εικόνα 1.2, για παράδειγμα, γίνεται πρώτα δειγματοληψία στο θ και το π ανεξάρτητα, αφού κανένας από τους δύο κόμβους δεν έχει γονείς, και μετά πραγματοποιείται δειγματοληψία στο ορθογώνιο, όπου επιλέγεται ένα ζεύγος (xi, yi) δεδομένων των γονέων κόμβων θ και π.

1.4.1. Γεννητικό μοντέλο

Η γεννητική τεχνική ορίζει αρχικά τον υπολογισμό των δεσμευμένων κατανομών πιθανότητας P(x|y) ρητά, έτσι ώστε η εκτίμηση της P(x) να προκύπτει από κοινού από της δεσμευμένες αυτές πιθανότητες. Από τις εκτιμήσεις αυτές, και τις εκτιμήσεις της P(y), λαμβάνεται μια εκτίμηση για την P(y|x) με χρήση της φόρμουλας Bayes. Οι γεννητικές μέθοδοι χρησιμοποιούν το μοντέλο της εικόνας. Η κατανομή της κλάσης P(x|y) χρησιμοποιεί οικογένειες μοντέλων {P(x|y,θ)} και την εκ των προτέρων γνώση για την κλάση P(y) με $\pi_y = P(y|\pi),\ \pi = (\pi_y)_y$.



Εικόνα 1.2: Γεννητικό μοντέλο

Η διαδικασία που περιγράφεται στην εικόνα καλείται και *μοντέλο συλλογικής πυκνότητας (joint density model)*. Για κάθε παράμετρο $\hat{\theta}$ *και* $\hat{\pi}$ μια εκτίμηση της P (x | y) μπορεί να υπολογιστεί από τον τύπο Bayes για την δεσμευμένη πιθανότητα,

$$P\left(y|x,\hat{\theta},\hat{\pi}\right) = \frac{\hat{\pi}_y P\left(x|y,\hat{\theta}\right)}{\sum_{y'=1}^{M} \hat{\pi}_{y'}\, P(x|y',\hat{\theta}}. \quad (1.1)$$

Η εκτίμηση για το P(x) προκύπτει από

$$P(x|\theta,\pi) = \sum_{y=1}^{M} \pi_y P(x|y,\theta). \quad (1.2)$$

Εάν labeled και unlabeled δεδομένα είναι διαθέσιμα, προκύπτει ένα φυσικό κριτήριο, η από κοινού λογαριθμική πιθανοφάνεια των δύο

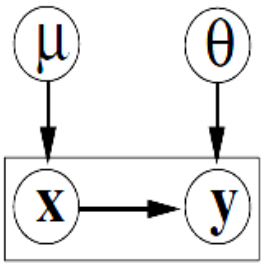
$$\sum_{i=1}^{n} \log \pi_{yi} P(x_i|y_i,\theta) + \sum_{i=n+1}^{n+m} log \sum_{y=1}^{M} \pi_y P(x_i|y,\theta). \quad (1.3)$$

Μέθοδοι που διατυπώθηκαν και αναλύθηκαν σύμφωνα με αυτή τη μέθοδο, κρατώντας αυτή τη βασική ιδέα και προσθέτοντας διάφορες βελτιώσεις είναι η

μέθοδος των Castelli και Cover το 1995, οι Titterington κ.α. το 1985, οι Shahshahani και Landgrebe το 1994, οι Nigam κ.α. το 2000 με εφαρμογή σε αναγνώριση κειμένου, οι Corduneanu και Jaakkola το 2002 και αρκετοί άλλοι. Ο καθένας στηρίχθηκε σε στατιστικά μοντέλα που βελτίωναν την απόδοση ή μελέτησε το πρόβλημα από διαφορές οπτικές. Η βασική ιδέα ήταν η γεννητική μέθοδος.

1.4.2. Διαγνωστική μέθοδος

Στο διαγνωστικό μοντέλο, η $P(y|x)$ προκύπτει κατευθείαν από την οικογένεια {P(y|x,θ)}. Για τον σχηματισμού του πλήρους μοντέλου δειγματοληψίας των δεδομένων θα πρέπει να μοντελοποιηθεί η P(x) από μια οικογένεια P(x|μ). Όμως, εάν ενδιαφερόμαστε μόνο για την ανανέωση της γνώσης σχετικά με το θ ή για την πρόβλεψη του y σε δεδομένα που δεν έχουμε δει, η P(x) δεν είναι απαραίτητη.



Εικόνα 1.3: Διαγνωστικό μοντέλο

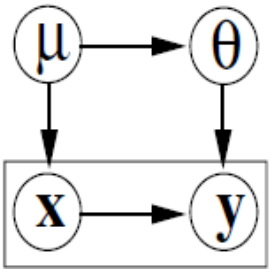
Στο παραπάνω διάγραμμα τα θ και μ είναι εκ των προτέρων γνωστά, ασυμβίβαστα ενδεχόμενα, δηλαδή ισχύει για αυτά P(θ,μ) = P(θ)P(μ). Η πιθανοφάνεια είναι η

$$P(D_l, D_u|\theta, \mu) = P(Y_l|X_l, \theta)P(X_l, D_u|\mu), \qquad (1.4)$$

όπου ο τύπος δείχνει ότι η $P(D_l, D_u|\theta, \mu) \propto P(Y_l|X_l, \theta)P(\theta)$ και τα θ και μ είναι ανεξάρτητα και εκ των υστέρων. Επίσης, $P(\theta|D_l, \mu) = P(\theta|D_l)$, δηλαδή ότι η γνώση των unlabeled δεδομένων D_u ή η γνώση για το μ, δεν αλλάζουν την εκ των προτέρων γνώση για την $P(\theta|D_l)$ των labeled δεδομένων. Η εξαγωγή συμπεράσματος για τα unlabeled δεδομένα, δε μπορεί να πραγματοποιηθεί με το συγκεκριμένο μοντέλο, αφού δεν προβλέπει την εκτίμηση της P(x), σύμφωνα με την οποία θα αποδοθούν ετικέτες στα άγνωστα δεδομένα.

Είναι αδύνατον, όμως, να εφαρμοστεί ssl μέθοδος χωρίς υπολογισμό της P(x). Απαιτείται, επομένως, μια τροποποίηση του παραπάνω μοντέλου, έτσι ώστε να υπολογίζεται και η P(x), και παρ' ότι δεν θα υπακούει πλήρως στο κλασικό μοντέλο διαγνωστικής διαδικασίας, θεωρείται διαγνωστική μέθοδος. Η εκ των προτέρων

ανεξαρτησία του μ και θ έχει ως αποτέλεσμα η οικογένεια {P(y|x,θ)} να είναι κανονικοποιημένη ανεξάρτητα από την κατανομή εισόδου. Για να γίνει χρήση των πρόσθετων μη επισημασμένων δεδομένων σε συνδυασμό με διαγνωστικές Bayesian επιτηρούμενες τεχνικές, θα πρέπει να υπεισέρθει μια εκ των προτέρων γνώση σχετικά με τα άγνωστα δεδομένα και την λανθάνουσα συνάρτηση που τα αντιπροσωπεύει. Αν επιτραπούν εκ των προτέρων εξαρτήσεις μεταξύ θ και μ, δηλαδή πληροφορίες για το μ να μεταφέρονται στο θ όπως στο σχήμα, τότε στον υπολογισμό των χαρακτηρισμών των unlabeled δεδομένων θα λάβει μέρος και η έκφραση P(θ,μ) = P(θ|μ)P(μ), με P(θ) = $\int$ P(θ|μ)P(μ) $d\mu$. Το διάγραμμα σε αυτήν την περίπτωση είναι

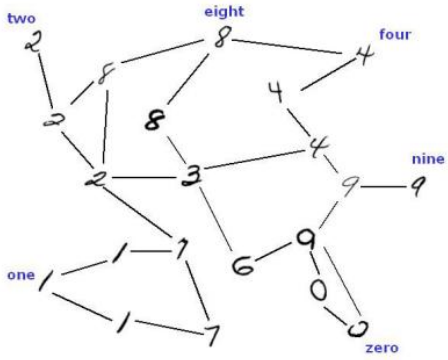


Εικόνα 1.4: Εμπλουτισμένο διαγνωστικό μοντέλο

Τα unlabeled δεδομένα μπορούν να αλλάξουν την εκ των προτέρων γνώση για το θ. Ουσιαστικά, εφαρμόζεται κανονικοποίηση στην λανθάνουσα συνάρτηση η οποία εξαρτάται από την εισερχόμενη κατανομή.

Εκφραστές της μεθόδου είναι οι Tong και Koller το 2000, οι Chapelle κ.α. το 2001, σύμφωνα με την παραδοχή της ομαδοποίησης και τις διάφορες εκφάνσεις που μπορεί να πάρει οι Seeger κ. α. το 2000, οι Szummer και Jaakola το 2002, οι αλγόριθμοι των οποίων η μέθοδος αναλύεται εκτενώς στην συνέχεια στην επόμενη ενότητα, αλγόριθμοι που χρησιμοποιούν τον πυρήνα Fisher, οι Co-training τεχνικές και αρκετές άλλες.[1][5]

2. Μέθοδοι Βασισμένες σε Γράφους



Οι μέθοδοι βασισμένες σε γράφο ξεκινούν με τη δημιουργία ενός γράφου εγγύτητας δεδομένων, του οποίου οι κόμβοι είναι τα ταξινομημένα και τα μη ταξινομημένα δεδομένα. Τα δεδομένα – κόμβοι ενώνονται με ακμές. Οι ακμές είναι σταθμισμένες, δηλώνοντας την ομοιότητα των κόμβων που ενώνουν. Οι κόμβοι που ενώνονται με ακμές μεγάλου βάρους είναι αρκετά πιθανόν να ταξινομούνται στην ίδια ομάδα, να πρέπει δηλαδή να λάβουν ίδια ετικέτα (χαρακτηρισμό). Οι ιδιότητες της θεωρίας γραφημάτων κληρονομούνται από τις συγκεκριμένες μεθόδους. Οι γραφικές μέθοδοι είναι, κατά κύριο λόγο, μη παραμετρικές, διακριτές και μεταγωγικές.

2.1. Κατηγορίες μεθόδων βασισμένων σε γράφους

Αναφέρονται συνοπτικά μερικές μέθοδοι απόδοσης ετικετών που έχουν χρησιμοποιηθεί στις γραφικές μεθόδους.

Οι Blum και Chawla το 2001 πρότειναν ένα είδος ssl αλγορίθμου γνωστό ως *st-cut ή mincut*. Στη δυαδική περίπτωση οι θετικές ετικέτες λειτουργούν ως πηγές και οι αρνητικές ως υποδοχείς (δέκτες). Οι κόμβοι που ενώνονται με πηγές χαρακτηρίζονται ως θετικοί και εκείνοι που ενώνονται με υποδοχείς ως αρνητικοί. Η συνάρτηση που επιδιώκεται να ελαχιστοποιηθεί στο yi είναι η

$$\infty \sum_{i \in L} (y_i - y_i|L)^2 + \frac{1}{2} \sum_{i,j} w_i, j \left(y_i - y_j^2\right). \quad (1.5)$$

Οι τιμές που αποδίδονται στις ετικέτες είναι 0 και 1. Το 2004 οι δημιουργοί του πρόσθεσαν μια διόρθωση, μέσω διαταραχών στο γράφο, ενώ βελτίωση της μεθόδου προτάθηκε και από τους Pang και Lee το 2004.

Οι Μηχανές Boltzman (Διακριτά Τυχαία Markov πεδία) αποτελούνν επίσης ssl μέθοδο. Υπολογίζονται οι οριακές πιθανότητες των τυχαίων πεδίων Markov. Εφαρμόστηκε από τους Zhu και Ghahramani το 2002 και από τους Getz κ.α. το 2005. Οι τελευταίοι βελτίωσαν τη μέθοδο δειγματοληψίας και πρότειναν διαδικασία με την οποία είναι δυνατή η εύρεση νέων κλάσεων στο δείγμα.

Τα συνεχή ανάλογα της παραπάνω κατηγορίας είναι Γκαουσιανά Τυχαία πεδία και Αρμονικές συναρτήσεις. Χαρακτηρίζονται από την ύπαρξη τετραγωνικής συνάρτησης απωλειών με άπειρο βάρος και από την κανονικοποίηση της συνάρτησης απόδοσης ετικετών με βάση τον συνδιασμένο Laplacian. Αναλύθηκε από τον Zhu, εκ νέου το 2005, ως προς τις ιδιότητες που μπορούν να προκύψουν από τη μέθοδο. Τέτοιες μέθοδοι εφαρμόστηκαν από τους Grady κ.α. σε ιατρικές εικόνες. Άλλοι ερευνητές τις εφάρμοσαν σε χρωματισμό ασπρόμαυρων εικόνων, σε αναγνώριση κειμένου και σε πρόβλεψη βαθμολογίας ταινιών.

Άλλος ένας τρόπος που χρησιμοποιήθηκε από τους Zhou κ.α. το 2004 είναι Τοπική και Ολική Σταθερότητα. Πέραν της συνάρτησης απωλειών, χρησιμοποιείται ο κανονικοποιημένος Laplacian για κανονικοποίηση της συνάρτησης απόδοσης ετικετών. Ο συγκεκριμένος αλγόριθμος αναλύεται εκτενώς στη συνέχεια. Επίσης από τους Belkin κ.α. έγινε χρήση της Tikhonov Κανονικοποίησης (με συνάρτηση απωλειών την $\frac{1}{k}\sum_i (f_i - y_i)^2 + \gamma f^T S f$) και της Manifold κανονικοποίησης. Αρκετά διαδεδομένη μέθοδος ssl μάθησης είναι και η χρήση πυρήνων και του φάσματος της Laplacian του γράφου. Η λίστα των μεθόδων είναι αρκετά εκτενής, δεν αποτελείται μόνο από όσες αναφέρθηκαν και συνεχώς αυξάνεται.

Οι μέθοδοι που υλοποιήθηκαν στην συγκεκριμένη εργασία επιδιώκουν να προσεγγίσουν με βέλτιστο τρόπο τα δεδομένα ενός γράφου δεδομένων, ικανοποιώντας δύο βασικές συνθήκες. Οι συνθήκες σχετίζονται με το κύριο τμήμα των μεθόδων, δηλαδή με τον τρόπο προσέγγισης του γράφου των δεδομένων. Η πρώτη συνθήκη λέει ότι η μέθοδος θα πρέπει να προσεγγίζει καλά, να μην αλλοιώνει, δηλαδή σε μεγάλο βαθμό, τις τιμές που έχουν αρχικά οι ετικέτες των ταξινομημένων δεδομένων, ενώ σύμφωνα με τη δεύτερη θα πρέπει η μέθοδος να είναι ομαλή σε όλο τον γράφο. Οι δύο αυτές συνθήκες είναι παρούσες στο σώμα των αλγορίθμων με την χρήση μιας συνάρτησης απωλειών και ενός τρόπου ομαλοποίησης της συνάρτησης.

2.2. Βασικοί ορισμοί

Ακολουθούν μερικοί ορισμοί, ιδιαίτερα χρήσιμοι για την πλήρη κατανόηση και ανάλυση των αλγορίθμων. Θα αναφερθούν βασικά σημεία της θεωρίας

γραφημάτων καθώς και ο ορισμός του λαπλασιανού τελεστή, ο οποίος επεκτείνεται αργότερα στον λαπλασιανό τελεστή του γράφου.

2.2.1. Τελεστής Laplace Διανυσματικής Συνάρτησης

Ορισμός: Έστω $\vec{F}: A \to \mathbb{R}^3$ μια διανυσματική συνάρτηση τριών ανεξάρτητων πραγματικών μεταβλητών x, y, z με πεδίο ορισμού το ανοικτό υποσύνολο A του $\mathbb{R}^3$, η οποία ορίζεται με τύπο

$$\vec{F}(x,y,z) = F_1(x,y,z)\vec{\imath} + F_2(x,y,z)\vec{\jmath}F_3(x,y,z)\vec{k}. \quad (2.1)$$

Οι συναρτήσεις F_1, F_2, F_3 είναι συνεχείς και έχουν συνεχείς μερικές παραγώγους δεύτερης τάξης στο A. Ο τελεστής Laplace για τη διανυσματική συνάρτηση του $\vec{F}$ συμβολίζεται με $\nabla^2\vec{F}$ ή $\Delta\vec{F}$ και ορίζεται ως

$$\nabla^2\vec{F} = \nabla^2\left(F_1(x,y,z)\vec{\imath} + F_2(x,y,z)\vec{\jmath}F_3(x,y,z)\vec{k}\right) \quad (2.2)$$

Ο τελεστής Laplace για τη διανυσματική συνάρτηση $\vec{F}$ ονομάζεται επίσης λαπλασιανή της $\vec{F}$.

2.2.2. Θεωρία γραφημάτων

Ορισμός: Ο *γράφος ή το γράφημα* των δεδομένων αντιπροσωπεύει την γεωμετρία των δεδομένων. Συμβολίζεται με $g = (V,E)$, όπου οι *κόμβοι ή κορυφές* V = 1,2,…,n είναι τα δεδομένα (μη κενό σύνολο στοιχείων), και E ένα σύνολο μη διατεταγμένων ζευγών από ακμές ή πλευρές, οι οποίες υποδηλώνουν την ομοιότητα των δεδομένων.

Οι πλευρές συμβολίζονται και με e = {u,v}, με u και v δύο κορυφές του γράφου, οι όποιες λέγονται και άκρα της πλευράς e. Δύο πλευρές που προσπίπτουν στην ίδια κορυφή είναι *γειτονικές (adjacent)*.

Ο αριθμός των κορυφών ενός γράφου g = (V,E) ονομάζεται *τάξη* του g και συμβολίζεται με $|V|$. Ο αριθμός των πλευρών του γράφου ονομάζεται *μέγεθος* και συμβολίζεται με $|E|$, αν και πολλές φορές, κυρίως στην Πληροφορική, μέγεθος του γράφου ονομάζεται ο αριθμός των κορυφών.

Ορισμός: *Βαθμός* μιας κορυφής u ∈ V ονομάζεται ο αριθμός των πλευρών του g που προσπίπτουν στην u και συμβολίζεται με d(u). Ένας γράφος για τον οποίο ισχύει d(u) = k για κάθε κορυφή του, λέγεται k-κανονικός γράφος. Εάν ο βαθμός ενός κόμβου είναι μηδέν, τότε ο κόμβος είναι *απομονωμένος*.

Δύο κορυφές u και v του γράφου είναι συνδεδεμένες, αν υπάρχει τουλάχιστον ένα μονοπάτι uv. Ένας μη κατευθυνόμενος γράφος, είναι ένας γράφος για τον οποίο ισχύει (u, v) αν και μόνο αν (v, u). Αν στον ορισμό ενός γράφου αντικατασταθούν τα στοιχεία του E με διατεταγμένα ζεύγη στοιχείων του V, λαμβάνεται ένας *προσανατολισμένος* ή *κατευθυνόμενος γράφος (directed graph, digraph)*, δηλαδή $E \subseteq V \times V$.

Ορισμός: Ένας γράφος λέγεται *συνεκτικός* αν για κάθε ζεύγος κορυφών του γράφου υπάρχει ένα μονοπάτι που τις συνδέει.

Ορισμός: Έστω ένας γράφος g΄ = (V΄, E΄). Ο g΄ είναι *υπογράφος* του g εάν ισχύει και V' $\subseteq V$ και $E' \subseteq E$.

Σε ένα γράφο, μια πεπερασμένη ακολουθία εναλλάξ κορυφών και πλευρών του, που αρχίζει και τελειώνει σε κορυφή και που κάθε πλευρά που περιέχεται στην ακολουθία προσπίπτει στη κορυφή που προηγείται και σε αυτή που έπεται, λέγεται *δρόμος ή διαδρομή (walk)* του g. Αν σε έναν δρόμο κάθε πλευρά του δρόμου εμφανίζεται μόνο μία φορά, ο δρόμος λέγεται *δρομίσκος ή μονοπάτι* (trail). Ένας δρόμος με αρχή και τέλος στην ίδια κορυφή, λέγεται κλειστός δρόμος, αλλιώς λέγεται ανοιχτός. Ένας δρόμος που είναι κλειστό μονοπάτι λέγεται κύκλος.

Ορισμός: Δένδρο ονομάζεται ένας συνεκτικός γράφος που δεν περιέχει κύκλους. Κάθε γράφος που δεν περιέχει κύκλους ονομάζεται δάσος. Ισχύει ότι, ένας γράφος είναι δάσος αν είναι υπογράφος ενός δένδρου.

2.2.3. Πίνακας γειτνίασης

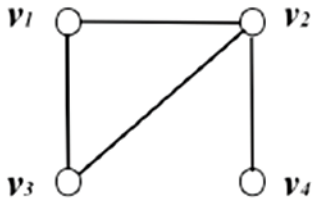
Η μέτρηση της ομοιότητας των δεδομένων γίνεται με χρήση του πίνακα γειτνίασης. *Ο πίνακας γειτνίασης (adjacency matrix)* ή πίνακας βάρους ή πίνακας συνάφειας, μπορεί να έχει διάφορες μορφές. Αν i και j δύο κόμβοι του γράφου, σκοπός της συνάρτησης βάρους είναι η κατασκευή ενός πίνακα συνάφειας ή βάρους Wij, για κάθε ζεύγος κόμβων του γράφου.

Μια μορφή που μπορεί να λάβει είναι να παίρνει την τιμή ένα, εάν δύο κόμβοι γειτονεύουν ή την τιμή μηδέν αλλιώς, δηλαδή

$$Wij = \begin{cases} 1, & \text{εάν } xi, xj \text{ γείτονες} \\ 0, & \text{εάν } xi, xj \text{ όχι γείτονες} \end{cases} \quad (2.3)$$

μορφή που στην ουσία αποτελεί μια ειδική περίπτωση, μιας και τα βάρη παίρνουν δύο συγκεκριμένες τιμές. Στην εικόνα φαίνεται ένα παράδειγμα κατασκευής πίνακα γειτνίασης.

$$W = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$



Εικόνα 2.1: Πίνακας γειτνίασης και αντίστοιχος γράφος

Ένας διαφορετικός ορισμός του πίνακα γειτνίασης είναι, να λαμβάνουν τα στοιχεία του την τιμή ένα, εάν ο i κόμβος ανήκει στους k κοντινότερους γείτονες, ή την τιμή μηδέν αλλιώς,

$$Wij = \begin{cases} 1, & \text{εάν } xi \in k - \text{κοντινότερους γείτονες του } xj \\ 0, & \text{αλλιώς} \end{cases} \quad (2.4)$$

Ο γράφος που κατασκευάζεται συνήθως από τον συγκεκριμένο ορισμό είναι κατευθυνόμενος, δηλαδή μη συμμετρικός, αν και με αμοιβαία ένωση των k – πλησιέστερων γειτόνων προκύπτει συμμετρικός πίνακας.

Μια γενικευμένη μορφή του πίνακα W δίνεται αν οριστεί w(e), με e να δηλώνει την πλευρά ενός γράφου, ως

$$Wij = \begin{cases} \text{w(e)}, & \text{εάν } e = (i,j) \in E \\ 0, & \text{αλλιώς} \end{cases} \quad (2.5)$$

όπου το e είναι και η ελάχιστη απόσταση που μπορεί να έχουν τα σημεία. Το αποτέλεσμα είναι ένας συνεκτικός γράφος, αλλά κάποια σημεία του γράφου μένουν ασύνδετα.

Τέλος, ο πίνακας γειτνίασης μπορεί να υπολογιστεί με χρήση του γκαουσιανού πυρήνα, διασποράς σ,

$$Wij = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}. \quad (2.6)$$

Ο συγκεκριμένος τρόπος σύνδεσης των δεδομένων ενώνει όλα τα σημεία μεταξύ τους, κατασκευάζοντας τελικά έναν πλήρη γράφο. Η τιμή που λαμβάνεται για παρόμοια δεδομένα αναμένεται μεγάλη, και όσο μικρότερη είναι η τιμή αυτή, τόσο τα δεδομένα δεν μοιάζουν μεταξύ τους.

Ο τελευταίος τρόπος ορισμού χρησιμοποιήθηκε για την υλοποίηση των αλγορίθμων που θα αναλυθούν στη συνέχεια. Ο πίνακας συνάφειας υπολογίζεται αρχικά, και είναι βασική μεταβλητή στο σώμα των μεθόδων που υλοποιήθηκαν. Η Ευκλείδεια απόσταση για κάθε ζεύγος δεδομένων υπολογίζεται και, σε συνδυασμό με την τιμή της διασποράς των δεδομένων, κατασκευάζεται ο γράφος. Τα δεδομένα, αφού κανονικοποιηθούν, εισέρχονται σε μέθοδο που υπολογίζει τον πίνακα γειτνίασης, με βάση την παραπάνω σχέση. Ο πίνακας γειτνίασης μαζί με τις επιμέρους μεταβλητές κάθε αλγορίθμου οδηγούν στο αποτέλεσμα. Να σημειωθεί ότι οι τρεις παραπάνω τρόποι υπολογισμού του πίνακα (ο οποίος ουσιαστικά αποτελεί τον γράφο) δεν είναι οι μόνοι. Ο τρόπος κατασκευής ολοένα και καταλληλότερων γράφων είναι ένα ανοιχτό πρόβλημα.

2.2.4. Λαπλασιανός τελεστής πίνακα

Για την υλοποίηση των αλγορίθμων ορίζεται ο πίνακας D, που είναι ο βαθμός *(degree matrix)* του πίνακα συνδεσιμότητας W, τέτοιος ώστε $\mathrm{Dii} = \sum_j Wij = di.$ Ο υπολογισμός του βαθμού του γράφου είναι το άθροισμα όλων των στοιχείων των στηλών της κάθε γραμμής του πίνακα συνδεσιμότητας. Ο πίνακας συνδεσιμότητας θεωρείται ότι δεν έχει κύκλους ή πολλαπλές κορυφές. Ο υπολογισμός του συγκεκριμένου διανύσματος υπεισέρχεται στον υπολογισμό του πίνακα Laplacian ή λαπλασιανού πίνακα του γράφου, ο οποίος συναντάτε σε δύο μορφές, τον κανονικοποιημένο και τον μη κανονικοποιημένο. Πρόκειται για έναν πίνακα που εκφράζει τη σύνδεση των δεδομένων και τη μετάβαση από το ένα στο άλλο. Είναι πολύ βασικό χαρακτηριστικό, και χρησιμοποιείται για την τελική ανάθεση ετικετών στα δεδομένα. Ουσιαστικά αποτελεί ένα διακριτό ανάλογο του διαφορικού τελεστή Laplace.

Ο πίνακας Laplacian ενός γραφήματος και οι ιδιοτιμές του χρησιμοποιούνται σε διάφορους τομείς των μαθηματικών, της φυσικής και της χημείας. Οι πρώτες αναφορές σχετικά με την σημασία του συγκεκριμένου πίνακα εμφανίζονται σε μια δημοσίευση του Kirchhoff[2] το 1847, σχετικά με τον ηλεκτρισμό, για αυτό το λόγο ονομάζεται αλλιώς και πίνακας Kirchhoff. Λεπτομερής ανάλυσή του έγινε μεταγενέστερα σε κείμενα σχετικά με την *φασματική θεωρία γραφημάτων (spectral graph theory)*, αφού κυρίως παρουσιάζουν ενδιαφέρουν οι ιδιότητες των ιδιοτιμών και των ιδιοσυναρτήσεών του. Η φασματική θεωρία γραφημάτων, και οι ιδιότητες που προκύπτουν από την ανάλυση των γράφων δεδομένων και των ιδιοτιμών τους,

[2] Gustav Robert Kirchhoff, (1824 – 1887), Γερμανός φυσικός με μεγάλη συνεισφορά στη μελέτη των ηλεκτρικών κυκλωμάτων, της ακτινοβολίας του μέλανος σώματος και στη φασματοσκοπία.

βρίσκει αρκετές εφαρμογές, όπως στην τμηματοποίηση εικόνας, στην εξόρυξη κειμένου και σε εφαρμογές web (collaborative recommendation, κατηγοριοποίηση κειμένου), ανάλυση Manifold και άλλες. Την τελευταία δεκαετία πληθώρα δημοσιεύσεων αναφέρονται σε επίλυση τέτοιων προβλημάτων μέσω της φασματικής θεωρίας γραφημάτων.

Ο ακριβής ορισμός του Laplacian θα ακολουθήσει στη συνέχεια. Για να οριστεί ο τελεστής Laplace με τρόπο ανάλογο των συνεχών συναρτήσεων θα αναφερθούμε πρώτα στην έννοια της κλίσης μιας συνάρτησης επί των κόμβων του γράφου. Θεωρήσουμε μια πραγματική συνάρτηση f :V $\rightarrow \mathbb{R}$. Η συνάρτηση αυτή αναθέτει έναν πραγματικό ακέραιο αριθμό σε κάθε κόμβο του γράφου. Η f είναι ένα διάνυσμα αριθμημένο με τους κόμβους του γράφου, είναι δηλαδή f $\in \mathbb{R}^n$. Επίσης, ορίζεται η *μήτρα πρόσπτωσης (incidence matrix)* ενός γράφου, η οποία είναι ένας $|E| \times |V|$ $(m \times n)$πίνακας, ως

$$\nabla := \begin{cases} \nabla\, e(v) = -1 & \text{εάν το } v \text{ είναι ο κόμβος αρχής της ακμής } e \\ \nabla e(v) = 1 & \text{εάν το } v \text{ είναι ο κόμβος πέρατος της ακμής } e \\ \nabla e(v) = 0 & \text{εάν ο } v \text{ δεν ανήκει στην ακμή } e. \end{cases} \quad (2.7)$$

Η *συνοριακή χαρτογράφηση* (co – boundary) του γράφου ορίζεται ως η κλίση ή grad της f,

$$(\nabla f)(\mathrm{e}) = \mathrm{f}(u) - \mathrm{f}(\mathrm{v}) \quad (2.8)$$

με u, v τους κόμβους της ακμής e. Επί του γράφου, η κλίση μεταφράζεται σε μέτρηση της αλλαγής της συνάρτησης από ακμή σε ακμή.

Ο λαπλασιανός τελεστής ορίστηκε προηγουμένως. Η δράση του στην συνάρτηση f δίνεται από τη σχέση,

$$(Lf)(v) = \sum_{u \sim v} \big(f(u) - f(v)\big), \quad (2.9)$$

όπου $u \sim v$ δηλώνει ότι οι κόμβοι u και v είναι γειτονικοί. . Η τετραγωνική μορφή του λαπλασιανού τελεστή είναι $L = \nabla^{\mathrm{T}}\nabla$ και σε αυτή την περίπτωση η δράση του είναι

$$f^T L f = \frac{1}{2}\sum_{u \sim v} \big(f(u) - f(v)\big)^2 \quad (2.10)$$

Αναλυτικότερα, σε ένα γράφο g, έστω d(u) ο βαθμός του κόμβου u. Εάν W είναι ο πίνακας γειτνίασης, ο οποίος περιέχει τιμές από μηδέν έως ένα, τότε ορίζεται ο *μη κανονικοποιημένος Laplacian* (λαπλασιανός) πίνακας του γράφου

$$L\ (g) = D(g) - W(g) \quad (2.11)$$

με D τον διαγώνιο πίνακα των βαθμών, των κόμβων του γράφου και W ο πίνακας γειτνίασης του γράφου χωρίς βάρη. Ο L(g) είναι θετικός, ημιορισμένος και μονοσήμαντος M-πίνακας[3]. Πρόκειται, ουσιαστικά για έναν τελεστή, με μια ενδιαφέρουσα δράση επί του γράφου δεδομένων.

Οι ιδιοτιμές του L ονομάζονται Laplacian ιδιοτιμές ή απλά ιδιοτιμές του g και αν τις συμβολίσουμε με $\lambda_1, \lambda_2, \ldots, \lambda_n$ τότε ισχύει

$$0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n .$$

Οι ιδιοτιμές του είναι πραγματικές, θετικές και καλούνται φάσμα του L ή φάσμα του σχετιζόμενου γράφου g. Βασική ιδιότητα των ιδιοσυναρτήσεων του Laplacian είναι ότι όσο πιο μικρής τάξης είναι τόσο μεγαλύτερη ομαλότητα παρουσιάζουν σχετικά με τον γράφο. Αυτό σημαίνει είναι πιο αντιπροσωπευτικές ως προς τα χαρακτηριστικά του γράφου και άρα πιο σημαντικές. Επίσης, για παρόμοιους γράφους, οι διαφορές τους είναι εμφανείς σε συναρτήσεις μεγαλύτερης τάξης. Οι ιδιοσυναρτήσεις, αποτελούν ορθοκανονική βάση, που μπορεί να προσεγγίσει κάθε τετραγωνικά ολοκληρώσιμη συνάρτηση στο γράφο. Δηλαδή, κάθε συνάρτηση στο g μπορεί να αναλυθεί ως άθροισμα από ιδιοσυναρτήσεις του πίνακα L. Από τον ορισμό του πίνακα προκύπτει ότι

$$L(u,v) = \begin{cases} d(u) & \text{εάν } u = v \\ -1 & \text{εάν } u \text{ και } v \text{ είναι γειτονικά.} \\ 0 & \text{αλλιώς} \end{cases} \quad (2.12)$$

Ο *κανονικοποιημένος Laplacian* του g ορίζεται ως

$$\mathcal{L}(u,v) = \begin{cases} 1 & \text{εάν } u = v \text{ και } d(u) \neq 0 \\ -\dfrac{1}{\sqrt{d(u)d(v)}} & \text{εάν } u \text{ και } v \text{ είναι γειτονικά.} \\ 0 & \text{αλλιώς} \end{cases} \quad (2.13)$$

Αν ο πίνακας γειτνίασης W έχει υπολογιστεί, και ο γράφος δεν περιέχει απομονωμένους κόμβους, τότε ο $\mathcal{L}$ γράφετε

[3] Τα πραγματικά μέρη των ιδιοτιμών του είναι θετικά.

$$\mathcal{L} = I - D^{-1/2}\, W\ D^{-1/2}\ = D^{-1/2}\, L(g) D^{-\frac{1}{2}}. \quad (2.14)$$

Αποδεικνύεται ότι η ελαχιστοποίηση της ποσότητας $\sum_{u\sim v}\big(f(u) - f(v)\big)^2$ ή απλούστερα, αλλά ομοίως, της $\sum_{u\sim v}\big(f(u) - f(v)\big)$, είναι ανάλογο πρόβλημα με την εύρεση ιδιοτιμών του L[4], δηλαδή

$$Lf = \lambda f \ (2.15)$$

με λ τις ιδιοτιμές και f τις ιδιοσυναρτήσεις. Η ποσότητα $\sum_{u\sim v}\big(f(u) - f(v)\big)^2$ονομάζεται και άθροισμα Dirichlet του γράφου. Για την εύρεση ιδιοτιμών του $\mathcal{L}$ ορίζεται συνάρτηση g = $D^{-1/2}$f, με g τις ιδιοσυναρτήσεις του $\mathcal{L}$, και η διαδικασία υπολογισμού ιδιοσυναρήσεων και ιδιοτιμών είναι ίδια με του L.

Στην περίπτωση γράφου με βάρη, δηλαδή ενός γράφου που εμπεριέχει και μια συνάρτηση w:$V \times V \to \mathbb{R}$ η οποία ικανοποιεί τη σχέση $w(u,v) = w(v,u)$ και $w(u,v) \geq 0$, οι παραπάνω σχέσεις γίνονται,

$$L(u,v) = \begin{cases} d(u) - w(u,u) & \varepsilon\acute{\alpha}\nu\ u = v \\ -w(u,v) & \varepsilon\acute{\alpha}\nu\ u\ \kappa\alpha\iota\ v\ \varepsilon\acute{\iota}\nu\alpha\iota\ \gamma\varepsilon\iota\tau o\nu\iota\kappa\acute{\alpha} \\ 0 & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (2.16)$$

και, προφανώς, για μια πραγματική συνάρτηση που εφαρμόζεται στους κόμβους του γράφου ισχύει

$$(Lf)(v) = \sum_{u\sim v} (\mathrm{f(u)} - \mathrm{f(v)} w(u,v) \quad (2.17)$$

ή σε τετραγωνική μορφή

$$f^T Lf = \frac{1}{2}\sum_{u\sim v} w(u,v)\big(f(u) - f(v)\big)^2 \quad (2.18)$$

Ομοίως, ο μη κανονικοποιημένος Laplacian γίνεται

$$\mathcal{L}(u,v) = \begin{cases} 1 - \dfrac{w(u,u)}{d(u)} & \varepsilon\acute{\alpha}\nu\ u = v\ \kappa\alpha\iota\ d(u) \neq 0 \\ -\dfrac{w(u,v)}{\sqrt{d(u)d(v)}} & \varepsilon\acute{\alpha}\nu\ u\ \kappa\alpha\iota\ v\ \varepsilon\acute{\iota}\nu\alpha\iota\ \gamma\varepsilon\iota\tau o\nu\iota\kappa\acute{\alpha} \\ 0 & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \quad (2.19)$$

[4] Chung, 1997, κεφάλαιο 1°

Εάν με D συμβολίσουμε τον διαγώνιο πίνακα με το στοιχείο (u,u) να έχει τιμή d(u), τότε

$$\mathcal{L} = D^{-1/2}\, L(g) D^{-1/2}. \quad (2.20)$$

Είναι προφανές, ότι οι αβαρείς γράφοι είναι ειδική περίπτωση των γράφων με βάρος, όπου όλα τα βάρη είναι μηδέν ή ένα. Στις τελευταίες σχέσεις ο βαθμός ενός κόμβου είναι $d(v) = \sum_u w(u, v)$.

2.3. Θεωρητική θεμελίωση αλγορίθμων

Το πρόβλημα που επιλύουν οι υλοποιημένοι αλγόριθμοι είναι διακριτό, καθώς καλούνται να χαρακτηρίσουν τα δεδομένα με ετικέτες -1 ή 1. Δεδομένου ενός γράφου g = (V,E), ενός συνόλου ετικετών Y = {-1, 1} και οι κόμβοι v σε ένα υποσύνολο S ⊂ V έχουν λάβει ετικέτες y(v)∈ Y, επιδιώκουμε να χαρακτηρίσουμε με αντίστοιχες ετικέτες και τους υπόλοιπους κόμβους που ανήκουν στο συμπλήρωμα του S. Μια καλή συνάρτηση κατηγοριοποίησης, θα πρέπει να αλλάζει όσο πιο αργά γίνεται για συγγενή δεδομένα – κόμβους και προφανώς να αλλάζει όσο το δυνατόν λιγότερο τις αρχικές ετικέτες των δεδομένων. Τα δύο αυτά ζητήματα θα πρέπει να ικανοποιούν οι αλγόριθμοι.

Σε κάθε εφαρμογή του Laplacian, ο συγκεκριμένος τελεστής μπορεί να οριστεί με λίγο διαφορετικό, αλλά, σχεδόν πάντα, ανάλογο τρόπο με την προηγούμενη ανάλυση του τελεστή. Στην βιβλιογραφία εμφανίζονται διάφορες εκφράσεις του τελεστή, ανάλογα με το προς επίλυση πρόβλημα. Για την ανάπτυξη των αλγορίθμων θεωρήθηκε ο τελεστής ως ανάλογο του Laplace – Beltrami[5] τελεστή σε Riemannian manifolds. Για την ακρίβεια ορίστηκε ως

$$\Delta f \coloneqq -\frac{1}{2} div(\nabla f). \quad (2.21)$$

Ο τελεστής αυτός αποδεικνύεται ότι δίνει Laplacian γράφο αντίστοιχο του ορισμού της προηγούμενης ενότητας. Δηλαδή ισχύει

$$(\nabla f)([u, v]) \coloneqq \sqrt{w([u, v])}(f(v) - f(v), για\ κάθε\ κόμβο\ \epsilon\ E. \quad (2.22)$$

[5] Ο τελεστής Laplace – Beltrami Δ ορίζεται ως Δf = -div(∇f). Το ½ στην συγκεκριμένη έκφραση προκύπτει από το γεγονός ότι δε θέλουμε κάθε ακμή να μετρηθεί δύο φορές.

Επίσης, ορίζεται η γραφική κλίση σε κάθε κόμβο. Δοσμένης μιας συνάρτησης f του χώρου Hilbert[6] η κλίση της συνάρτησης είναι $\nabla f(u) = \{(\nabla f)([v,u]), για\ κάθε\ κόμβο\ \epsilon\ E$. Η νόρμα αυτής της κλίσης είναι $\|\nabla f(u)\| := (\sum_{u\sim v}(\nabla f)^2([u,v]))^{\frac{1}{2}}$. Η p – Dirichlet μορφή της συνάρτησης είναι

$$S_p(f) := \frac{1}{2}\sum_{v\epsilon V}\|\nabla f(u)\|^p. \quad (2.23)$$

Η τελευταία δηλώνει την ομαλότητα της f πάνω στον γράφο. Θεωρώντας τώρα μια συνάρτηση f τέτοια ώστε y (v) = 1 ή -1 για δεδομένα με ετικέτες και μηδέν για δεδομένα χωρίς ετικέτες η βελτιστοποίηση του προβλήματος

$$f = argmin\{S_p(f) + \mu\ \|f - y\|^2\} \quad (2.24)$$

οδηγεί τελικά στην πλήρη θεωρητική θεμελίωση των αλγορίθμων που θα υλοποιηθούν στη συνέχεια. Ο πρώτος όρος ονομάζεται *όρος ομαλότητας* (smoothness term) και απαιτεί η συνάρτηση να είναι όσο πιο ομαλή γίνεται. Ο δεύτερος όρος είναι ο *όρος συμφωνίας* (fitting term) και απαιτεί η συνάρτηση να είναι όσο το δυνατόν πιο σύμφωνη με τις αρχικές ετικέτες των δεδομένων. Το μ είναι παράμετρος που αφορά τα δύο ζητήματα που πρέπει να ικανοποιηθούν.
Η κανονικοποίηση της 2.23 για p = 2 εκφράζεται ως

$$E(f) = \frac{1}{2}\sum_{u,v} w([u,v])\big(f(u) - f(v)\big)^2. \quad (2.25)$$

Η σχέση ονομάζεται και τετραγωνική συνάρτηση ενέργειας.
Θεώρημα: Η λύση της (2.24) ικανοποιεί τη σχέση Δf + μ (f – y) = 0

Η εξίσωση του θεωρήματος μπορεί να θεωρηθεί ως ένα διακριτό ανάλογο της εξίσωσης Euler – Lagrange[7]. Μια λύση κλειστής μορφής είναι η $f = \mu(\Delta + \mu I)^{-1}$. Δεδομένης μιας συνάρτησης c που εξαρτάται από τις ακμές και τα βάρη των ακμών, και η οποία είναι ίση με μ/(1+μ) για ακμή από τον κόμβο v στον κόμβο v, αποδεικνύεται ότι η επανάληψη

$$f^{t+1}(v) = \sum_{u\sim v} c([u,v])f^t(u) + c(u,v)y(v) \quad (2.26)$$

[6] Διανυσματικός χώρος εφοδιασμένος με εσωτερικό γινόμενο.
[7] Διαφορική εξίσωση για τις λύσεις της οποίας ένα δεδομένο συναρτησοειδές παρουσιάζει ακρότατο.

σε όλους τους κόμβους, συγκλίνει σε μια λύση κλειστής μορφής. Σε κάθε βήμα της επανάληψης σε κάθε κόμβο του γνωστοποιείται η ετικέτα των γειτονικών του κόμβων και μέσω αυτής της πληροφορίας θα ληφθεί η απόφαση για την δική του ετικέτα, ενώ λαμβάνεται υπόψη και η προηγούμενη πληροφορία του κόμβου.[1][2][3][4]

Στη βιβλιογραφία βρέθηκαν, και εφαρμόστηκαν με κατάλληλο τρόπο, σε πίνακες δεδομένων οι τέσσερις αλγόριθμοι που θα αναλυθούν παρακάτω. Πρόκειται για τρεις επαναληπτικούς αλγορίθμους και έναν μη επαναληπτικό, οι οποίοι αποδίδουν ετικέτες σύμφωνα με την ελαχιστοποίηση συγκεκριμένων συναρτήσεων κόστους. Πριν την πλήρη ανάλυση και εξήγηση των αλγορίθμων θα γίνει αναφορά στους Τυχαίους Περιπάτους Markov.

2.4. Τυχαίοι περίπατοι Markov – Ηλεκτρικό ανάλογο

Θα ήταν παράληψη να μην γίνει αναφορά στους *τυχαίους περίπατους Markov*[8] (Markov Random Walks). Οι τυχαίοι περίπατοι είναι η μεθοδολογία πάνω στην οποία στηρίχθηκαν οι αλγόριθμοι της επόμενης ενότητας. Δεδομένου ενός γράφου και ενός κόμβου έναρξης, επιλέγεται με τυχαίο τρόπο ένας γείτονας του αρχικού κόμβου και γίνεται μεταπήδηση σε αυτόν. Στη συνέχεια, επιλέγεται ξανά ένας γείτονας με τυχαίο τρόπο και γίνεται μετακίνηση σε αυτόν. Η τυχαία αυτή ακολουθία των διαδοχικών κόμβων είναι ένας *τυχαίος περίπατος (random walk)* σε έναν γράφο. Ένας τυχαίος περίπατος είναι μια ορισμένη, χρονικά αναστρέψιμη *μαρκοβιανή αλυσίδα (Markov chain)*.

Έστω g = (V,E) ένας συνεκτικός γράφος με n κόμβους και m κορυφές και, έστω, ένας τυχαίος περίπατος στον g με σημείο έναρξης τον κόμβο u_o. Εάν μετά από t βήματα βρισκόμαστε τον κόμβο u_t (με u_t : t = 0,1,…), μετακινούμαστε στον γείτονα του u_t με πιθανότητα $1/d(u_t)$. Η ακολουθία των τυχαίων κόμβων είναι μια μαρκοβιανή αλυσίδα. Ο κόμβος έναρξης μπορεί να είναι συγκεκριμένος ή να ανήκει σε μια κάποια αρχική κατανομή Po. Η κατανομή P_t του u_t είναι

$$P_t(i) = Prob\ (v_t = i) \quad (2.26)$$

και δηλώνει την κατανομή μετά από t βήματα.

8 Andrey Andreyevich Markov (14 Ιουνίου 1856 – 20 Ιουλίου 1922), Ρώσος μαθηματικός με σπουδαία συνεισφορά στις Στοχαστικές Διαδικασίες. Προς τιμήν του δόθηκε η ονομασία μαρκοβιανές αλυσίδες και μαρκοβιανές διαδικασίες σε αντίστοιχα θέματα που είχε μελετήσει.

Θεωρούμε $M = (p_{ij})$ με $i, j \in V$ τον πίνακα πιθανοτήτων μετάβασης αυτής της μαρκοβιανής αλυσίδας, δηλαδή

$$p_{ij} = \begin{cases} \frac{1}{d(i)}, & \varepsilon\acute{\alpha}\nu\ i, j \in V \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma. \end{cases} \quad (2.27)$$

Θεωρούμε επίσης A τον πίνακα γειτνίασης του γράφου g και D τον διαγώνιο πίνακα με $(D)_{ii} = 1/d(i)$, τότε M = DA_G. Εάν ο g είναι d – κανονικός, τότε M = (1/d) A_G. Ο κανόνας του περιπάτου είναι

$$P_{t+1} = M^T P_t, \quad (2.28)$$

και άρα

$$P_t = M^T P_0. \quad (2.29)$$

Η κατανομή του t κατά σειρά σημείου μπορεί να θεωρηθεί ως ένα διάνυσμα στον $\mathbb{R}^V$. Το σημείο i.j του πίνακας M^T είναι η πιθανότητα $p_{i,j}^t$ να ολοκληρωθεί η διαδρομή I,j, με σημείο αρχής το i και πέρατος το j, σε t βήματα.

Εάν ο γράφος είναι κανονικός, τότε η μαρκοβιανή του αλυσίδα είναι *συμμετρική*, δηλαδή η πιθανότητα για μετακινηθούμε στον κόμβο u, δεδομένου ότι είμαστε στο κόμβο υ, είναι ίδια με την πιθανότητα να μετακινηθούμε στον κόμβο υ, ενώ είμαστε στον u. Για μη κανονικούς γράφους αυτή η ιδιότητα αντικαθιστάτε από *χρονική αντιστρεψιμότητα*, η οποία σημαίνει ένας τυχαίος περίπατος αντίθετης φοράς είναι και αυτός τυχαίος περίπατος. Η κατανομή πιθανότητας που λαμβάνεται από τον περίπατο αντίθετης φοράς είναι ίδια με την κατανομή πιθανότητας που λαμβάνεται εάν ο τυχαίος περίπατος ξεκινήσει από την κατανομή P_t.

Η κατανομές πιθανότητας $P_0, P_1, \ldots$ είναι διαφορετικές γενικά. Εάν ισχύει $P_0 = P_1$ για τον γράφο g, η κατανομή P_0 είναι *στατική* ή *σταθερής κατάστασης*. Προφανώς, σε αυτή την περίπτωση $P_0 = P_t$ για όλα τα t ≥ 0. Ο περίπατος που αντιστοιχεί από αυτήν την κατάσταση καλείται στατικός περίπατος. Για κάθε γράφο g, η κατανομή

$$\pi(\upsilon) = \frac{d(\upsilon)}{2m} \quad (2.30)$$

είναι στατική. Αποδεικνύεται ότι η στατική κατανομή είναι μοναδική. Με δεδομένη την στατική κατανομή η αρχή της χρονικής αντιστροφής ορίζεται για κάθε ζεύγος κόμβων I, j ∈ V $\pi(i)p_{ij} = \pi(j)p_{ji}$.

Η σύνδεση με τις επαναληπτικές μεθόδους που θα αναλυθούν στην συνέχεια ξεκινάει αν ορίσουμε με μορφή πινάκων την πιθανότητα μετακίνησης από τον κόμβο i στον κόμβο j ως

$$p_{ij} = \frac{W_{ij}}{\sum_k W_{ik}} \quad (2.31)$$

με σκοπό να προσεγγιστούν οι πιθανότητες των ετικετών των κλάσεων. Στην παραπάνω σχέση όπου W_{ij} υπολογίζεται από τον γκαουσιανό πυρήνα. Κάθε στοιχείο x_i συσχετίζεται με μια πιθανότητα $P(y = 1|i)$ με την οποία ανήκει στην κλάση 1. Δεδομένου ενός στοιχείου x_k,μπορεί να υπολογιστεί η πιθανότητα $P^t(y_{start} = 1|k)$,όπου ξεκινάμε από ένα σημείο της κλάσης $y_{start} = 1$, δεδομένου ότι φτάνουμε στο x_k, μετά από t βήματα τυχαίου περιπάτου βάση της σχέσης

$$P^t(y_{start} = 1|k) = \sum_{i=1}^{n} P\,(y = 1|i)\, P_{o|t}(i|k), \quad (2.32)$$

όπου $P_{o|t}(i|k)$ είναι η πιθανότητα που ξεκινάμε από x_i και φτάνουμε στον k μετά από t βήματα του τυχαίου περιπάτου. Το x_k κατηγοριοποιείται στην κλάση 1 ή στην -1 σύμφωνα με ένα κατώφλι. Η μέθοδος απόδοσης ετικέτας που μόλις αναλύθηκε προτάθηκε από τους Szummer και Jaakkola το 2002.

Ένας εναλλακτικός τρόπος να γίνει χρήση των τυχαίων περιπάτων είναι να αποδοθεί σε ένα στοιχείο x_i μια ετικέτα που εξαρτάται από την πιθανότητα να φτάσουμε σε ένα στοιχείο θετικής κλάσης όταν εκτελούμε έναν τυχαίο περίπατο, ξεκινώντας από το x_i και μέχρι να φτάσουμε σε ένα δεδομένο που έχει ετικέτα. Σε αυτήν ακριβώς την αρχή βασίζεται ο πρώτος αλγόριθμος.

Όταν το x_i είναι δεδομένο με ετικέτα ισχύει $P\,(y_{start} = 1|i) = \delta_{yi1}$, με δ την συνάρτηση δέλτα Kronecker,[1] και όταν είναι δεδομένο χωρίς ετικέτα ισχύει $P\,(y_{end} = 1|i) = \sum_{j=1}^{n} P\,(y_{end} = 1|j)p_{ij}$. Αν η πιθανότητα εκφραστεί στην μορφή πινάκων $P = D^{-1}W$, θεωρηθεί ο μη κανονικοποιημένος λαπλασιανός $L = D - W$ και οι σχέσεις $\hat{z}_\iota = P\,(y_{end} = 1|i)$ και $\hat{Z} = (\hat{Z}_l, \hat{Z}_u)$ προκύπτει το γραμμικό σύστημα

$$L_{uu}\hat{Z}_u = W_{ul}\hat{Z}_l. \quad (2.33)$$

Το σύστημα έχει λύση την $(\hat{Z}_l, \hat{Z}_u)$ και την $(\hat{Y}_l, \hat{Y}_u)$ με

$$\hat{Y}_u = 2\hat{Z}_u - (1, 1, \dots, 1)^T \quad (2.34)$$
$$\hat{Y}_l = 2\hat{Z}_l - (1, 1, \dots, 1)^T = Y_l. \quad (2.35)$$

Έτσι το γραμμικό σύστημα μπορεί να γραφεί ως

$$L_{uu}\hat{Y}_u = W_{ul}\hat{Y}_l \quad (2.36)$$

και η λύση του συστήματος είναι

$$\hat{Y}_u = -L_{uu}^{-1}L_{ul}Y_l \quad (2.37)$$

με το πρόσημο κάθε στοιχείου yi του $\hat{Y}_u$ να δίνει την υπολογισμένη ετικέτα του x_i. Η λύση αυτού του αλγορίθμου τυχαίου περιπάτου δίνεται σε κλειστή μορφή από ένα γραμμικό σύστημα, είναι δηλαδή ισοδύναμη με τον επαναληπτικό αλγόριθμο 1 ή τον 2 με μ να τείνει στο μηδέν και ε = 0. Η ετικέτα που αποδίδεται εξαρτάται από το αν η πιθανότητα να ανήκει το στοιχείο στη θετική ή αρνητική κλάση ξεπερνά, όπως ήδη αναφέρθηκε, ένα επιλεγμένο κατώφλι.

Οι αλυσίδες Markov έχουν εφαρμογές σε πολλούς τομείς όπως την φυσική, την χημεία, την ιατρική, την μουσική, την θεωρία παιγνίων, τον αθλητισμό, την οικονομία και άλλους. Στη φυσική μαρκοβιανά συστήματα εμφανίζονται στη θερμοδυναμική, στη στατιστική φυσική και μηχανική και γενικά όπου γίνεται χρήση μοντέλων πιθανότητας για την αναπαράσταση μη απολύτως ορισμένων λεπτομερειών ενός φυσικού προβλήματος. Στη χημεία το κλασικό μοντέλο για την δράση των ενζύμων Michaelis-Menten μπορεί να μελετηθεί ως μαρκοβιανή αλυσίδα και γενικά όποιες διαδικασίες είναι άγνωστο το μακρινό παρελθόν τους αλλά μόνο η ακριβώς προηγούμενη κατάσταση. Επίσης, οι μαρκοβιανές αλυσίδες χρησιμοποιούνται στα τηλεφωνικά δίκτυα και την Θεωρία Ουρών, στην αναγνώριση φωνής, σε εφαρμογές του παγκόσμιου ιστού, όπως για παράδειγμα το Page Ranking που χρησιμοποιεί η Google, στην μελέτη συμπεριφοράς μετοχών, στις Κοινωνικές Επιστήμες και την ανάλυση της εξέλιξης μιας κοινωνίας, στην Γενετική Πληθυσμών για την περιγραφή γενετικών παρεκκλίσεων, στην υλοποίηση εφαρμογών παραγωγής τυχαίου κειμένου, προφανώς στην Στατιστική για την περιγραφή πολύπλοκων μοντέλων και σε πολλές άλλες εφαρμογές. Κοινό χαρακτηριστικό όλων των εφαρμογών είναι η μελέτη του αποτελέσματος χωρίς να είναι γνωστές όλες οι λεπτομέρειες και πτυχές του προβλήματος, η μελέτη, δηλαδή, της τυχαίας έκβασης ενός γεγονότος με βάση τους ήδη γνωστούς, αλλά και άγνωστους παράγοντες, η εξαγωγή συμπεράσματος με ελλιπείς πληροφορίες.

Οι μέθοδοι διάδοσης ετικέτας που ακολουθούν μπορούν, επίσης, να συσχετιστούν με τα ηλεκτρικά κυκλώματα και θεωρηθούν ανάλογό τους, σύμφωνα με τους Zhou κ. α. αλλά και τους Doyle και Snell (1984). Ο πίνακας βάρους μπορεί να θεωρηθεί ως πίνακας τιμών αγωγιμότητας μεταξύ των κόμβων, για παράδειγμα το

Wij δηλώνει την αγωγιμότητα μεταξύ των κόμβων i και j. Η αγωγιμότητα είναι το αντίστροφο της αντίστασης και δηλώνει το πόσο εύκολο το ηλεκτρικό ρεύμα διαρρέει έναν αγωγό. Οι κόμβοι αντικαθιστώνται με αντιστάσεις. Κόμβοι με θετική ετικέτα είναι εκείνοι που είναι συνδεδεμένοι με θετική πηγή τάσης και αρνητική ετικέτα έχουν όσοι συνδέονται με αρνητική. Στόχος είναι ο υπολογισμός της τάσης των unlabeled δεδομένων. Έστω Iij η ένταση του ρεύματος που διαρρέει την περιοχή μεταξύ i και j και

$$Vij = \hat{y}_j - \hat{y}_i \quad (2.38)$$

η διαφορά δυναμικού μεταξύ των i και j, τότε σύμφωνα με το νόμο του Ωμ ισχύει

$$Iij = Wij\, Vij. \quad (2.39)$$

Ο κανόνας του Kirchoff για τις εντάσεις των ρευμάτων ορίζει ότι

$$\sum_j Iij = 0, \quad (2.40)$$

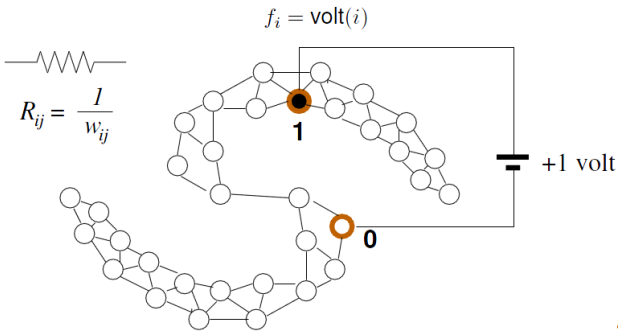
δηλαδή ότι το άθροισμα των ρευμάτων που εισέρχονται από έναν κόμβο είναι ίσο με το άθροισμα των εντάσεων των ρευμάτων που εξέρχονται από αυτόν. Ο κανόνας εφαρμόζεται στα unlabeled δεδομένα, δηλαδή i > l, αφού για αυτά αναζητείται η τιμή της τάσης, και η σχέση γράφεται λόγω της ως

$$0 = \sum_j Wij \left(\hat{y}_j - \hat{y}_i\right) = -\left(L\hat{Y}\right)_i. \quad (2.41)$$

Σε μορφή πινάκων η τελευταία σχέση γίνεται

$$L_{ul}Y_l + L_{uu}\hat{Y}_u = 0, \qquad (2.42)$$

ίδια με την λύση που προέκυψε από τους περιπάτους Markov. Συνοψίζοντας, οι γραφικές μέθοδοι μπορούν να θεωρηθούν ως ηλεκτρικά ανάλογα με τις τιμές των συναφειών να αντιστοιχούν στους αντιστάτες και τις τιμές των ετικετών να αντιστοιχούν σε πηγή τάσης.



Εικόνα 2.2: Ηλεκτρικό ανάλογο

2.5. Επαναληπτικοί αλγόριθμοι

2.5.1. Διάδοση ετικέτας

Σκοπός των αλγορίθμων είναι να χαρακτηρίσουν διαδοχικά τα δεδομένα του γράφου. Η διαδικασία που ακολουθείται είναι επαναληπτική. Λαμβάνονται, αρχικά, υπόψη τα ήδη χαρακτηρισμένα δεδομένα, και δίνονται ετικέτες στα μη χαρακτηρισμένα, μέχρι να προσπελαστεί όλος ο γράφος. Τα εκτιμώμενα δεδομένα – κόμβοι έχουν τη μορφή $\hat{Y} = (\widehat{Y_l}, \widehat{Y_u})$, όπου $\widehat{Y_l}$ τα δεδομένα που έχουν ετικέτες και $\widehat{Y_u}$ τα δεδομένα χωρίς ετικέτες. Τα προβλήματα που επιλύουν οι αλγόριθμοι είναι δύο κλάσεων, ενώ για προβλήματα περισσότερων κλάσεων είναι δυνατόν να επεκταθούν με κατάλληλες τροποποιήσεις.

Η *διάδοση ετικετών* (label propagation) στα δεδομένα του γράφου είναι η μέθοδος που βασίζονται οι τρεις πρώτοι αλγόριθμοι. Κάθε κόμβος διαδίδει στους γειτονικούς του την ετικέτα του και η διαδικασία επαναλαμβάνεται μέχρι να επέλθει σύγκλιση, δηλαδή όλοι οι κόμβοι να προσπελαστούν. Η επανάληψη ξεκινά από τα δεδομένα που έχουν ετικέτες και συνεχίζεται μέχρις ότου διαδοθεί πληροφορία και στον τελευταίο μη χαρακτηρισμένο κόμβο.

Ο πρώτος αλγόριθμός προτάθηκε από τους Zhu και Ghahramani (2002) έχει την εξής μορφή

Αλγόριθμος 1, Διάδοση ετικέτας, *Zhu και Ghahramani (2002)*

Υπολογισμός του πίνακα συνάφειας W
Υπολογισμός του διαγώνιου πίνακα D, όπου $Dii = \sum_j Wij$
Αρχικοποίηση πίνακα δεδομένων, έτσι ώστε τα μη χαρακτηρισμένα δεδομένα να είναι μηδενικά, $\hat{Y}^{(0)} \leftarrow (y_{1,\dots,}y_l, 0, \dots, 0)$
Επανάληψη για τον υπολογισμό των:
$\hat{Y}^{(t+1)} \leftarrow D^{-1}W\,\hat{Y}^{(t)}$

και $\hat{Y}^{(t+1)} \leftarrow Y_l$
μέχρι να επέλθει σύγκλιση στο $\hat{Y}^{(\infty)}$
Απόδοση ετικετών στο δεδομένα xi, σύμφωνα με το πρόσημο του $\hat{y}_i^{(\infty)}$

Χαρακτηριστικό γνώρισμά του είναι ότι αφήνει τις ετικέτες των γνωστών δεδομένων ανεπηρέαστες, επαναφέρει δηλαδή τις αρχικές τιμές τους, και δεν επηρεάζονται από τον ίδιο τον αλγόριθμο. Επίσης, απαιτεί σημεία με μικρή απόσταση να έχουν παρόμοιες ετικέτες. Η τιμή, τελικά, που αποδίδεται σε κάθε κόμβο χωρίς αρχική ετικέτα είναι ο μέσος όρος των ετικετών των γειτόνων του, ισχύει δηλαδή

$$f(u) = \frac{1}{du}\sum_{u\sim v} w(u,v)f(v)\,, \qquad (2.43)$$

με u να συμβολίζονται οι κόμβοι χωρίς ετικέτες και f η πραγματική συνάρτηση που αναθέτει τιμές στους κόμβους, η οποία αναφέρθηκε προηγουμένως. Εκφρασμένη σε μορφή πινάκων η συνάρτηση αυτή γίνεται f = Pf με $P = D^1W$. Αξίζει να συμπληρωθεί ότι η f είναι *μοναδική* σύμφωνα με την μέγιστη αρχή των αρμονικών συναρτήσεων, και είναι είτε σταθερή, είτε λαμβάνει τιμές μεταξύ μηδέν και ένα [10].

2.5.2. Σύγκριση αλγορίθμου 1

Το 2010 πραγματοποιήθηκε ερεύνα σχετικά με την απόδοση του αλγόριθμο 1 σε σχέση με άλλους δύο, τον αλγόριθμο *Προσρόφηση (Adsorption)* και τον αλγόριθμο *Τροποποιημένης Προσρόφησης (Modified Adsorption*), η οποία παραθέτεται συνοπτικά. Ο αλγόριθμος διάδοσης ετικέτας των Zhu κ.α. (2002) θεωρείται ο πιο αποτελεσματικός βασισμένος σε γράφο ssl αλγόριθμος και αποτέλεσε την βάση για πολλές παραλλαγές και προεκτάσεις. Η επιλογή των συγκεκριμένων αλγορίθμων έγινε γιατί έχει αποδειχτεί ότι είναι κατάλληλοι για την ssl νεπεξεργασία μεγάλων βάσεων δεδομένων.
Ο αλγόριθμος *Adsorption* προτάθηκε από τους Baluja κ.α. το 2009 και είναι ένας βασισμένος σε γράφο ssl αλγόριθμος. Είναι επαναληπτικός και οι ετικέτες που έχουν αποδοθεί στον κόμβο v στην (t+h) επανάληψη ανανεώνονται με χρήση προσεγγίσεων στην t επανάληψη σύμφωνα με τη σχέση

$$\hat{Y}_v^{t+1} \leftarrow p_v^{inj} \times Y_v + p_v^{cont} \times B_v^{(t)} + p_v^{abnd} \times r \quad (2.44)$$

με

$$B_v^{(t)} = \sum_u \frac{W_{uv}}{\sum_{u'} W_{u'v}} \hat{Y}_u^{(t)}. \quad (2.45)$$

Οι p_v^{inj}, p_v^{cont} και p_v^{abnd} είναι πιθανότητας που ορίζονται για κάθε κόμβο v του γράφου από τον αλγόριθμο και το άθροισμά τους είναι ίσο με ένα. Το r είναι ένα διάνυσμα που εκφράζει την αβεβαιότητα των ετικετών σε έναν κόμβο. Η κεντρική ιδέα του Adsortpion είναι ο στενότερος έλεγχος της διάδοσης ετικετών με μείωση του όγκου πληροφοριών που περνάνε μέσα από κάθε κόμβο.
Ο αλγόριθμος Modified Adsorption προτάθηκε από τους Talukdar και Crammer το 2009. Αποτελεί επέκταση του προηγούμενου, καθώς έχει όλα τα χαρακτηριστικά του, αλλά επίσης μπορεί αν εκφραστεί ως ένα μη περιορισμένο πρόβλημα βελτιστοποίησης,

$$\hat{Y}^{min} \quad \sum_{l\in c}[\mu 1\ (Y_l - \widehat{Y_l})^T S\ (Y_l - \widehat{Y_l})\ + \mu_2 \widehat{Y_l}^T L' \widehat{Y_l} + \mu_3 \left\|\widehat{Y_l} - R_l\right\|^2], \qquad (2.46)$$

με μ1, μ2 και μ3 να είναι υπερπαραμετρικές και L΄ ο Laplacian.

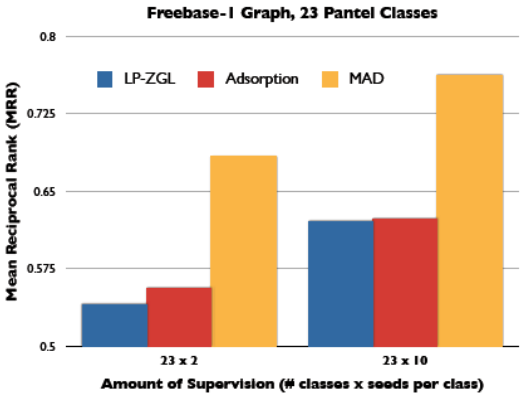
Στη σύγκριση λήφθηκε υπόψη το μέγεθος Mean Reciprocal Rank (MRR),

$$\text{MRR} = \frac{1}{|\text{Q}|}\sum_{v\in Q}\frac{1}{r_v}, \ (2.47)$$

με $\text{Q} \subseteq V$ το πλήθος των κόμβων που χρησιμοποιήθηκαν στην δοκιμή και r_v ένας αριθμός που σχετίζεται με την ετικέτα που αποδόθηκε στον κόμβο v. Όσο μεγαλύτερη η τιμή αυτού του δείκτη, τόσο καλύτερη η απόδοση του αλγορίθμου.

Η πρώτη δοκιμή έγινε με χρήση δεδομένων από την Freebase (Metaweb Technologies, 2009)[9], μιας μεγάλης ελεύθερης συλλογής σχεσιακών βάσεων δεδομένων. Ο γράφος κατασκευάστηκε με κόμβους κάθε μεμονωμένη τιμή των κελιών των πινάκων που χρησιμοποιήθηκαν και κάθε μεμονωμένη τιμή ιδιότητας οι οποίοι ενώνονται με ακμές ομοιότητας. Ο γράφος κατασκευάστηκε για 18 διαφορετικά πεδία διαφόρων θεματολογιών, όπως αστρονομία, βιολογία, βιβλία, φαγητό και άλλα. Οι Pantel κ.α. το 2009 εξήγαγαν δεδομένα από την Wikipedia τα οποία συγκρίθηκαν με την προηγούμενη περιγραφή και πραγματοποιήθηκαν δοκιμές σε δεδομένα 23 κλάσεων με 2 ή 10 labeled δεδομένα ανά κλάση. Ο αλγόριθμος 1 και ο Adsorption είχαν παρόμοια απόδοση, ενώ η απόδοση του Modified Absorption ήταν αρκετά καλύτερός τους. Τα αποτελέσματα συνοψίζονται στο παρακάτω γράφημα.

[9] http://www.freebase.com/



Εικόνα 2.3: Σύγκριση αλγορίθμου 1 με αλγόριθμο Adsorption και Modified Adsorption

Όμοιο πείραμα έγινε με μεγαλύτερο όγκο δεδομένων και τα αποτελέσματα ήταν ανάλογα. Πραγματοποιήθηκε, επίσης, πείραμα με χρήση γράφου που κατασκευάστηκε από δεδομένα του TextRunner[10]. Η κατασκευή του γράφου έγινε με διαφορετικό τρόπο, λόγω της φύσης των δεδομένων. Η απόδοση του Modified Adsorption ήταν καλύτερη, αλλά η διαφορά του από τους υπολοίπους ήταν αρκετά μικρότερη. Οι δοκιμές περιείχαν στοιχεία επικαλυπτόμενων κλάσεων.

Ο Modified Adsorption έχει πολύ καλή απόδοση σε γράφους μεγάλου βαθμού. Η συνάρτηση προς ελαχιστοποίηση στο εσωτερικό του επιτυγχάνει μεγαλύτερη ακρίβεια από την συνάρτηση του αλγορίθμου 1, λόγω των περιορισμών που επιβάλλει η καθεμία. Από το πείραμα προέκυψε ότι θα πρέπει να λαμβάνεται υπόψη ο βαθμός του γράφου, προκειμένου να επιλεγεί ο καταλληλότερος ssl αλγόριθμος για τα δεδομένα. Επίσης, η χρήση συνδυασμού ιδιοτήτων-οντοτήτων αποδίδει βέλτιστα αποτελέσματα όταν λαμβάνεται υπόψη η επιρροή των χαρακτηριστικών των βάσεων στην απόκτηση γνώσης των οντοτήτων των κλάσεων.[13]

2.5.3. Διάδοση Ετικέτας, άλλοι αλγόριθμοι

Στηριζόμενοι στον παραπάνω αλγόριθμο, καθώς στην επαναληπτική μέθοδο επίλυσης γραμμικών συστημάτων Jacobi, οι συγγραφείς όρισαν έναν καινούριο αλγόριθμο που διαφέρει από τον προηγούμενο σε κάποια βασικά σημεία. Οι τιμές των γνωστών δεδομένων επηρεάζονται από την εφαρμογή του αλγορίθμου, τα διαγώνια στοιχεία του πίνακα συνάφειας θέτονται ίσα με μηδέν και χρησιμοποιούνται συστηματικές σταθερές για καλύτερα αποτελέσματα.

[10] Προτάθηκε από τους Banko κ.α. το 2007 και είναι ένα ανοιχτό σύστημα *Εξαγωγής Πληροφοριών (Information Extraction)*.

Αλγόριθμος 2, ***Διάδοση ετικέτας, (Εμπνευσμένος από την μέθοδο Jacobi)***

Υπολογισμός του πίνακα συνάφειας W με Wii = 0 *Υπολογισμός του διαγώνιου πίνακα D, όπου Dii =* $\sum_j Wij$ *Επιλογή παραμέτρου α μεταξύ μηδέν και ένα και παραμέτρου ε > 0* *Υπολογισμός σταθεράς* $\mu \leftarrow \frac{\alpha}{1-\alpha} \in (0, +\infty)$ *Υπολογισμός του διαγώνιου πίνακα A έτσι ώστε* $Aii \leftarrow I_{[l]}(i) + \mu Dii + με$ *Αρχικοποίηση πίνακα δεδομένων, έτσι ώστε τα μη χαρακτηρισμένα δεδομένα να είναι μηδενικά,* $\hat{Y}^{(0)} \leftarrow (y_{1,\dots,}y_l, 0, \dots, 0)$ *Επανάληψη για τον υπολογισμό των ετικετών:* $\hat{Y}^{(t+1)} \leftarrow A^{-1}\left(\mu W \hat{Y}^{(t)} + \hat{Y}^{(0)}\right)$ *μέχρι να επέλθει σύγκλιση στο* $\hat{Y}^{(\infty)}$ *Απόδοση ετικετών στο δεδομένα xi, σύμφωνα με το πρόσημο του* $\hat{y}_i^{(\infty)}$*[1]*

Πίνακας του γραμμικού συστήματος θεωρείται ο διαγώνιος πίνακας A. Ο ορισμός του προϋποθέτει την κατασκευή του $I_{[l]}(i)$, ο οποίος είναι ένας πίνακας γραμμή που περιέχει μονάδες στις θέσεις των labeled στοιχείων.

Οι επαναληπτικές μέθοδοι του αλγορίθμου μπορούν να γραφούν και ως αθροίσματα με τη μορφή

$$\hat{y}_i^{(t+1)} \leftarrow \frac{\left(\sum_j Wij \hat{y}_j^{(t)} + \frac{1}{\mu} y_i\right)}{\sum_j Wij + \frac{1}{\mu} + \varepsilon} \quad (2.48)$$

για δεδομένα με ετικέτες και

$$\hat{y}_i^{(t+1)} \leftarrow \frac{\left(\sum_j Wij \hat{y}_j^{(t)}\right)}{\sum_j Wij + \varepsilon} \quad (2.49)$$

για δεδομένα χωρίς ετικέτες. Οι δύο παραπάνω σχέσεις δηλώνουν ότι εάν ένα δεδομένα έχει ήδη ετικέτα λαμβάνεται υπόψη στον χαρακτηρισμό του και η προηγούμενη τιμή του, ενώ εάν δεν έχει χαρακτηρίζεται, ομοίως με πριν, από τον μέσο όρο των ετικετών των γειτονικών δεδομένων. Η σύγκλιση του συγκεκριμένου αλγορίθμου αποδεικνύεται καθώς αποτελεί ανάλογο της μεθόδου Jacobi.

Ο επόμενος αλγόριθμος που υλοποιήθηκε, ορίστηκε από τους Zhou κ.α. το 2004. Ο χαρακτηρισμός προκύπτει από τους γειτονικούς κόμβους και από μικρή συμμετοχή της αρχικής τιμής του κόμβου. Η συνάρτηση κόστους που επιδιώκει να ελαχιστοποιήσει ο αλγόριθμος τρία είναι λίγο διαφορετική από την συνάρτηση που ελαχιστοποιούν οι προηγούμενοι. Περισσότερα για τις συναρτήσεις κόστους θα αναφερθούν στην συνέχεια.

Αλγόριθμος 3, ***Διάδοση ετικέτας,*** *(Zhou et al., 2004)*

Υπολογισμός του πίνακα συνάφειας W με Wii = 0
Υπολογισμός του διαγώνιου πίνακα D, όπου Dii = $\sum_j Wij$
Υπολογισμός του πίνακα L = $D^{-1/2}W\,D^{-1/2}$
Αρχικοποίηση πίνακα δεδομένων, έτσι ώστε τα μη χαρακτηρισμένα δεδομένα να είναι μηδενικά, $\hat{Y}^{(0)} \leftarrow (y_{1,\dots,}y_l, 0, \dots, 0)$
Επιλογή μιας παραμέτρου $\alpha \in [0, 1)$
Επανάληψη για τον υπολογισμό των ετικετών:
$\hat{Y}^{t+1} \leftarrow a\mathcal{L}\,\hat{Y}^t + (1-a)\,\hat{Y}^{(0)}$ *μέχρι να επέλθει σύγκλιση στο* $\hat{Y}^{(\infty)}$
Απόδοση ετικετών στο δεδομένα xi, σύμφωνα με το πρόσημο του $\hat{Y}_l^{(\infty)}$ *[1]*

Το κόστος σύγκλισης αναμένεται να είναι, στην χειρότερη περίπτωση, της τάξης του $O(kn^2)$, με k τον αριθμό των γειτόνων ενός σημείου του γράφου.

2.6. Σύγκληση αλγορίθμων

Η σύγκλιση των παραπάνω αλγορίθμων δίνει τελικά τα αποτελέσματα για τις ετικέτες των μη χαρακτηρισμένων δεδομένων. Επιδιώκεται η εύρεση της λύσης με την χρήση επαναληπτικών μεθόδων επίλυσης γραμμικών συστημάτων. Ο λόγος που επιλέχθηκαν οι επαναληπτικές μέθοδοι είναι γιατί οι πίνακες είναι μεγάλης τάξης και πολλά στοιχεία τους είναι ίσα με μηδέν. Ο αλγόριθμος της Γενικής Επαναληπτικής μεθόδου, δηλαδή της μεθόδου πάνω στην οποία βασίζονται οι περισσότερες μέθοδοι αυτής της κατηγορίας, ορίζει μια ακολουθία διανυσμάτων $\{x^{(k)}\}$, k = 0,1,… η οποία, αν συγκλίνει, θα συγκλίνει στην μοναδική λύση του γραμμικού συστήματος.

Για τον πρώτο αλγόριθμο επιθυμείται η σύγκλιση της σχέσης $\hat{Y}^{(t+1)} \leftarrow D^{-1}W\,\hat{Y}^{(t)}$. Ισχύει ότι $\|D^{-1}W\| < 1$, οπότε σύμφωνα με την Γενική Επαναληπτική μέθοδο συγκλίνει για οποιοδήποτε αρχικό διάνυσμα.

Για τον δεύτερο αλγόριθμο απαιτείται η σύγκλιση της σχέσης

$$\hat{Y}^{(t+1)} \leftarrow A^{-1}\left(\mu W\hat{Y}^{(t)} + \hat{Y}^{(0)}\right) \quad (2.50)$$

η οποία αποτελεί ανάλογο της μεθόδου επίλυσης γραμμικών συστημάτων Jacobi. Συγκεκριμένα, αν θεωρηθεί το n x n γραμμικό σύστημα Ax = b, όπου A ομαλός και ικανοποιεί τους περιορισμούς $A_{ii} \neq 0, i = 1,2,\dots,n$, η επαναληπτική σχέση που δίνει την τιμή κάθε συνιστώσας της λύσης είναι

$$x_i^{t+1} = \frac{1}{A_{ii}}\left(b - \sum_{j\neq i} A_{ij}x_j^t\right). \quad (2.51)$$

Με αντικατάσταση των $x \coloneqq \hat{Y}$, $b \coloneqq SY$ και $A \coloneqq S + \mu L + \mu\varepsilon I$ προκύπτει

$$\widehat{y_i}^{t+1} = \frac{1}{I_{[l]}(i) + \mu \sum_{j \neq i} W_{ij} + \mu\varepsilon} \left(I_{[l]}(i) y_i + \mu \sum_{j \neq i} W_{ij}\, \widehat{y_j}^{t} \right), \quad (2.52)$$

δηλαδή η σχέση που χρησιμοποιείται για τον χαρακτηρισμό των δεδομένων.

Θεώρημα: Αν ο πίνακας A του γραμμικού συστήματος έχει αυστηρή *διαγώνια υπεροχή κατά γραμμή*, δηλαδή

$$|A_{ii}| > \sum_{i \neq j} |A_{ij}|, \qquad (2.53)$$

τότε η επαναληπτική μέθοδος Jacobi συγκλίνει για οποιοδήποτε αρχικό διάνυσμα.
Η διαγώνια υπεροχή για τον πίνακα $A \coloneqq S + \mu L + \mu\varepsilon I$ ισχύει, αφού L = D –W και Dii = $\sum_{i \neq j} W_{ij}$, με $W_{ij} > 0$.

Ο αλγόριθμος 3 στηρίζεται σε μια λύση της 2.24 της μορφής

$$f = (1 - a)(I - aS)^{-1} y \quad (2.54)$$

με α = 1/(1+μ), λύση αντίστοιχης της κλειστής λύσης. Ο τελεστής Laplace για τον αλγόριθμο ορίστηκε ως $\Delta = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}$ = I-$D^{-\frac{1}{2}}WD^{-\frac{1}{2}} = I - S$, με $S = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. Πιο συγκεκριμένα, χρησιμοποιείται η σχέση

$$\hat{Y}^{t+1} \leftarrow a\mathcal{L}\,\hat{Y}^{t} + (1 - a)\hat{Y}^{(0)} \quad (2.55)$$

που γράφεται και

$$\hat{Y}^{t+1} \leftarrow a\mathcal{L}^{t}\,\hat{Y}^{(0)} + (1 - a)\sum_{i=0}^{t} a\mathcal{L}\,\hat{Y}^{(0)}. \quad (2.56)$$

Ο $\mathcal{L}$ είναι παρόμοιος του P = $D^{-1}W = D^{-1/2}\mathcal{L}D^{-1/2}$ και άρα έχει τις ίδιες ιδιοτιμές. Αφού ο P είναι στοχαστικός[11] πίνακας, οι ιδιοτιμές του ανήκουν στο διάστημα [-1, 1], ιδιότητα που αποδεικνύεται για τους στοχαστικούς πίνακες. Κατά συνέπεια οι ιδιοτιμές του $a\mathcal{L}$ ανήκουν στο διάστημα (-1, 1), αν ληφθεί υπόψη και ότι α < 1. Από την τελευταία σχέση αν $t \to \infty$, $(a\mathcal{L})^t \to 0$ [1][3], και άρα

[11] Εάν το άθροισμα των στοιχείων σε κάθε γραμμή ή στήλη ενός n τάξης μη αρνητικού πίνακα είναι 1 ο πίνακας ονομάζεται *στοχαστικός γραμμής ή στήλης* αντίστοιχα.

$$\sum_{i=0}^{t}(a\mathcal{L})^i \rightarrow (I - a\mathcal{L})^{-1} \quad (2.57)$$

δηλαδή

$$\hat{Y}^t \rightarrow \hat{Y}^{(\infty)} = (1-a)(I - \alpha\mathcal{L})^{-1}\hat{Y}^{(0)}, \quad (2.58)$$

με $\hat{Y}^{(0)}$ το αρχικό δοσμένο διάνυσμα.

2.7. Αλγόριθμος αναστροφής πινάκων

Οι Belkin και Niyogi (2003b) πρότειναν επίσης έναν αλγόριθμο ημιεπιτηρούμενης μάθησης βασισμένο, στην ιδέα της κανονικοποίησης του γράφου δεδομένων. Οι μέχρι τώρα αλγόριθμοι τείνουν να ελαχιστοποιήσουν το τετραγωνικό σφάλμα των εκτιμώμενων ετικετών. Αυτό που προτείνεται από τους συγγραφείς είναι να γίνει χρήση των ιδιοτήτων του γραφήματος δεδομένων και συγκεκριμένα του λαμπασιανού πίνακα L. Οι ιδιοτιμές των ιδιοσυναρτήσεων αποτελούν μέτρο ομαλότητας ως προς την προσέγγιση του συνόλου των δεδομένων. Το λαπλασιανό γράφημα είναι στενά συνδεδεμένο με τη λαπλασιανή του συνόλου των δεδομένων, της οποίας οι ιδιοσυναρτήσεις αποτελούν μια βάση του χώρου Hilbert για L^2 συναρτήσεις του συνόλου (Rosenberg, 1997). Αποδεικνύεται ότι η προβολή οποιασδήποτε συνάρτησης L^2 στις πρώτες p ιδιοσυναρτήσεις, οι οποιές έχουν διαταχθεί κατά σειρά αυξανόμενης ιδιοτιμής είναι ένας τρόπος ομαλοποίησης της συνάρτησης πάνω στο σύνολο των δεδομένων[1]. Την αρχή αυτή χρησιμοποίησαν για την ανάπτυξη του αλγορίθμου οι Belkin και Niyogi ο οποίος προτείνει τον υπολογισμό των p πρώτων ιδιοδιανυσμάτων του λαπλασιανού γράφου, και την εύρεση του γραμμικού συνδυασμού αυτών των ιδιοδιανυσμάτων που προβλέπουν καλύτερα τις ετικέτες. Επιδιώκεται η εύρεση ομαλής συνάρτησης που ταιριάζει καλύτερα στα δεδομένα (η καλύτερη συνάρτηση θα βρεθεί μέσω του τετραγωνικού σφάλματος), αφού θα έχει υπολογιστεί από τις p ομαλότερες ιδιοσυναρτήσεις της λαπλασιανής. Ο αλγόριθμος δεν βασίζεται αποκλειστικά στην ελαχιστοποίηση του τετραγωνικού σφάλματος και δεν έχει απόλυτη σύνδεση με τους αλγορίθμους διάδοσης ετικέτας που αναφέρθηκαν μέχρι τώρα, βασίζεται όμως σε παρόμοιες παραδοχές και επιδιώξεις σχετικά με τον γράφο των δεδομένων.

Αλγόριθμος 4 **Διάδοση ετικέτας,** (Belkin και Niyogi, 2003b)

Υπολογισμός του πίνακα συνάφειας W με Wii = 0 *Υπολογισμός του διαγώνιου πίνακα D, όπου Dii =* $\sum_j Wij$ *Υπολογισμός του πίνακα L = D – W*

Υπολογισμός των p πρώτων ιδιοδιανυσμάτων που αντιστοιχούν στις p μικρότερες ιδιοτιμές του L. *Ελαχιστοποίηση στις ιδιοτιμές με χρήση του τετραγωνικού κριτηρίου* $\sum_{i=1}^{l}(y_i - \sum_{i=1}^{p} a_j e_{j,i})^2$ *Απόδοση ετικετών στα δεδομένα xi* $(1 \leq i \leq n)$ *σύμφωνα με το πρόσημο του* $\sum_{i=1}^{p} a_j e_{j,i}$*[1]*

Ο λόγος που ο συγκεκριμένος αλγόριθμος ανήκει στους αλγορίθμους αναστροφής πινάκων είναι γιατί η διαδικασία μπορεί να υλοποιηθεί και με χρήση μόνο πινάκων συνεχίζοντας μετά τον υπολογισμό του πίνακα L ως:

- Αρχικοποίηση πίνακα δεδομένων, έτσι ώστε τα μη χαρακτηρισμένα δεδομένα να είναι μηδενικά, $\hat{Y}^{(0)} \leftarrow (Y_{1,\dots,}Y_l, 0, \dots, 0)$
- Για τον υπολογισμό των ετικετών χρήση της ισότητας: $\hat{Y} = (S + \mu L + \mu\varepsilon I)^{-1} SY$ [1]

Στην ενότητα *Κριτήρια Κόστους*, αναφέρεται αναλυτικά η απόδειξη της έκφρασης.

2.8. Κριτήρια κόστους

Οι αλγόριθμοι βασίζονται στην ελαχιστοποίηση κριτηρίων απαιτούμενου κόστους για τον χαρακτηρισμό των δεδομένων, τα οποία αναλύονται στη συνέχεια. Αν θεωρήσουμε το τελικό διάνυσμα ετικετών $\hat{Y} = \left(\hat{Y_l}, \hat{Y_u}\right)$ και το αρχικό $Y = (Y_l, Y_u)$, οι δύο εκφράσεις κριτηρίων κόστους είναι

$$C(\hat{Y}) = \left\|\hat{Y_l} - Y_l\right\|^2 + \mu\hat{Y}^T L\hat{Y_l} + \mu\varepsilon\left\|\hat{Y}\right\|^2, \quad (2.59)$$

το οποίο προτάθηκε από τους Belkin et al. (2004b), Delalleau et al. (2005), και η παρόμοια έκφραση του κριτηρίου που προτάθηκε από τους Zhou et al. (2004),

$$C'^{(\hat{Y})} = \left\|\hat{Y_l} - SY_l\right\|^2 + \frac{\mu}{2}\sum_{i,j} W_{ij}\left(\frac{\hat{y_i}}{\sqrt{D_{ii}}} - \frac{\hat{y_j}}{\sqrt{D_{jj}}}\right)^2. \quad (2.60)$$

Οι δύο συναρτήσεις κόστους πηγάζουν από την αναγκαιότητα να υπάρχει συνάφεια μεταξύ των αρχικών ετικετών των δεδομένων, αλλά και με τον γράφο που δημιουργείται από όλα τα δεδομένα, χαρακτηρισμένα και μη.

Στην πρώτη, η ευκλείδεια απόσταση μεταξύ των αρχικών και τελικών ετικετών των χαρακτηρισμένων δεδομένων είναι ο πρώτος όρος της συνάρτησης. Οι

άλλοι δύο όροι προκύπτουν από την παραδοχή της ομαλότητας του γράφου των δεδομένων, αλλά και από την προσπάθεια να αποφευχθούν φαινόμενα εκφυλισμού, που είναι πιθανό να παρουσιαστούν κατά την κατασκευή του γράφου. Τέτοια φαινόμενα παρουσιάζονται, αν, για παράδειγμα, ένα δεδομένο έχει ενωθεί μόνο με μη χαρακτηρισμένους όρους. Τα μ και ε είναι σταθερές, ενώ, όπου L = D–W είναι η μη κανονικοποιημένος Laplacian του γράφου.

Για την ελαχιστοποίηση της συνάρτηση υπολογίζεται η πρώτη παράγωγος ως προς $\hat{Y}$και τίθεται ίση με μηδέν, με σκοπό την εύρεση μιας έκφρασης ως προς $\hat{Y}$. Θεωρείται αρχικά ο διαγώνιος πίνακας S (n x n), τέτοιος ώστε $S_{ii} = I_{[l]}(i)$. Το πρώτο μέρος του της συνάρτησης γράφεται με χρήση του S ως $\left\|S\hat{Y} - SY\right\|^2$. Η πρώτη παράγωγος είναι

$$\frac{1}{2}\frac{\partial C(\hat{Y})}{\partial \hat{Y}} = (S + \mu L + \mu\varepsilon I)\hat{Y} - SY \quad (2.61)$$

και η δεύτερη

$$\frac{1}{2}\frac{\partial^2 C(\hat{Y})}{\partial \hat{Y}\partial \hat{Y}^T} = S + \mu L + \mu\varepsilon I, \quad (2.62)$$

ποσότητα η οποία είναι θετική όταν ε > 0. Από τον μηδενισμό της πρώτης παραγώγου προκύπτει η έκφραση για τον τελικό πίνακα ετικετών, αν η σχέση λυθεί απλά ως προς $\hat{Y}$,

$$\hat{Y} = (S + \mu L + \mu\varepsilon I)^{-1} SY, \quad (2.63)$$

δηλαδή ο τρόπος υπολογισμού των ετικετών στον αλγόριθμο 4. Οι ετικέτες που δίνονται στα δεδομένα εξαρτώνται από τον γράφο αποκλειστικά και όχι από τις αρχικές ετικέτες.

Η δεύτερη συνάρτηση κόστους γράφεται ως

$$C'^{(\hat{Y})} = \left\|\hat{Y}_l - Y_l\right\|^2 + \left\|\widehat{Y_u}\right\|^2 + \mu\hat{Y}^T(I - \mathcal{L})\hat{Y} \Longrightarrow$$
$$C'^{(\hat{Y})} = \left\|\hat{Y}_l - Y_l\right\|^2 + \left\|\widehat{Y_u}\right\|^2 + \mu(D^{-\frac{1}{2}}\hat{Y})^T L\left(D^{-\frac{1}{2}}\hat{Y}\right), \quad (2.64)$$

με D τον διαγώνιο πίνακα, έτσι όπως ορίζεται στους αλγορίθμους. Οι βασικές διαφορές της από την πρώτη είναι δύο. Πρώτον, ο όρος $\left\|\hat{Y}_l - Y_l\right\|^2 + \left\|\widehat{Y_u}\right\|^2$ επιδιώκει να μηδενίσει τις ετικέτες των μη χαρακτηρισμένων δεδομένων και, δεύτερον, οι ετικέτες κανονικοποιούνται από τη ρίζα των διαγώνιων στοιχείων του πίνακα D κατά

τον υπολογισμό της ομοιότητάς τους. Η παράγωγος του κριτηρίου ως προς $\widehat{Y}$ δίνει ως αποτέλεσμα

$$\frac{1}{2}\frac{\partial C(\widehat{Y})}{\partial \widehat{Y}} = \widehat{Y} - SY + \mu\left(\widehat{Y} - \mathcal{L}\widehat{Y}\right) \quad (2.65)$$

και τελικά με μηδενισμός λύση ως προς $\widehat{Y}$ προκύπτει

$$\widehat{Y} = \left((1+\mu)I - \mu\,\mathcal{L}\right)^{-1} SY. \quad (2.66)$$

Η έκφραση για την τελική ανάθεση ετικετών είναι $\hat{Y} = \left((1+\mu)I - \mu\mathcal{L}\right)^{-1} S\,Y$ και, αν συγκριθεί με την $\hat{Y}^t \to \hat{Y}^{(\infty)} = (1-a)(I - \alpha\mathcal{L})^{-1}\hat{Y}^{(0)}$ του αλγορίθμου 3, είναι σχεδόν ίδια έως κάποια τιμή του μ,. Οι μικρές διαφορές που ίσως εμφανιστούν στην απόδοσης ετικετών είναι μικρής σημασίας, καθώς το πρόσημο λαμβάνεται τελικά υπόψη και όχι η ακριβής τιμή της ετικέτας [1].

.

3. Αποτελέσματα αλγορίθμων

3.1. Βασικοί αλγόριθμοι

Οι τέσσερις αλγόριθμοι υλοποιήθηκαν και εφαρμόστηκαν σε πίνακες δεδομένων προκειμένου να μελετηθεί η αποτελεσματικότητα τους σε δεδομένα που ανήκουν σε κλάσεις μεγάλης ταξικής ανισορροπίας. Είναι αναμφίβολη η επιτυχία τους σε χαρακτηρισμό δεδομένων προερχόμενα από κλάσεις με μικρές διαφορές πληθυσμού, για αυτό και δεν μελετάται στην παρούσα εργασία. Οι αλγόριθμοι έχουν την δυνατότητα επίλυσης προβλημάτων δύο κλάσεων.

Η εφαρμογή τους έγινε σε δεδομένα κατάλληλα διαμορφωμένα. Ζητούμενο ήταν να εισαχθεί σε κάθε αλγόριθμο, διάνυσμα του οποίου τα πρώτα στοιχεία είναι δεδομένα με ετικέτα, και τα τελευταία δεδομένα χωρίς ετικέτα. Οι πίνακες δεδομένων αποτελούνται από 10 κλάσεις. Προκειμένου να επιλυθεί πρόβλημα 10 κλάσεων με μεθόδους δύο κλάσεων ακολουθήθηκε η μέθοδος 10 fold cross validation. Εκατό προβλήματα δύο κλάσεων επιλύονται στο σύνολο. Στην συνέχεια θα γίνει λεπτομερής ανάλυση της μεθόδου.

Πρώτο βήμα αποτέλεσε ο σχεδιασμός των αλγορίθμων σε MATLAB (έκδοση 7.12.0.635 (R2011a)), όπου, με βάση την θεωρητική διατύπωση γράφτηκαν αντίστοιχοι αλγόριθμοι. Οι τρείς πρώτοι αποδίδουν ετικέτες μέσω επαναληπτικών μεθόδων και ο τελευταίος χαρακτηρίζει τα δεδομένα μέσω γραμμικής σχέσης. Τα ορίσματα για κάθε αλγόριθμο είναι αναλυτικά:

- Αλγόριθμος 1:
    - X: ένα (m x n) διάνυσμα δεδομένων, όπου κάθε γραμμή αντιστοιχεί σε ένα δεδομένα με n αριθμό χαρακτηριστικών. Το X εισάγεται στη μορφή X = [Xl, Xu], με (l x m) να είναι ο υποπίνακας του X με τα labeled στοιχεία και (u x m) ο υποπίνακας με τα unlabeled στοιχεία.
    - Yl: είναι το (l x 1) δεδομένο διάνυσμα, το οποίο περιέχει τις πραγματικές ετικέτες των labeled δεδομένων.
    - Sigma: παράμετρος που σχετίζεται με την κατασκευή του πίνακα συνάφειας του γραφήματος.
    - Theta: παράμετρος για έλεγχο της σύγκλισης του αλγορίθμου.

- Αλγόριθμος 2: X, Yl, Theta, ομοίως με παραπάνω και επίσης
    - Alpha: παράμετρος που ανήκει στο διάστημα (0,1) και χρησιμοποιείται για τον ορισμό του πίνακα A, ο οποίος είναι ο πίνακας του συστήματος προς επίλυση.
    - Epsilon: παράμετρος που χρησιμοποιείται για τον ορισμό του πίνακα A.
- Αλγόριθμος 3: X, Yl, sigma, alpha, theta, ομοίως με παραπάνω.
- Αλγόριθμος 4: X , Yl, sigma, alpha, epsilon, ομοίως με παραπάνω.

Ο πίνακας δεδομένων X αποτελεί κοινό όρισμα σε όλους και αποτελείται από 1000 γραμμές και 30 στήλες. Κάθε γραμμή είναι ένα διάνυσμα 30 χαρακτηριστικών. Για επιβεβαίωση της λειτουργικότητας και αποτελεσματικότητας των αλγορίθμων έγιναν δοκιμές τόσο σε μικρότερους όσο και σε μεγαλύτερους πίνακες δεδομένων. Οι δοκιμές επιβεβαιώνουν πλήρως τα αποτελέσματα και τα συμπεράσματα που εξήχθησαν, αλλά στα πλαίσια της διπλωματικής αυτής εργασίας θα εστιάσουμε στα αποτελέσματα των αλγορίθμων για τον πίνακα X, τα οποία είναι και απόλυτα αντιπροσωπευτικά.

Αναλόγως με την μέθοδο στην οποία στηρίχθηκαν, καθένας λαμβάνει ως ορίσματα και κάποιες ξεχωριστές παραμέτρους. Κοινά χαρακτηριστικό είναι, επίσης, ο υπολογισμός του πίνακα συνάφειας W και του διαγώνιου πίνακα D. Ο W υπολογίζεται σε ξεχωριστή μέθοδο την ConnectivityWeightMatrix.m, σύμφωνα με την σχέση

$$Wij = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}, \quad (3.1)$$

με xi και xj να είναι στοιχεία του πίνακα X. Τα στοιχεία της κύριας διαγωνίου του W θέτονται ίσα με μηδέν ή διατηρούν τις τιμές που προκύπτουν από τον γκαουσιανό πυρήνα. Ο διαγώνιος πίνακας D του W είναι τέτοιος ώστε Dii = $\sum_j Wij$. Τα στοιχεία της κύριας διαγωνίου του πίνακα D είναι το άθροισμα όλων των στοιχείων της στήλης j.

Τα δεδομένα είναι διανύσματα, και κάθε γραμμή αποτελεί ένα ξεχωριστό στοιχείο με ιδιαίτερα χαρακτηριστικά, οι τιμές των οποίων αναγράφονται στις στήλες του πίνακα. Η διαμόρφωση των πινάκων δεδομένων έγινε ώστε να υπάρχει απόλυτη συμφωνία με το πρότυπο που ορίζουν οι αλγόριθμοι σχετικά με το διάνυσμα δεδομένων, δηλαδή ότι $\hat{Y} = (\widehat{Y_l}, \widehat{Y_u})$, με $\widehat{Y_l}$ τα δεδομένα που έχουν ετικέτες και $\widehat{Y_u}$ τα δεδομένα χωρίς ετικέτες. Το 90% των δεδομένων θεωρήθηκε ότι είχε γνωστή ετικέτα και του 10% αναζητείται μέσω του αλγορίθμου.

Αρχικά ο πίνακας δεδομένων διασπάτε και συναρμολογείται ξανά έτσι ώστε να έχει την εξής μορφή

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_l \\ X_{l+1} \\ \vdots \\ X_{l+u} \end{bmatrix},$$

με l ο αριθμός των δεδομένων με ετικέτα, u ο αριθμός των στοιχείων χωρίς ετικέτα και ι+u ο συνολικός αριθμός των γραμμών m. Κάθε Xi είναι διάνυσμα n στηλών. Τα στοιχεία από ι+1 έως ι+u είναι αυτά που χρησιμοποιούνται για εξαγωγή συμπεράσματος σχετικά με την ορθότητα των αποτελεσμάτων των μεθόδων. Το 90% των Xi ανήκει στην αρνητική κλάση -1 και το 10% αυτών στην θετική κλάση +1. Οι δύο κλάσεις έχουν πολύ μεγάλη διαφορά ως προς την σχετική τους συχνότητα, γεγονός που αυξάνει κατά πολύ την πιθανότητα λάθους. Το τελευταίο επιβεβαιώνεται απόλυτα και από τα αποτελέσματα που προέκυψαν, και τα οποία αναλύονται στη συνέχεια. Μετά την ολοκλήρωση όλων των διαδικασιών εισαγωγής και επεξεργασίας, και την λήξη της διαδικασίας χαρακτηρισμού το διάνυσμα επιθυμούμαι να έχει την μορφή

$$Y = \begin{bmatrix} Y1 \\ Y2 \\ \vdots \\ Y_{10\%*l} \\ Y_{(10\%*l)+1} \\ \vdots \\ Y_l \\ Y_{l+1} \\ \vdots \\ Y_{10\%*u} \\ Y_{(10\%*u)+1} \\ \vdots \\ Y_u \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \\ 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix},$$

με $10\% * l$ να σημαίνει ότι το 10% των χαρακτηρισμένων στοιχείων ανήκει στην θετική κλάση. Επίσης η έκφραση $Y_{(10\%*l)+1}$ σημαίνει το 90% των ι δεδομένων (από 10% έως 90%) ανήκει στην αρνητική κλάση και ομοίως το 10% των μη χαρακτηρισμένων στοιχείων ανήκει στην θετική κλάση και το υπόλοιπο 90% στην αρνητική. Η σχέση μεταξύ ι και u είναι u = 0.1m και ι = 0.9 m, με m ο συνολικός αριθμός των στοιχείων, δηλαδή ως δεδομένα με ετικέτες θεωρείται το 90% των Xi, i = 1,2, …, m. Όσο πιο κοντά στην παραπάνω μορφή υπολογιστεί το διάνυσμα Y, τόσο πιο επιτυχής ως προς την απόδοση ετικετών είναι η μέθοδος.

Οι αλγόριθμοι υπολογίζουν από τα δεδομένα που τους δίνονται τον αριθμό των στοιχείων με ετικέτα και των στοιχείων χωρίς ετικέτα. Συγκεκριμένα εισάγεται το διάνυσμα στήλη Y_l, το οποίο περιέχει την πληροφορία των τιμών, αλλά και, προφανώς, του αριθμού των γνωστών στοιχείων. Ο Y_l είναι ο

$$Y_l = \begin{bmatrix} Y_1 \\ \vdots \\ Y_{90} \\ Y_{11} \\ \vdots \\ Y_{900} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix}.$$

Η διαφορά του αριθμού των γραμμών του αρχικού πίνακα X και του Y_l δίνει τον αριθμό των άγνωστων δεδομένων (u), δηλαδή υπολογίζεται το u = m – l, με m τον αριθμό των γραμμών του πίνακα X. Ο πίνακας X, ο οποίος έχει 1000 γραμμές, αποτελείται πάντα από u = 100 στοιχεία χωρίς ετικέτα, εκ των οποίων τα 90 ανήκουν στην αρνητική κλάση και τα 10 στην θετική κλάση. Επίσης 810 χαρακτηρισμένα στοιχεία γνωρίζει η μέθοδος ότι ανήκουν στην αρνητική κλάση και 90 στη θετική.

Η εξαγωγή αποτελεσμάτων, για καθένα από τα δέκα προβλήματα δύο κλάσεων, δεν γίνεται με μόνο μία είσοδο του πίνακα X της παραπάνω μορφής. Ο X εισάγεται 10 φορές, κατά τις οποίες εναλλάσσονται τα μη χαρακτηρισμένα στοιχεία, δηλαδή τα στοιχεία χωρίς ετικέτα. Σε κάθε είσοδο 10 διαφορετικά στοιχεία θεωρούνται μη χαρακτηρισμένα για την θετική κλάση και 90 διαφορετικά για την αρνητική. Η διαδικασία επαναλαμβάνεται 10 φορές για κάθε πρόβλημα δύο κλάσεων, και τελικά πραγματοποιούνται 100 επαναλήψεις μέχρι την ολοκλήρωση της απόδοσης ετικετών σε όλα τα στοιχεία, εφαρμόζεται, δηλαδή, η μέθοδος του 10 fold cross – validation.

Ο πίνακας δεδομένων εισάγεται σε μέθοδο προκειμένου να εφαρμοστούν οι αλγόριθμοι, αλλά και οι 100 επαναλήψεις. Η διαμόρφωση του πίνακα και ο ορισμός των ετικετών πραγματοποιείται μέσα σε αυτήν την μέθοδο. Αρχικά ο πίνακας δεδομένων κανονικοποιείται μέσω της μεθόδου get_normalized_matrix. Από το μέγεθός του, και θεωρώντας δεδομένο ότι όλες οι κλάσεις αποτελούνται από 100 στοιχεία, υπολογίζονται ο αριθμός των κλάσεων (10 για τον X), ο αριθμός των στοιχείων με ετικέτα και των στοιχείων χωρίς ετικέτα, βάση των ποσοστών 90% και 10% και, ομοίως, ο αριθμός των στοιχείων της αρνητικής και της θετικής κλάσης. Με αυτό τον τρόπο η μέθοδος μπορεί να εφαρμοστεί για κάθε πίνακα X που αποτελείται από κλάσεις μεγέθους 100 γραμμών.

Στη συνέχεια, με την βοήθεια κατάλληλων διανυσμάτων, όπου αποθηκεύονται οι απαιτούμενοι δείκτες, δύο επαναληπτικές μέθοδοι αποδίδουν ετικέτες σε όλα τα

στοιχεία. Η μία επανάληψη ορίζει το k πρόβλημα, με k = 1, 2, …,10. Η δεύτερη είναι εμφωλευμένη στην πρώτη και ορίζει το f fold, με f = 1, 2, …, 10. Στην εξωτερική επανάληψη διαχωρίζονται τα στοιχεία της θετικής από τα στοιχεία της αρνητικής, και ορίζεται με αυτό τον τρόπο το τρέχων πρόβλημα δύο κλάσεων. Στην εσωτερική επανάληψη τα δεδομένα διαχωρίζονται σε γνωστά και άγνωστα για κάθε κλάση. Τα στοιχεία που θεωρούνται γνωστά αλλάζουν 10 φορές, μέχρις ότου όλα τα στοιχεία να ληφθούν υπόψη μια φορά ως δεδομένα χωρίς ετικέτα. Σε κάθε επανάληψη, ο πίνακας X γίνεται,

$$X = \begin{bmatrix} Xl \\ Xu \end{bmatrix} \kappa\alpha\iota\, Xl = \begin{bmatrix} X_1 \\ \vdots \\ X_l \end{bmatrix} \kappa\alpha\iota\, Xu = \begin{bmatrix} X_{l+1} \\ \vdots \\ X_{l+u\,=\,n} \end{bmatrix}$$

με

$$X_{l+} = \begin{bmatrix} X_1 \\ \vdots \\ X_{10\%\ \tau o \upsilon\ l} \end{bmatrix} \kappa\alpha\iota\, X_{l-} = \begin{bmatrix} X_{(10\%\ \tau o \upsilon\ l)\,+\,1} \\ \vdots \\ X_l \end{bmatrix}$$

και

$$X_{u+} = \begin{bmatrix} X_{l+1} \\ \vdots \\ X_{10\%\ \tau o \upsilon\ u} \end{bmatrix} \kappa\alpha\iota\, X_{u-} = \begin{bmatrix} X_{(10\%\ \tau o \upsilon\ u)+1} \\ \vdots \\ X_{l\,+\,u} \end{bmatrix}$$

ενώ στην δεύτερη εμφωλευμένη επανάληψη γίνεται

$$X_{u+} = \begin{bmatrix} X_{l\,+\,(10\%\ \tau o \upsilon\ u)} \\ \vdots \\ X_{l\,+\,(10\%\ \tau o \upsilon\ u)\,+\,(10\%\ \tau o \upsilon\ u)} \end{bmatrix} \kappa\alpha\iota\, X_{u-} = \begin{bmatrix} X_{l\,+\,1} \\ \vdots \\ X_{l\,+\,(10\%\ \tau o \upsilon\ u)-\,1} \\ X_{l\,+\,(10\%\ \tau o \upsilon\ u)\,+\,(10\%\ \tau o \upsilon\ u)+1} \\ \vdots \\ X_{l\,+u} \end{bmatrix}.$$

Ομοίως συνεχίζεται η εναλλαγή των δεδομένων κάθε k προβλήματος f φορές, δηλαδή κάθε πρόβλημα λύνεται 10 φορές, αφού το f για εμάς είναι 10. Αν αντικατασταθούν οι παραμετρικοί δείκτες με αριθμητικούς προκύπτει για την πρώτη εξωτερική επανάληψη,

$$X_+ = \begin{bmatrix} X_1 \\ \vdots \\ X_{100} \end{bmatrix} \kappa\alpha\iota\, X_- = \begin{bmatrix} X_{101} \\ X_{1000} \end{bmatrix}.$$

Στην πρώτη εμφωλευμένη επανάληψη,

$$X_{l+} = \begin{bmatrix} X_1 \\ \vdots \\ X_{91} \end{bmatrix} \kappa\alpha\iota\, X_{l-} = \begin{bmatrix} X_{92} \\ \vdots \\ X_{900} \end{bmatrix}$$

και

$$X_{u+} = \begin{bmatrix} X_{901} \\ \vdots \\ X_{910} \end{bmatrix} \kappa\alpha\iota\, X_{u-} = \begin{bmatrix} X_{911} \\ \vdots \\ X_{1000} \end{bmatrix},$$

ενώ στην δεύτερη εμφωλευμένη επανάληψη γίνεται

$$X_{u+} = \begin{bmatrix} X_{911} \\ \vdots \\ X_{920} \end{bmatrix} \kappa\alpha\iota\, X_{u-} = \begin{bmatrix} X_{901} \\ \vdots \\ X_{910} \\ X_{921} \\ \vdots \\ X_{1000} \end{bmatrix},$$

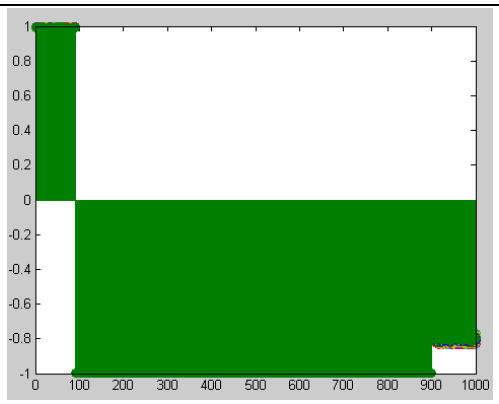
με το κάθε Xi να είναι διάνυσμα 30 στηλών.
Στην δεύτερη εξωτερική επανάληψη, και άρα στο δεύτερο πρόβλημα, η εξής διαμόρφωση εφαρμόζεται,

$$X_{+} = \begin{bmatrix} X_{101} \\ \vdots \\ X_{200} \end{bmatrix} \kappa\alpha\iota\, X_{-} = \begin{bmatrix} X_1 \\ \vdots \\ X_{100} \\ X_{201} \\ \vdots \\ X_{1000} \end{bmatrix}.$$

Ομοίως με παραπάνω διαμορφώνονται και τα στοιχεία l και u, και τα l_+, l_, u_+, u_.

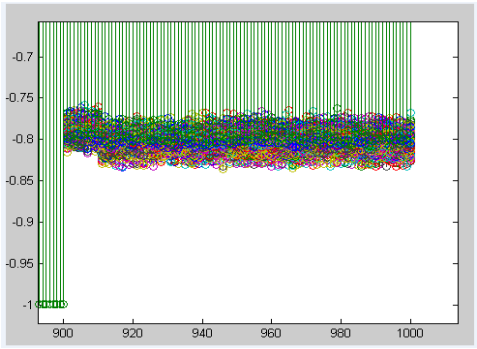
Οι αλγόριθμοι βρίσκονται σαν συναρτήσεις στο εσωτερικό των μεθόδων που πραγματοποιούν την παραπάνω διαδικασία, δέχθηκαν ως είσοδο τον κατάλληλα διαμορφωμένο X και προέκυψαν τα αποτελέσματα που φαίνονται στα παρακάτω διαγράμματα. Για κάθε αλγόριθμο, συνολικά όλες μαζί οι επαναλήψεις απέδωσαν ένα πίνακα 1000 x 100, όπου 1000 είναι οι γραμμές του πίνακα και οι στήλες ανά δέκα αντιστοιχούν στα δέκα προβλήματα που επιλύθηκαν. Αναλυτικά παρουσιάζονται τα αποτελέσματα για τον αλγόριθμο 2, ο οποίος επιλέχθηκε τυχαία για αναλυτική παρουσίαση αποτελεσμάτων. Αποδεικνύεται θεωρητικά ότι όλοι είναι ισοδύναμοι.

Η εκτέλεση της μεθόδου ApplyinJacobi.m δίνει ως αποτέλεσμα πίνακα με τις τιμές των ετικετών. Η γραφική παράσταση των αποτελεσμάτων φαίνεται στην παρακάτω εικόνα για κάθε k πρόβλημα και f fold.



Εικόνα 3.1: Τελική απόδοση ετικετών από αλγόριθμο 2

Τα πρώτα 900 στοιχεία λαμβάνουν τις σωστές ετικέτες. Τα τελευταία 100 είναι τα unlabeled στοιχεία και χαρακτηρίζονται ως στοιχεία της αρνητικής κλάσης. Ο οριζόντιος άξονας αντιστοιχεί στον αριθμό των στοιχείων και ο κάθετος στις ετικέτες. Στα πρώτα 900 στοιχεία αποδίδονται ετικέτες -1 και 1, δεν επηρεάζονται δηλαδή οι τιμές των labeled στοιχείων, αν και ο συγκεκριμένος αλγόριθμος αλλάζει στο $4^{ο}$ και $5^{ο}$ δεκαδικό στοιχείο την τιμή των labeled στοιχείων, αλλαγή, όμως, που είναι αμελητέα. Στα τελευταία 100 οι ετικέτες που αποδίδονται είναι -1, αν ληφθεί υπόψη μόνο το πρόσημο, αν όμως αφήσουμε τις τιμές που αποδίδει στα στοιχεία ο αλγόριθμος έτσι όπως αυτές υπολογίστηκαν, λαμβάνεται η παραπάνω εικόνα. Το διάγραμμα κατασκευάστηκε στο Matlab. Οι τιμές των τελευταίων 100 στοιχείων σε μεγέθυνση είναι



Εικόνα 3.2: Ετικέτες τελευταίων 100 στοιχείων

Από τις ετικέτες της εικόνας 1 (και 2) κατασκευάστηκε πίνακας, όπου έχει υπολογιστεί το ποσοστό επιτυχίας στην απόδοση ετικετών ως συνάρτηση της διαφοράς της καθαρής τιμής της ετικέτας που αποδίδεται από τον αλγόριθμο στο στοιχείο, με την πραγματική ετικέτα του στοιχείου. Για κάθε ki και fi φαίνονται τα ποσοστά επιτυχίας για τα πρώτα 10 στοιχεία της θετικής κλάσης και τα 90 της

αρνητικής. Για κάθε ki η πρώτη γραμμή δίνει το ποσοστό που πλησιάζεται η τιμή 1 για τα στοιχεία θετικής κλάσης και η δεύτερη γραμμή το κατά πόσο πλησιάζεται η τιμή -1 για τα 90 τελευταία στοιχεία. Είναι, δηλαδή, *ποσοστά εγγύτητας τιμής*. Τα ποσοστά υπολογίστηκαν με βάση το μέσο όρο των αποδιδόμενων ετικετών. Η πρώτη γραμμή αποτελείται από τιμές κοντά στο 10% και η δεύτερη από τιμές κοντά στο 90%. Η επιτυχία απόδοσης τιμής κοντά στην πραγματική για τα αρνητικά στοιχεία είναι πολύ μεγάλη. Δεν συμβαίνει όμως το ίδιο και για την αναγνώριση στοιχείων της θετικής κλάσης. Ο πίνακας είναι ο παρακάτω.

Πίνακας 1: Αποτελέσματα ποσοστού επιτυχίας εγγύτητας τιμής ετικέτας σε μορφή πίνακα. Για κάθε ki η πρώτη γραμμή αντιστοιχεί στην θετική κλάση και η δεύτερη στην αρνητική.

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
k1	10,28%	10,38%	10,42%	10,33%	10,20%	10,24%	10,14%	10,22%	10,37%	10,33%
	89,93%	89,73%	89,70%	89,67%	89,83%	89,88%	89,84%	89,71%	89,68%	89,73%
k2	11,00%	10,90%	10,91%	10,67%	10,92%	10,66%	10,85%	10,76%	10,66%	10,56%
	90,85%	90,54%	91,04%	91,08%	90,37%	90,23%	90,83%	90,98%	91,02%	90,70%
k3	10,22%	10,39%	10,44%	10,36%	10,21%	10,34%	10,42%	10,45%	10,46%	10,50%
	89,69%	89,71%	89,92%	89,96%	90,01%	89,96%	89,98%	90,02%	89,97%	89,80%
k4	10,74%	10,52%	10,60%	10,79%	10,58%	10,63%	10,70%	10,82%	10,65%	10,74%
	89,57%	90,14%	89,86%	89,51%	89,84%	89,91%	89,54%	89,39%	89,50%	89,61%
k5	10,84%	10,37%	10,50%	10,67%	10,74%	10,50%	10,71%	10,82%	10,68%	10,63%
	89,72%	90,26%	90,06%	89,60%	89,96%	90,11%	89,82%	89,31%	89,36%	89,91%
k6	10,54%	10,68%	10,71%	10,58%	10,56%	10,60%	10,36%	10,50%	10,53%	10,68%
	90,26%	89,34%	90,00%	90,51%	90,60%	90,41%	90,33%	90,46%	90,45%	90,19%
k7	10,12%	10,14%	9,99%	10,12%	10,23%	10,14%	10,15%	9,92%	10,11%	9,91%
	90,50%	90,57%	90,57%	90,45%	90,58%	90,65%	90,52%	90,65%	90,70%	90,37%
k8	10,64%	10,49%	10,96%	11,10%	10,96%	11,14%	10,72%	11,09%	10,87%	10,90%
	89,63%	90,02%	89,88%	89,40%	89,35%	89,59%	89,83%	89,76%	89,51%	89,83%
k9	10,92%	10,69%	10,79%	10,74%	10,77%	10,84%	10,63%	10,54%	10,81%	10,54%
	89,58%	90,10%	89,87%	89,68%	89,31%	89,59%	90,07%	89,94%	89,54%	89,78%
k10	10,07%	10,02%	10,11%	10,36%	10,24%	10,19%	10,02%	10,04%	10,15%	10,26%
	89,99%	90,07%	89,97%	89,98%	89,96%	90,07%	90,06%	89,87%	90,16%	90,04%

Οι στήλες του πίνακα αντιστοιχούν στο fi fold και οι γραμμές στο ki πρόβλημα. Ο μέσος όρος επιτυχίας για κάθε πρόβλημα είναι

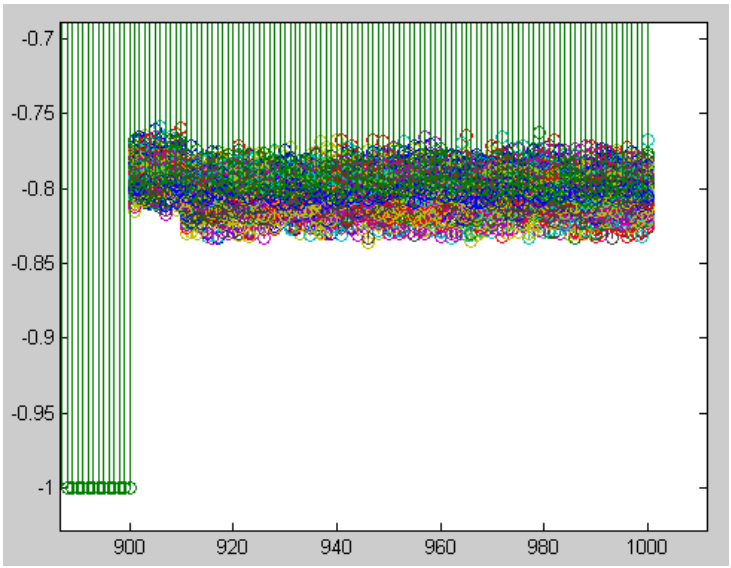
Πίνακας 2: Μέσος όρος fold για κάθε k πρόβλημα αλγορίθμου 2

k1	**10,29%**
	89,77%
k2	**10,79%**
	90,76%
k3	**10,38%**
	89,90%
k4	**10,68%**
	89,69%
k5	**10,64%**
	89,81%
k6	**10,57%**
	90,25%
k7	**10,08%**
	90,56%
k8	**10,89%**
	89,68%
k9	**10,72%**
	89,75%
k10	**10,15%**
	90,02%

Στον τελευταίο πίνακα υπολογίστηκαν οι μέσοι όροι των ποσοστών του πίνακα 1. Αν από τις τιμές που αποδόθηκαν στα δεδομένα ληφθούν υπόψη μόνο τα πρόσημα προκύπτει ότι *τα ποσοστά επιτυχίας για την αρνητική κλάση θα είναι 100% και για την θετική 0%*. Η διατήρηση των υπολογισμένων τιμών και η παρουσίασή τους θα οδηγήσει τελικά στην εναλλακτική μέθοδο που προτείνεται για σωστότερη απόδοση ετικετών παρακάτω.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα για όλους του αλγορίθμους. Θα δειχθούν για κάθε αλγόριθμο η μεγέθυνση της τελικής εικόνας των υπολογισμένων τιμών των ετικετών, δηλαδή εικόνες αντίστοιχες της εικόνας 2, και ο πίνακας με τους μέσους όρους των ποσοστών επιτυχίας. Αναλυτικότερα αποτελέσματα μπορούν να ληφθούν από την μελέτη του πίνακα AllLabels που προκύπτει από κάθε μέθοδο. Οι κώδικές βρίσκονται στο τέλος της εργασίας.

Αλγόριθμος 1, ***Διάδοση ετικέτας,*** *(Zhu και Ghahramani (2002))*



Εικόνα 3.3: Αποτελέσματα απόδοσης ετικετών για αλγόριθμο 1. Φαίνονται οι καθαρές τιμές των ετικετών και όχι το πρόσημο. Στον κάθετο άξονα τοποθετούνται οι τιμές των ετικετών. Στον οριζόντιο τα στοιχεία.

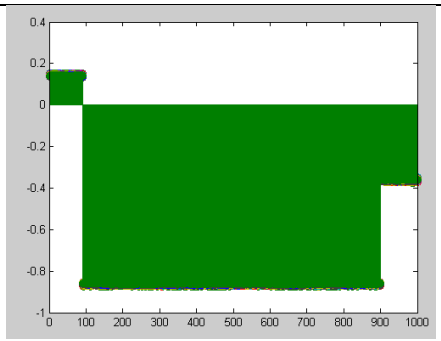
Πίνακας 3: Μέσος όρος εγγύτητας πραγματικής τιμής από αλγόριθμο 1.

k1	**10,29%**
	89,77%
k2	**10,79%**
	90,77%
k3	**10,38%**
	89,90%
k4	**10,68%**
	89,68%
k5	**10,65%**
	89,81%
k6	**10,58%**
	90,26%
k7	**10,08%**
	90,56%
k8	**10,89%**
	89,68%
k9	**10,73%**
	89,75%
k10	**10,15%**
	90,02%

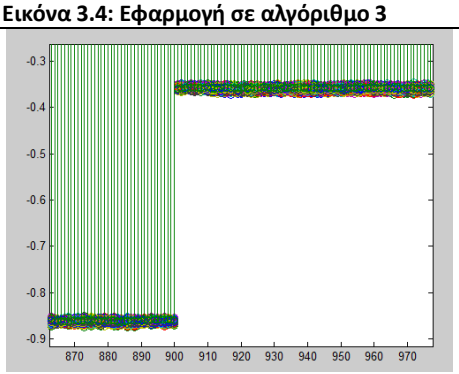
Ποσοστό επιτυχίας εύρεσης θετικού πρόσημου σε θετικά δεδομένα : 0%

Ποσοστό επιτυχίας εύρεσης αρνητικού πρόσημου σε αρνητικά δεδομένα: 100%

*Αλγόριθμος 3, **Διάδοση ετικέτας,** (Zhou et al., 2004)*



Εικόνα 3.4: Εφαρμογή σε αλγόριθμο 3



Εικόνα 3.5: Εφαρμογή σε αλγόριθμο 3, μεγέθυνση

Πίνακας 4: Μέσος όρος εγγύτητας πραγματικής τιμής από αλγόριθμο 1

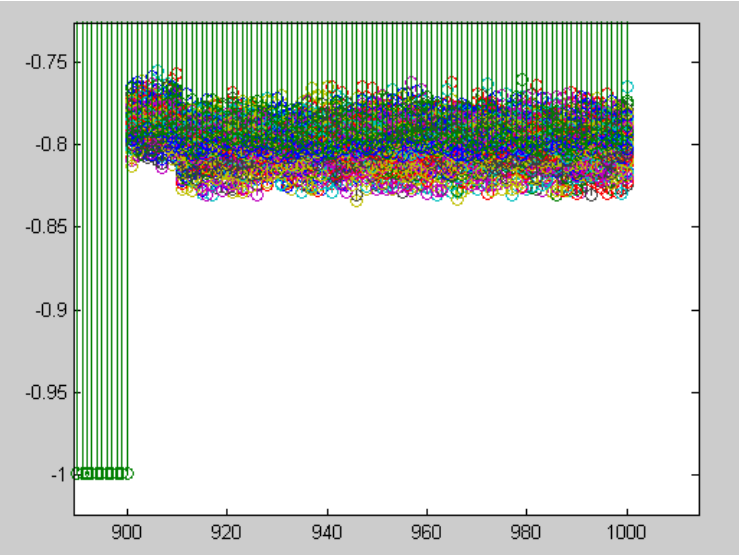
k1	32,10%
	67,87%
k2	32,06%
	68,31%
k3	32,11%
	67,93%
k4	32,23%
	67,83%
k5	32,20%
	67,88%
k6	32,09%
	68,08%
k7	31,91%
	68,22%
k8	32,29%
	67,82%
k9	32,24%
	67,85%
k10	32,02%
	67,98%

Ποσοστό επιτυχίας εύρεσης θετικού πρόσημου σε θετικά δεδομένα : 0%

Ποσοστό επιτυχίας εύρεσης αρνητικού πρόσημου σε αρνητικά δεδομένα: 100%

Ο συγκεκριμένος αλγόριθμος παρουσιάσει μεγάλο ενδιαφέρων. Αποδείχθηκε θεωρητικά ότι είναι ανάλογος των υπολοίπων, αλλά μέχρι κάποια τιμή των παραμέτρων του. Οι δοκιμές με διαφορετικές τιμές στην παράμετρο *α* το απέδειξαν και στην πράξη. Μάλιστα, η ομοιότητά του με τους υπόλοιπους αλγορίθμους γίνεται μεγαλύτερη για μεγαλύτερα άλφα, σχετικά κοντά στο ένα. Είναι, επίσης, ο μόνος από τους αλγορίθμους που αλλάζει τόσο πολύ την τιμή των labeled δεδομένων, αν και αυτό δεν είναι ιδιαίτερα σημαντικό, διότι, σύμφωνα με τον ορισμό των αλγορίθμων, λαμβάνεται υπόψη, μόνο το πρόσημο που αποδίδεται στα δεδομένα. Το ποσοστό επιτυχίας εύρεσης αρνητικού πρόσημου σε αρνητικά δεδομένα παραμένει ίσο με 100%

Αλγόριθμος 4, ***Διάδοση ετικέτας,*** *(Belkin και Niyogi, 2003b)*



Εικόνα 3.6: Αποτελέσματα απόδοσης ετικετών για αλγόριθμο 4, μεγέθυνση τελευταίων ετικετών.

Πίνακας 5: Μέσος όρος εγγύτητας σωστής τιμής για δεδομένα αλγορίθμου 4

k1	10,41%
	89,65%
k2	**10,92%**
	90,64%
k3	**10,50%**
	89,78%
k4	**10,79%**
	89,56%
k5	**10,76%**
	89,69%
k6	**10,70%**
	90,13%
k7	**10,21%**
	90,44%
k8	**11,00%**
	89,56%
k9	**10,84%**
	89,62%
k10	**10,27%**
	89,90%

Ποσοστό επιτυχίας εύρεσης θετικού πρόσημου σε θετικά δεδομένα : 0%
Ποσοστό επιτυχίας εύρεσης αρνητικού πρόσημου σε αρνητικά δεδομένα: 100%

3.2. Χρήση εκ των προτέρων γνώσης

Οι μέθοδοι δίνουν πολύ καλά αποτελέσματα για προβλήματα ταξινόμησης όπου οι κλάσεις δεν έχουν μεγάλη ανομοιογένεια ως προς τον αριθμό των στοιχείων που ανήκουν σε αυτές. Αν η μία κλάση διαφέρει αρκετά από την άλλη, απαιτείται, για την εξαγωγή σωστότερου αποτελέσματος, να γίνει χρήση της υπάρχουσας γνώσης για τις κλάσεις. Η συγκεκριμένη μεθοδολογία προτάθηκε από τους Zhu et al (2003b), έτσι ώστε το βάρος κάθε κλάσης να χρησιμοποιείται στην εξαγωγή του αποτελέσματος. Για το σκοπό αυτό χρησιμοποιήθηκε αντί του $\hat{y} \in [-1, 1]$ ένα διάνυσμα M διαστάσεων, με M να αντιπροσωπεύει τον αριθμό των κλάσεων, και $\hat{y}_{i,k}$ κάθε στοιχείο αυτού του διανύσματος, μεταξύ 0 και 1, που δίνει το βάρος κάθε k κλάσης. Εάν το M = 2 τότε το $\hat{y}_i \in [-1, 1]$ αναπαριστάτε με το διάνυσμα (1/2 (1 + $\hat{y}_i$,) 1/2 (1-$\hat{y}_i$)) ^{T} , όπου το πρώτο όρισμα είναι ο βαθμός της κλάσης 1 για το στοιχείο και το δεύτερο ο βαθμός της κλάσης -1 για το στοιχείο.

Η μέθοδος, στη συνέχεια, υπολογίζει την πιθανότητα ένα labeled στοιχείο να ανήκει στην κάθε κλάση k, και την πιθανότητα ένα unlabeled στοιχείο να ανήκει στην κλάση k. Υπολογίζεται από τις δύο πιθανότητες η μάζα της κλάσης, ένα πηλίκο που

δηλώνει την «δύναμη» της κλάσης μέσα στο δείγμα. Ο αριθμός αυτός χρησιμοποιείται για την απόδοση ετικέτας σε κάθε στοιχείο του δείγματος.

Αναλυτικότερα, αφού αποδοθούν ετικέτες με κάποιον αλγόριθμο στο σύνολο των δεδομένων, από τα lablebed δεδομένα υπολογίζεται η αρχική πιθανότητα της κλάσης k από τη σχέση

$$p_k = \frac{1}{l}\sum_{i=1}^{l} y_{i,k}.$$

Η μάζα της κλάσης k είναι η το ποσοστό των δεδομένων που αρχικά δεν είχαν ετικέτα και στα οποία δόθηκε ετικέτα της κλάσης k, δηλαδή

$$m_k = \frac{1}{u}\sum_{i=l+1}^{n} y_{i,k}.$$

Η πιθανότητα της κλάσης κανονικοποιείται αν διαιρεθεί με την μάζα της. Έτσι το πηλίκο

$$\omega_k = \frac{p_k}{m_k}$$

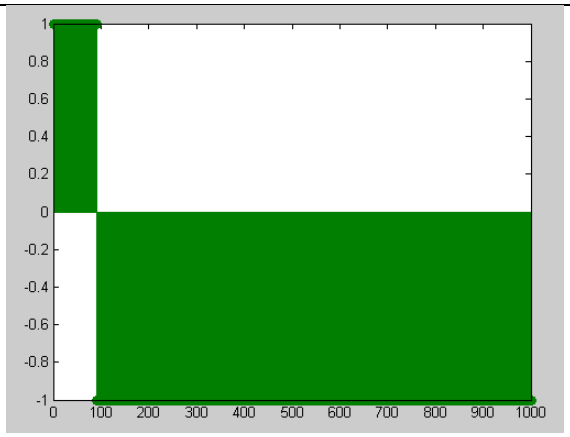
δίνει τον παράγοντα εκτίμησης της κάθε κλάσης k. Στην απόδοση ετικετών το ω_k λαμβάνει μέρος στην τελική εκτίμηση. Στο παράδειγμα του προβλήματος δύο κλάσεων, για παράδειγμα, θα μπορούσαν να εκτιμώνται οι ετικέτες από μια έκφραση της μορφή $argmax_k\left(w_k \widehat{y_{i,k}}\right)$.

Η διαδικασία αυτή έχει ως σκοπό να εναρμονίσει την αρχική κατανομή των κλάσεων με την τελική εκτιμώμενη κατανομή για τις κλάσεις. Βασική συνθήκη ώστε η μέθοδος να είναι αποτελεσματική είναι ότι τα χαρακτηρισμένα και τα μη χαρακτηρισμένα δεδομένα προέρχονται από την ίδια κατανομή. Από τις παραπάνω σχέσεις προκύπτει, ότι, για κάθε k, το ποσοστό ένα στοιχείο να ανήκει στην k είναι

$$p_k = \frac{w_k m_k}{\sum_{j=1}^{M} w_j m_j}.$$

Φυσικά η μέθοδος έχει αποτέλεσμα σε περιπτώσεις παραδειγμάτων που η μάζα κάθε κλάσης είναι διαφορετική από τις μάζες των υπολοίπων ή σε περιπτώσεις όπου οι μάζες των κλάσεων δεν έχουν μεγάλη απόκλιση, τέτοια ώστε το σύστημα μάθησης να αγνοεί τελείως κάποια εξ αυτών.

Η συγκεκριμένη μέθοδος σχεδιάστηκε και εκτελέστηκε στο Matlab και τα αποτελέσματα σχετικά με την επιτυχή απόδοση ετικέτας φαίνονται παρακάτω. Αρχικά, με κάποιον από τους αλγορίθμους αποδίδονται ετικέτες και υπολογίζεται το τελικό διάνυσμα ετικετών. Για το ή τα διανύσματα αυτά υπολογίζονται τα ζεύγη (1/2 (1 + $\hat{y}_i$,) 1/2 (1- $\hat{y}_i$)), τα οποία αποθηκεύονται στους πίνακες *Posprop* και *Negprop* αντίστοιχα, τα p_{ki} κάθε κλάσης, τα οποία αποθηκεύονται στους πίνακες *Pk1* και *Pk2* και τα ω_{ki} ομοίως στου πίνακες *W1* και *W2*. Τα επιμέρους προβλήματα είναι πάλι δύο κλάσεων, μιας αρνητικής κλάσης που ανήκουν σε αυτή το 90% των δεδομένων και μιας θετικής που μόνο το 10% των δεδομένων της ανήκουν. Τα διανύσματα επανεξετάζονται και κάθε στοιχείο τους λαμβάνει ετικέτα σύμφωνα με το $argmax_k(w_k \widehat{y_{i,k}})$. Τα αποτελέσματα της μεθόδου για τον αλγόριθμο 1 φαίνονται στην παρακάτω εικόνα, όπου στον κάθετο άξονα είναι και πάλι οι τιμές των ετικετών. Στο τελικό αποτέλεσμα λήφθηκε υπόψη το πρόσημο που αποδόθηκε σε κάθε στοιχείο. Τα πρώτα 100 είναι τα θετικά labeled δεδομένα, και τα υπόλοιπα είναι τα αρνητικά labeled δεδομένα, στα οποία ο αλγόριθμος 1 επαναφέρει τα αρχικά πρόσημα. Όλα τα τελευταία 100 χαρακτηρίζονται ως αρνητικά.



Εικόνα 3.7: Αποτέλεσμα απόδοσης ετικετών με χρήση υπάρχουσας γνώσης

Οι ετικέτες αποδίδονται σωστά στα labeled δεδομένα, ενώ όλα τα unlabeled χαρακτηρίζονται ως αρνητικά.

Η ίδια ακριβώς εικόνα λαμβάνεται και από όλους τους υπόλοιπους αλγορίθμους.

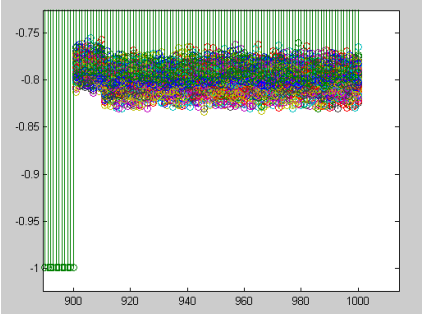
Η διορθωτική αυτή μέθοδος φαίνεται να αποτυγχάνει όταν οι κλάσεις εμφανίζουν μεγάλη ταξική ανισορροπία. Είναι προφανές ότι η μέθοδος αγνοεί 100% την μικρή θετική κλάση των δεδομένων, όπως ακριβώς και οι αλγόριθμοι της προηγούμενης ενότητας. Το βάρος της είναι πολύ μικρότερο της μεγάλης κλάσης και

οδηγεί τον αλγόριθμο στη λάθος απόφαση. Ο κώδικας της συγκεκριμένης υλοποίησης βρίσκεται στην ενότητα Κώδικες.

3.3. Εναλλακτική χρήση της εκ των προτέρων γνώσης

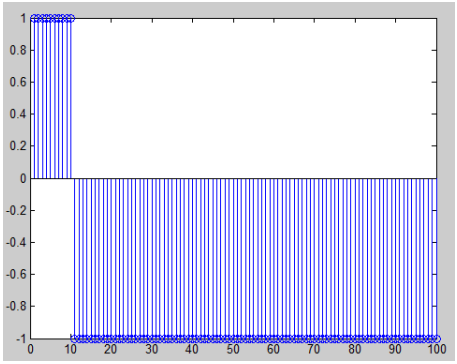
Η γνώση για το ποσοστό κάθε κλάσης, εάν παρατηρηθούν τιμές του m_k του προηγούμενου παραδείγματος είναι, με ικανοποιητική ακρίβεια, κοντά στους αριθμούς που επιθυμούνται για το m_k. Οι επιθυμητές τιμές για το m_k είναι εκείνες που αντιπροσωπεύουν τα ποσοστά κάθε κλάσης k στο δείγμα. Έτσι, για το παράδειγμα, το m_1 υπολογίζεται περίπου ίσο με 0,9, δηλαδή 90%, ενώ το m_2 περίπου ίσο με 0,1, δηλαδή 10%. Η ορθότητα των τιμών αυτών έδωσε την ώθηση για περεταίρω μελέτη των δεδομένων. Η μελέτη εστιάστηκε στο γεγονός ότι οι τιμές αυτές υπολογίζονται και είναι αντιπροσωπευτικές, άρα θα μπορούσε να βρεθεί ένας τρόπος καλύτερης εκμετάλλευσης τους, για την εξαγωγή ακριβέστερων αποτελεσμάτων.

Η μεγάλη ταξική ανισσοροπία των κλάσεων είναι ο λόγος που κανένα unlabeled δεδομένο δεν ταξινομείται στην θετική κλάση. Πάντα οι γείτονές του θα είναι κατά συντριπτική πλειοψηφία αρνητικά δεδομένα και έτσι κανένα δεδομένο από το διάνυσμα των unlabeled δε θα χαρακτηριστεί ως θετικό. Κατά τη διάρκεια των δοκιμών σε πίνακες παρατηρήθηκε ότι τα unlabeled δεδομένα που κανονικά ανήκουν στην θετική κλάση, χαρακτηρίζονται μεν ως αρνητικά ως προς το πρόσημο, έχουν, όμως, λάβει από τον αλγόριθμο τιμές μεγαλύτερες κατά μέσω όρο από εκείνες που έλαβαν τα αρνητικά unlabeled δεδομένα. Στο διάγραμμα της εικόνας 3.8 φαίνεται αρκετά καθαρά αυτό. Μάλιστα, ενώ αρχικά εναλλασσόταν μόνο το διάνυσμα των labeled θετικών δεδομένων, χωρίς να γίνονται εσωτερικές εναλλαγές στα unlabeled δεδομένα, η διαφορά φαινόταν εν μέρει, όταν εφαρμόστηκε σωστά η τεχνική του ten fold cross validation, και πραγματοποιήθηκαν συνολικά 100 αναθέσεις ετικετών, η διαφορά έγινε ιδιαίτερα εμφανής.



Εικόνα 3.8: Εικόνα από εφαρμογή σε αλγόριθμο 2, μεγέθυνση τελευταίων 100 ετικετών

Οι σωστές τιμές των mi, οι οποίες υπολογίζονται σύμφωνα με την υπόδειξη της βιβλιογραφίας, θα μπορούσαν να χρησιμοποιηθούν, όχι ως μέγιστο γινόμενο στην ανάθεση ετικετών, αλλά ως κατώφλι. Δεδομένων των mi για την θετική ή αρνητική κλάση σε κάθε υπολογισμένο πίνακα ετικετών εφαρμόστηκε αλγόριθμος στα unlabeled δεδομένα έτσι ώστε σύμφωνα με το κατώφλι που προκύπτει από το διάνυσμα των m1 και m2 να χαρακτηρίζονται εκ νέου. Αρχικά, απομονώθηκαν τα unlabeled στοιχεία, αφού πρώτα είχαν χαρακτηριστεί από κάποιον από τους αλγορίθμους. Κάθε στήλη μελετάται με χρήση επαναληπτικής μεθόδου χωριστά, αφού άλλωστε αποτελεί ένα διαφορετικό πρόβλημα. Κάθε στοιχείο της στήλης χαρακτηρίζεται εκ νέου με βάση ένα κατώφλι (ευθεία γραμμή παράλληλη στον xx' άξονα θα μπορούσε να χαρακτηριστεί) που προκύπτει από το διάνυσμα των τιμών που αρχικά αποδόθηκε στα δεδομένα. Το κατώφλι που δίνει αποτελέσματα με αντίστοιχα mi για κάθε πρόβλημα κρατείται σε πίνακα. Τα unlabeled δεδομένα χαρακτηρίζονται ξανά ως θετικά ή αρνητικά σύμφωνα με αυτό το κατώφλι. Οι ετικέτες για αυτά θα έπρεπε να παρουσιάζουν την παρακάτω εικόνα,

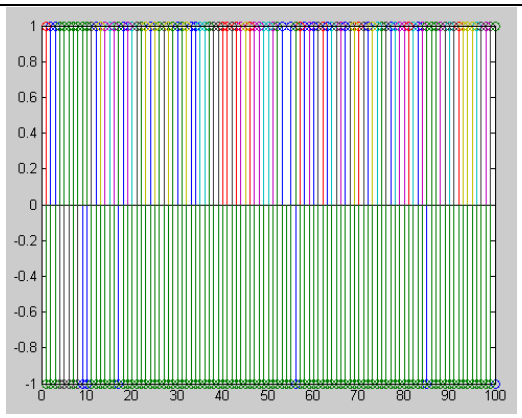


Εικόνα 3.9 : Επιθυμητό αποτέλεσμα

με 1 την ετικέτα των θετικών δεδομένων και -1 την ετικέτα των αρνητικών.

Παρακάτω φαίνονται τα αποτελέσματα μετά από εφαρμογή της μεθόδου στο χαρακτηρισμένο διάνυσμα για κάθε αλγόριθμο. Οι κώδικες της μεθόδου βρίσκονται στην ενότητα κώδικες.

Εφαρμογή σε αλγόριθμο 2



Εικόνα 3.10: Αποτέλεσμα διορθωτικής μεθόδου σε δεδομένα που είχαν χαρακτηριστεί από τον αλγόριθμο 2

Η παραπάνω εικόνα δείχνει ότι αρκετά θετικά δεδομένα έχουν χαρακτηριστεί σωστά ως θετικά, αλλά ως θετικά χαρακτηρίστηκαν και κάποια αρνητικά δεδομένα. Είναι, επίσης, εμφανές ότι το μεγαλύτερο ποσοστό των δεδομένων χαρακτηρίστηκε ως αρνητικό, κάτι που ήταν επιθυμητό και αναμενόμενο. Παρακάτω, σε αναλυτικούς πίνακες, φαίνονται τα ποσοστά επιτυχίας εύρεσης σωστής ετικέτας για την αρνητική και την θετική κλάση. Το ποσοστό εύρεσης σωστής θετικής ετικέτας αυξήθηκε πολύ, και το ποσοστό εύρεσης σωστής αρνητικής ετικέτας μειώθηκε σε σχέση με τις προηγούμενες μεθόδου. Η μείωση είναι αρκετά μικρή, μιας και ο μέσος όρος επιτυχίας εύρεσης σωστής αρνητικής ετικέτας δεν μειώνεται κάτω από 90% (δεν παύει όμως να μειώνεται). Τα αποτελέσματα παρουσιάζονται και με τη μορφή διαγραμμάτων.

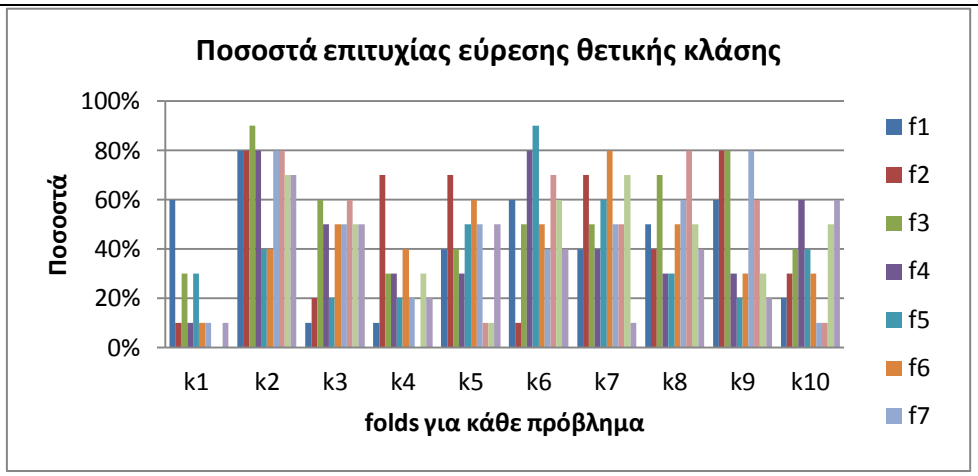
Τα ποσοστά επιτυχίας στις δύο κλάσεις για τους 100 πίνακες ετικετών που προκύπτουν από τα 100 συνολικά προβλήματα που διαμορφώθηκαν στην προηγούμενη ενότητα είναι μετά την εφαρμογή του κατωφλιού στον αλγόριθμο 2:

Ποσοστά επιτυχίας εύρεσης θετικής κλάσης

	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10
f1	50%	80%	10%	10%	40%	60%	60%	50%	60%	20%
f2	10%	80%	20%	80%	70%	10%	70%	40%	80%	30%
f3	30%	90%	60%	30%	40%	50%	50%	70%	80%	40%
f4	10%	80%	50%	30%	30%	80%	40%	30%	20%	60%
f5	20%	40%	20%	20%	50%	90%	60%	30%	20%	40%
f6	10%	40%	50%	40%	60%	50%	80%	50%	30%	30%
f7	10%	80%	50%	20%	50%	40%	50%	60%	80%	10%
f8	0%	80%	60%	50%	10%	70%	50%	80%	60%	10%
f9	0%	70%	50%	30%	10%	60%	70%	40%	30%	50%
f10	10%	70%	50%	20%	50%	40%	10%	70%	20%	60%

Πίνακας 6

ή σε μορφή γραφήματος



Εικόνα 3.11: Γράφημα ποσοστών επιτυχίας εύρεσης θετικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 2

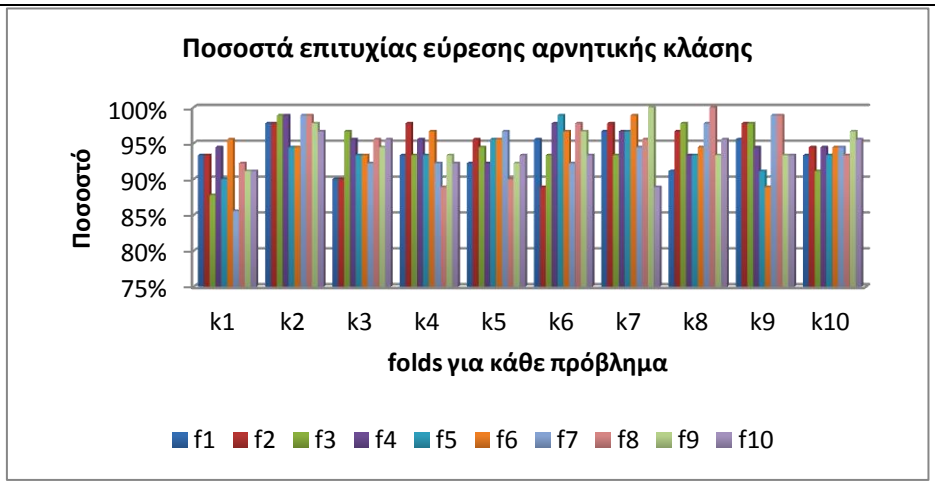
Κάποιες, ελάχιστες σε αριθμό, δοκιμές απέδωσαν μηδενικά ποσοστά επιτυχίας, όμως, σε αρκετές, το ποσοστό επιτυχίας είναι αρκετά ικανοποιητικό, ειδικά αν συγκριθεί με τα μηδενικά ποσοστά επιτυχίας εύρεσης θετικής κλάσης των προηγούμενων δύο τρόπων αντιμετώπισης του προβλήματος. Το μεγαλύτερο ποσοστό αγγίζει το 80%, ενώ αρκετά είναι και τα folds που είχαν ποσοστό επιτυχίας πάνω από 50%. Τα ποσοστά επιτυχίας για την αρνητική κλάση είναι:

Ποσοστά επιτυχίας εύρεσης αρνητικής κλάση

	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10
f1	92%	98%	94%	92%	97%	96%	93%	91%	94%	93%
f2	92%	98%	90%	97%	96%	90%	98%	97%	98%	93%
f3	88%	99%	96%	93%	98%	93%	94%	98%	98%	91%
f4	94%	99%	94%	94%	92%	98%	92%	93%	94%	94%
f5	91%	94%	93%	92%	97%	99%	97%	91%	91%	93%
f6	96%	96%	93%	96%	97%	97%	99%	94%	93%	93%
f7	96%	98%	93%	92%	97%	93%	94%	99%	99%	94%
f8	92%	99%	96%	89%	88%	98%	96%	100%	96%	94%
f9	88%	98%	94%	93%	92%	93%	100%	94%	94%	97%
f10	91%	97%	96%	90%	93%	93%	89%	94%	93%	97%

Πίνακας 7

ή



Εικόνα 3.12: Γράφημα ποσοστών επιτυχίας εύρεσης αρνητικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 2

Η μείωση του 100% που προέκυπτε για αυτήν στις προηγούμενες εφαρμογές είναι γεγονός. Είναι όμως αρκετά μικρή.

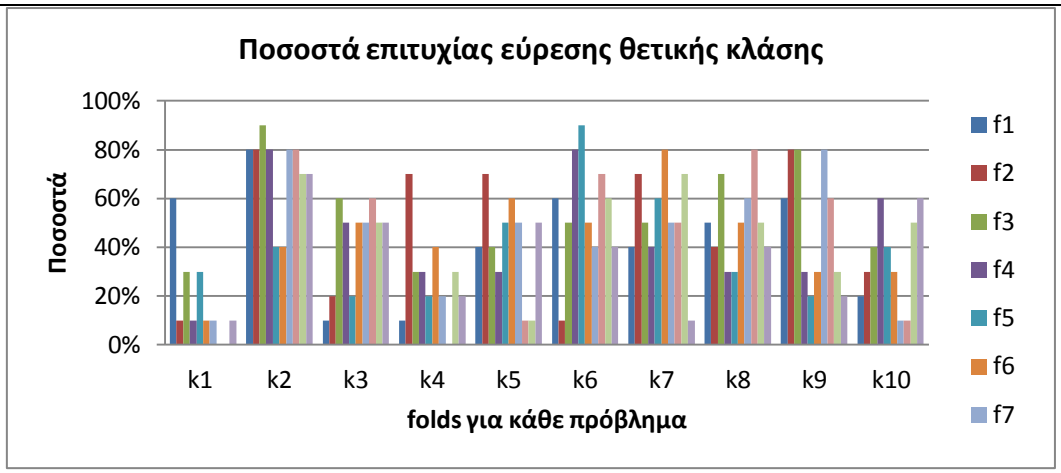
Εφαρμογή σε αλγόριθμο 1

Ποσοστά επιτυχίας εύρεσης θετικής κλάσης

	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10
f1	60%	80%	10%	10%	40%	60%	40%	50%	60%	20%
f2	10%	80%	20%	70%	70%	10%	70%	40%	80%	30%
f3	30%	90%	60%	30%	40%	50%	50%	70%	80%	40%
f4	10%	80%	50%	30%	30%	80%	40%	30%	30%	60%
f5	30%	40%	20%	20%	50%	90%	60%	30%	20%	40%
f6	10%	40%	50%	40%	60%	50%	80%	50%	30%	30%
f7	10%	80%	50%	20%	50%	40%	50%	60%	80%	10%
f8	0%	80%	60%	0%	10%	70%	50%	80%	60%	10%
f9	0%	70%	50%	30%	10%	60%	70%	50%	30%	50%
f10	10%	70%	50%	20%	50%	40%	10%	40%	20%	60%

Πίνακας 8

ή



Εικόνα 3.13: Γράφημα ποσοστών επιτυχίας εύρεσης θετικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 1

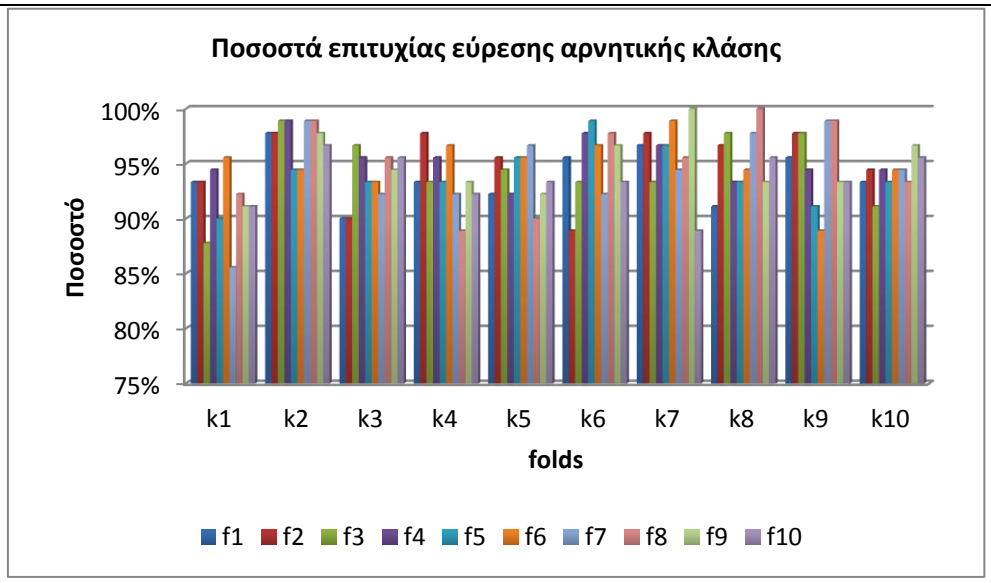
Μηδενικά ποσοστά επιτυχίας παρατηρούνται και εδώ, όμως, τα ποσοστά επιτυχίας είναι σαφώς βελτιωμένα, φτάνοντας μέχρι και 90% για κάποια fold.

Ποσοστά επιτυχίας εύρεσης αρνητικής κλάσης

	k1	k2	k3	k4	k5	k6	k7	k8	k9	k10
f1	92%	98%	94%	92%	97%	96%	97%	91%	93%	93%
f2	91%	98%	90%	98%	96%	90%	98%	97%	98%	93%
f3	87%	99%	96%	93%	98%	94%	93%	98%	97%	96%
f4	94%	99%	93%	94%	92%	98%	97%	93%	94%	94%
f5	90%	94%	93%	92%	97%	99%	97%	91%	90%	93%
f6	94%	96%	93%	96%	97%	97%	99%	94%	93%	93%
f7	93%	98%	93%	91%	97%	93%	94%	98%	99%	94%
f8	92%	99%	96%	92%	88%	98%	96%	100%	96%	94%
f9	88%	98%	94%	93%	92%	93%	100%	92%	94%	97%
f10	91%	97%	96%	90%	93%	93%	89%	98%	93%	96%

Πίνακας 9

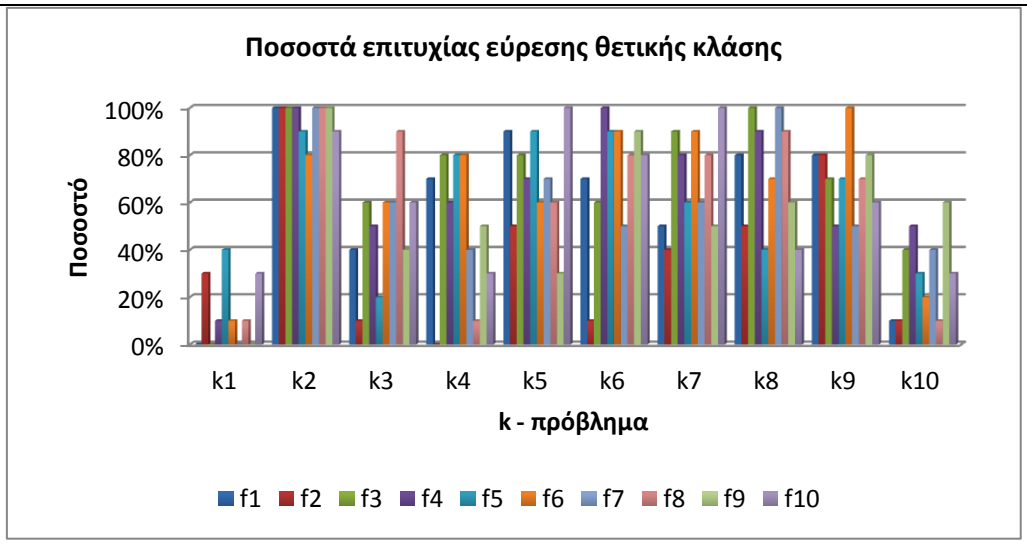
ή



Εικόνα 3.14: Γράφημα ποσοστών επιτυχίας εύρεσης αρνητικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 1

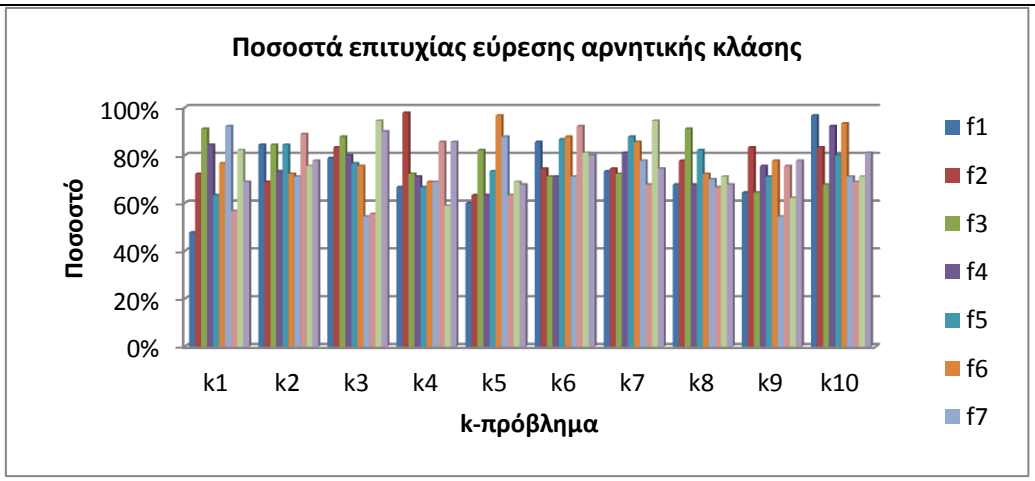
Τα αποτελέσματα είναι απολύτως όμοια με τα του αλγορίθμου 2. Για τους υπόλοιπους αλγορίθμους παραθέτονται μόνο τα διαγράμματα, για πιο συνοπτική παρουσίαση αποτελεσμάτων.

Εφαρμογή σε αλγόριθμο 3



Εικόνα 3.15: Γράφημα ποσοστών επιτυχίας εύρεσης θετικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 3

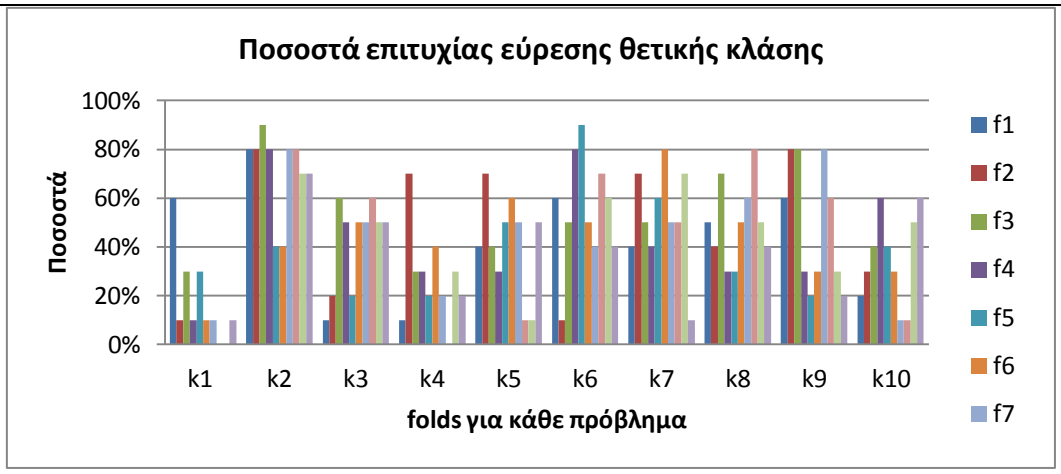
Οι τιμές που αποδίδει ο αλγόριθμος 3 στα δεδομένα της θετικής κλάσης διαφοροποιούνται λιγότερο από τις ετικέτες που αποδίδονται στα δεδομένα της αρνητικής. Ενώ η διαφορά για όλους τους υπόλοιπους είναι στο τρίτο δεκαδικό ψηφίο, για τον αλγόριθμο 3 είναι στο δεύτερο. Τα ποσοστά επιτυχίας του είναι, κατά μέσο όρο λίγο χαμηλότερα από των υπολοίπων.



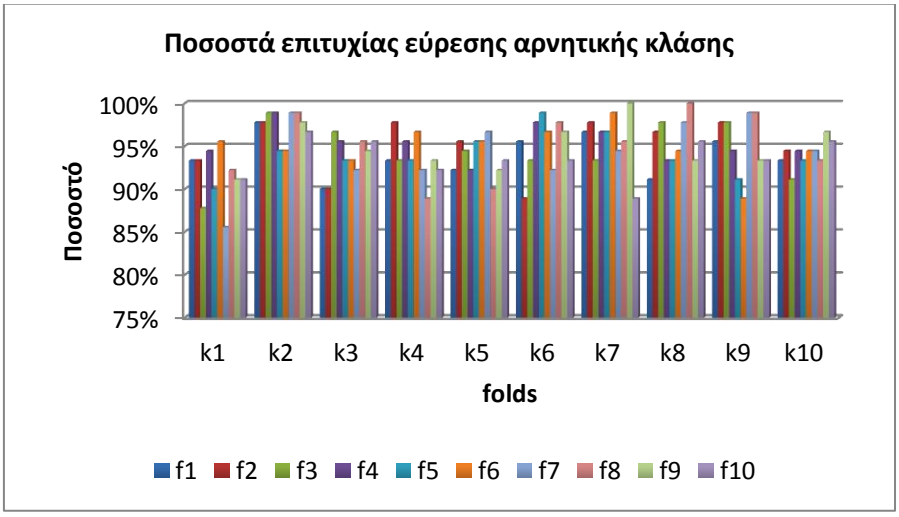
Εικόνα 3.16: Γράφημα ποσοστών επιτυχίας εύρεσης αρνητικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 3

Τα αποτελέσματα είναι αρκετά καλά και αν ληφθεί υπόψη μόνο το πρόσημο, είναι απόλυτα σωστά για την αρνητική κλάση. Η απόσταση από το -1 κάθε ετικέτας είναι, όμως, μεγαλύτερη, γεγονός στο οποίο οφείλεται η πτώση αυτών των ποσοστών σε σχέση με εκείνα των υπολοίπων αλγορίθμων.

Εφαρμογή σε αλγόριθμο 4



Εικόνα 3.17: Γράφημα ποσοστών επιτυχίας εύρεσης θετικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 4



Εικόνα 3.18: Γράφημα ποσοστών επιτυχίας εύρεσης αρνητικής κλάσης, εφαρμογή σε δεδομένα αλγόριθμου 4

3.4. Συμπεράσματα

Η απόλυτη αποτυχία εύρεσης της μικρής κλάσης στο πρόβλημα μεγάλης ταξικής ανισορροπίας, για το οποίο χρησιμοποιήθηκαν οι αλγόριθμοι, έκανε απόλυτα ξεκάθαρο ότι θα έπρεπε να γίνει χρήση της εκ των προτέρων γνώσης για το δείγμα. Η επιλογή της τιμής της ετικέτας, δηλαδή του κατάλληλου χαρακτηρισμού για κάθε δεδομένο, με βάση το μέγιστο γινόμενο ή απλούστερα με βάση την πλειοψηφία της γνώσης που λαμβάνει από γειτονικά δεδομένα έδινε αποτελέσματα που αγνοούσαν την θετική, μικρή κλάση.

Η προσεκτικότερη παρατήρηση των αποδιδόμενων ετικετών, οδήγησε στο συμπέρασμα ότι η υπάρχουσα γνώση για τις κλάσεις θα μπορούσε να χρησιμοποιηθεί με τρόπο ώστε να υποχρεώνει τα δεδομένα να λάβουν τις σωστές τιμές που

υποδεικνύει αυτή η γνώση. Στην τάξη του τρίτου δεκαδικού ψηφίου τα θετικά unlabeled δεδομένα λάμβαναν μεγαλύτερες τιμές από τα αρνητικά και αυτό δεν φαινόταν να ήταν κάτι τυχαίο. Έτσι χρησιμοποιήθηκε στα χαρακτηρισμένα δεδομένα ένα κατώφλι που κυμαινόταν από τη μέγιστη έως την ελάχιστη τιμή που λάμβαναν οι ετικέτες. Το κατώφλι που αντιστοιχούσε σε αποτελέσματα κοντινότερα στην υπάρχουσα γνώση χρησιμοποιήθηκε για τελική απόδοση ετικετών. Τα αποτελέσματα ήταν πολύ καλά στην πλειοψηφία τους, και η θετική κλάση δεν αγνοούνταν. Παρατηρήθηκε, όμως, μικρή μείωση στο ποσοστό επιτυχίας εύρεσης της αρνητικής κλάσης.

Περεταίρω επεκτάσεις θα μπορούσαν να περιλαμβάνουν την μείωση ακόμα περισσότερο στην αποτυχία εύρεσης των σωστών ετικετών των κλάσεων, την εκτενέστερη μελέτη την μεταβλητών κάθε αλγορίθμου, την επεκτασιμότητα σχετικά με το είδος των δεδομένων και τον τρόπο εισαγωγής του στους αλγορίθμους και η δυνατότητα εισαγωγής καινούριων άγνωστων δεδομένων στο σύστημα μετά την ολοκλήρωση της διαδικασίας.

4. Άλλες διατυπώσεις - Επεκτάσεις

Μια προσπάθεια αναφοράς σε άλλες αναλύσεις και διατυπώσεις, καθώς και σε πιθανές επεκτάσεις γίνεται στη συνέχεια. Ο μεγαλύτερος όγκος βιβλιογραφίας σχετικά με το θέμα των μερικώς επιτηρούμενων αλγορίθμων που βασίζονται σε γράφους είναι σχετικά πρόσφατος και η συγκέντρωση και ταξινόμησή του είναι δύσκολη. Επίσης, λόγω της πρακτικής φύσης του προβλήματος πολλές αναφορές γίνονται σε πειράματα ή επεκτάσεις και βελτιώσεις υπαρχόντων μεθόδων.

Παρουσιάζονται θέματα που θεωρήθηκαν σημαντικά και ενδιαφέρονται στα πλαίσια της συγκεκριμένης εργασίας, και όχι το σύνολο των σχετικών δημοσιεύσεων και προτάσεων. Αρχικά, περιγράφεται η διαφορά μεταξύ μεταγωγικών και επαγωγικών μεθόδων και ακολουθούν μερικοί περιορισμοί των μεθόδων και προσπάθειες ή προτάσεις βελτίωσης.

4.1. Διαφορά μεταξύ μεταγωγικής και επαγωγικής μάθησης

Το χαρακτηριστικό γνώρισμα της μάθηση με ημι – επιτήρηση είναι η χρήση ταξινομημένων και μη ταξινομημένων δεδομένων για την εκπαίδευση του μαθητευόμενου συστήματος. Το συγκεκριμένο είδος μάθηση μπορεί να είναι επαγωγικό ή μεταγωγικό.

Μεταγωγική (transductive) ονομάζεται η μάθηση που καθιστά το σύστημα δυνατό να ταξινομήσει μόνο δεδομένα που εισέρχονται σε αυτό εξ αρχής, μαζί με τα ταξινομημένα και μη ταξινομημένα δεδομένα εκπαίδευσης. Τα δεδομένα, δηλαδή, που χρησιμοποιούνται για δοκιμή ή απλά για ταξινόμηση από το σύστημα, υπάρχουν ήδη μέσα σε αυτό. Οι πρώτες γραφικές μέθοδοι ήταν μεταγωγικές.

Επαγωγική *(inductive)* ονομάζεται η μάθηση που καθιστά το σύστημα δυνατό να χαρακτηρίσει δεδομένα που εισέρχονται σε αυτό μετά την ολοκλήρωση της εκπαίδευσής του. Πρόκειται για δεδομένα που το σύστημα δεν έχει έρθει ξανά σε επαφή με αυτά, και καλείται να τα ταξινομήσει κατάλληλα, αφού έχει εκπαιδευτεί.

Οι αλγόριθμοι της προηγούμενης ενότητας είναι μεταγωγικοί. Έχουν την δυνατότητα να ταξινομήσουν δεδομένα που έχουν εισαχθεί από την αρχή στο σύστημα. Η επίλυση του συστήματος και η εξαγωγή αποτελεσμάτων σε περίπτωση που νέα δεδομένα εισέρχονται στο σύστημα συνεχώς, απαιτεί εκ νέου υπολογισμό και χρήση των αλγορίθμων, με αποτέλεσμα το κόστος για την απόδοση ετικετών να αυξάνεται πολύ. Το κριτήριο κόστους που παρουσιάστηκε αναλυτικά μετασχηματίζεται για την περίπτωση που συνεχώς εισέρχονται δεδομένα στη μορφή

$$\hat{y} = \frac{\left(\sum_j W_X(x, x_j)\hat{y}_j\right)}{\sum_j W_X(x, x_j) + \varepsilon}, \qquad (4.1)$$

όπου y1, . . . , ˆyn είναι οι ετικέτες των δεδομένων που έχουν ήδη ταξινομηθεί με χρήση των αλγορίθμων, $\hat{y}$ η ζητούμενη ετικέτα του καινούριου προς ταξινόμηση στοιχείου x και W_X η συνάρτηση που παρήγαγε τον πίνακα W του X = (x1,…,xn). Η σχέση προέκυψε, ομοίως με ανωτέρω, με μηδενισμό της παραγώγου ως προς $\hat{y}$ της σχέσης

$$C(\hat{y}_1, \dots, \hat{y}_n, \hat{y}) = \sigma\tau\alpha\theta\varepsilon\rho\acute{\alpha} + \mu\left(\sum_J W_X(x,\ x_j)\left(\hat{y} - \hat{y}_j\right)^2 + \varepsilon\hat{y}^2\right). \quad (4.2)$$

Ο τύπος είναι αρκετά απλός και οι υπολογιστικές του απαιτήσεις εξαρτώνται γραμμικά από τον αριθμό των δεδομένων που έχουν ήδη ταξινομηθεί από το σύστημα. Εάν ο W_X υπολογιστεί με βάση τον Γκαουσιανό πυρήνα, ο μεταγωγικός τύπος είναι παρόμοιος με την μέθοδο Nadaraya-Watson (1964), με τη διαφορά ότι χρησιμοποιούνται στοιχεία από δεδομένα που έχουν ταξινομηθεί με τους αλγορίθμους με ημι – επιτήρηση, και όχι δεδομένα εξ αρχής ταξινομημένα.

4.2. Περιορισμοί - Βελτιώσεις – Προτάσεις

Οι εφαρμογές των graph - based ssl μέθοδων αυξάνονται σε πλήθος συνεχώς, και εξελίσσονται, τόσο ως προς τον τρόπο χρήσης υπαρχόντων μεθόδων, όσο και ως προς το θεωρητικό πλαίσιο που τις διέπει, δίνοντας έναυσμα για εκ νέου μελέτη και καινούριες προτάσεις. Υπάρχουν, όμως, τρείς βασικοί κοινοί περιορισμοί που γίνονται προσπάθειες άρσης τους ή παράληψής τους. Οι περιορισμοί είναι:

i. Σύνθετα δεδομένα είναι αναγκασμένα να θεωρούνται ότι ανήκουν σε ένα μονό manifold
ii. Η μάθηση πρέπει να γίνεται κατά δεσμίδες (batch mode)
iii. Η ετικέτα στόχος θεωρείται ομαλή στο manifold

Ο πρώτος περιορισμός προκύπτει γιατί όλες οι μέθοδοι, μέχρι και αρκετά πρόσφατα, θεωρούν ότι το διάνυσμα των εισαγόμενων δεδομένων X είναι μονό manifold ή πολύ καλά διαχωρισμένα manifolds. Ως εκ τούτου, ο πίνακας γειτνίασης κατασκευάζεται με βάση τους k κοντινότερους γείτονες ή με βάση γκαουσιανούς πυρήνες, με την απόσταση μεταξύ των δεδομένων να υπολογίζεται ως ευκλείδεια απόσταση. Και στις δύο περιπτώσεις κοντινοί κόμβοι-δεδομένα συνδέονται με μεγάλα βάρη και αναμένεται να έχουν κοντινές ετικέτες. Όμως, αν πρόκειται για multimedia δεδομένα, η κατανομή των αντικειμένων μπορεί να διαμορφώνουν πολλαπλά manifolds τα οποία τέμνονται ή αλληλεπικαλύπτονται, όπως για παράδειγμα στην κατάτμηση κίνησης από εικόνες video, όπου κοντινά αντικείμενα από διαφορετικά manifolds δεν ικανοποιούν την παραδοχή της ομαλότητας στο manifold.

Σύμφωνα με τον δεύτερο περιορισμό οι υπάρχουσες μέθοδοι μαθαίνουν κατά δεσμίδες, δηλαδή απαιτούν το σύνολο των δεδομένων προς εκπαίδευση να είναι διαθέσιμο όλο σε μία είσοδο. Αυτή η δέσμευση δεν επιτρέπει στις μεθόδους να εφαρμόσουν άμεση μάθηση σε εφαρμογές όπως η αναγνώριση αντικειμένων από robot, το οποίο διαθέτει κάμερα και συλλέγει συνεχώς εικόνες για τα αντικείμενα που βρίσκονται γύρω του. Παρότι μπορεί να γίνει περιοδική και επιλεκτική ενημέρωση για το είδος των αντικειμένων που έχουν συλλεχθεί στις εικόνες (και αυτό είναι ημι επιτηρούμενη μάθηση), δεν έχει εφαρμοστεί άμεση ημι-επιτηρούμενη μάθηση, η οποία θα μειώσει και κατά πολύ τον όγκο των αποθηκευμένων δεδομένων – εικόνων.

Ο τρίτος περιορισμός αφορά την ομαλότητα των ετικετών στο γράφο. Η σχέση μεταξύ της ετικέτας και του αντίστοιχου γράφου του προβλήματος μελετάται από την πλευρά της αρμονικής ανάλυσης. Η παραδοχή της ομαλότητας είναι ανάλογη με την προτίμηση χαμηλής συχνότητας συνιστωσών του φάσματος του γραφήματος.

Πρόσφατες εξελίξεις στο comprehensive sensing[12] επιτρέπουν την μάθηση από έναν αυθαίρετο συνδυασμό συνιστωσών χαμηλής και υψηλής συχνότητας, αρκεί να είναι μικρός ο αριθμός των συνιστωσών.

Προτάσεις άρσεων των τριών περιορισμών έχουν γίνει από τους Zhou κ.α. το 2009 και αναλύονται στη συνέχεια.

Σχετικά με τον πρώτο περιορισμό προτάθηκε η χρήση, αντί μέτρου ομοιότητας, του μέτρου ανομοιότητας του τετραγώνου της απόστασης Hellinger (H) για σύγκριση τοπικών περιοχών, το οποίο μέτρο είναι ευαίσθητο σε τοπικές manifold δομές. Ορίζεται για δύο σημεία xi, xj, ως $H^2(p,q)$ ίση με $\frac{1}{2}\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2 dx$, με $p = \mathcal{N}(x: 0, \sum_{xi})$ και $q = \mathcal{N}(x: 0, \sum_{xj})$ να είναι μηδενικές μέσες Gaussians, είναι συμμετρική στο [0, 1], λαμβάνει μικρές τιμές όταν η τοπική γεωμετρία είναι όμοια και μεγάλες όταν εμφανίζεται μεγάλη διαφορά στην ένταση ή στην κατεύθυνση του manifold. Πριν τη χρήση της απόστασης Hellinger γίνεται εκτίμηση του τοπικού πίνακα συνδιακύμανσης $\sum_x$ γύρω από ένα τυχαία επιλεγμένο σύνολο από anchor αντικείμενα x. Το μέτρο, ύστερα, χρησιμοποιείται σε συνδυασμό με γράφο που κατασκευάζεται σύμφωνα με τους k κοντινότερους Mahalanobis γείτονες. Η κοινή χρήση των δύο μεθόδων αποδίδει γράφο ο οποίος έχει μεγάλα βάρη για κόμβους που είναι κοντά σύμφωνα με την απόσταση Mahalanobis και έχουν παρόμοια δομή συνδιακύμανσης. Η μέθοδος δοκιμάστηκε ως προς την αποτελεσματικότητά της στην αναγνώριση του συμβόλου του δολαρίου και μιας επιφάνειας στην οποία παρεμβάλλεται μια έλικα. Τα αποτελέσματα ήταν για το πρώτο σύμβολο παρόμοια με αυτά της επιτηρούμενης μεθόδου με την οποία συγκρίθηκε, ενώ για το δεύτερο σχήμα ήταν πολύ καλύτερα. Προτείνεται περαιτέρω μελέτη σχετικά με επιλογή των αρχικών anchor αντικειμένων, την επιλογή παραμέτρων και την αντικατάσταση της απόστασης Hellinger με άλλη.

Η άμεση μάθηση σε απευθείας σύνδεση (online ssl) μπορεί να επιτευχθεί μέσω αλγορίθμου που προτείνεται, και που αναθέτει ετικέτες στα δεδομένα βηματικά, με χρήση μιας συνάρτηση πρόβλεψης. Η πρόβλεψη της ετικέτας ενός επόμενου στοιχείου γίνεται ακόμα και αν δεν έχει προβλεφθεί το προηγούμενο. Κόμβοι που προσπελάστηκαν πολύ νωρίτερα δεν λαμβάνονται υπόψη στην πρόβλεψη. Επιτυγχάνεται επιθυμητή σταθερή πολυπλοκότητα μάθησης σε κάθε βήμα. Η ακρίβεια είναι συγκρίσιμη με την ακρίβεια της μάθησης κατά δεσμίδες (όχι καλύτερη). Η συγκεκριμένη μέθοδος δεν είναι βέλτιστη, εισάγει όμως την έννοια των

[12] *Compressed sensing (γνωστό και ως compressive sensing, compressive sampling ή sparse sampling)* είναι μία τεχνική ανάλυσης σήματος για την αποτελεσματική λήψη και ανακατασκευή ενός σήματος, με την εύρεση λύσης υποορισμένων γραμμικών συστημάτων. Το γεγονός αυτό εκμεταλλεύεται την σπανιότητα ή την συμπιεστότητα του σήματος σε κάποιο σημείο, επιτρέποντας σε όλο το σήμα να οριστεί από σχετικά λίγες μετρήσεις.

online ssl, ένα πεδίο πολλά υποσχόμενο αν μελετηθεί περισσότερο πρακτικά και θεωρητικά.

Τέλος, έχει ήδη αρχίσει να επιδιώκεται η άρση της απαίτηση της ομαλότητας των ετικετών στο manifold, με τη βοήθεια της μελέτης του compressive sensing. Η ομαλότητα των ετικετών δεσμεύει την εφαρμογή των μεθόδων σε πολυπλοκότερα δεδομένα, προερχόμενα από πολλές διαφορετικές κλάσεις ή cluster. Με απόδειξη ότι η επαγωγική μάθηση σε γράφους είναι αντίστοιχη του compressive sensing είναι δυνατόν να γίνουν προβλέψεις για τις τιμές των ετικετών από ένα σύνολο n labeled δεδομένων και όχι από το σύνολο των δεδομένων. Μια αρχική εφαρμογή της ιδέας είχε αρκετά καλά αποτελέσματα. Μελλοντικές επεκτάσεις αφορούν την χρήση άλλων, καταλληλότερων, φασμάτων πέραν του Laplacian, μελέτη σε τυχαίους πίνακες, περαιτέρω θεωρητική θεμελίωση.

Οι προτάσεις είναι ενδιαφέρουσες και αρκετά ενθαρρυντικές για νέες μελέτες. Δείχνουν βασικά μονοπάτια για την βελτίωση και επέκταση των γραφικών μεθόδων μάθησης με ημι-επιτήρηση.

Κώδικες Υλοποίησης

Κανονικοποίηση Πίνακα

```
function [NM] = get_normalized_matrix(M)
Min = min(M);
Max = max(M);
D = Max - Min;
D = 1./D;
Mm = [];
R = [];
[r,c] = size(M);
for k = 1:1:r
   R = [R;D];
   Mm = [Mm;Min];
end;
NM = M - Mm;
NM = NM.*R;
```

Κατασκευή πίνακα γειτνίασης

```
function [W] = ConnectivityWeightMatrix(X,type,param)
n = size(X,1);
D = dist(X,X');
switch type
   case 'full'
      sigma  = param;
      W = exp(-D/(2*sigma*sigma));
   case 'knn'
      k = ceil(param);
      if(k>n)
         error('Number of k nearest neighbors should be lesser than n');
      end;

      [SD,ID] = sort(D,2);
      F = zeros(n,n);
      for t = 1:1:n
         f(ID(t,:)) = [1:1:n];
         F(t,:) = f([1:1:n]);
      end;
      W = hardlim(k - F);
   case 'thres'
      thres = param;
      W = hardlim(thres - D);
   otherwise
      error('Invalid type parameter');
end;

end
```

Αλγόριθμος 1 Διάδοση ετικέτας, Zhu και Ghahramani (2002)

Συνάρτηση υλοποίησης αλγορίθμου

```
function [Y] = LabelPropagationZhu02(X,Yl,sigma,theta)
n = size(X,1);
l = length(Yl);
u = n - l;
W = ConnectivityWeightMatrix(X,'full',sigma);
D = diag(sum(W'));
Y = [Yl;zeros(u,1)];
diff  = 1;
while(diff > theta)
   Yold  = Y;
      Y = inv(D) * W * Y;
      Y([1:1:l]) = Yl;
      diff = dist(Y',Yold);
end;
end
```

Αλγόριθμος 2 **Διάδοση ετικέτας, (Εμπνευσμένος από τον αλγόριθμο του Jacobi)**

Συνάρτηση υλοποίησης αλγορίθμου

```
function [Y] = LabelPropagationJacobi(X,Yl,sigma,alpha,epsilon,theta)
n = size(X,1);
l = length(Yl);
u = n - l;
W = ConnectivityWeightMatrix(X,'full',sigma);
Idiag = [1:n+1:n*n];
W(Idiag) = 0;
D = diag(sum(W'));
Ddiag = D(Idiag);
mu = alpha / (1 - alpha);
Il = [ones(1,l),zeros(1,u)];
A = diag(Il + mu * Ddiag + mu * epsilon);
Y = [Yl;zeros(u,1)];
Yo = Y;
diff  = 1;
while(diff > theta)
   Yold  = Y;
   Y = inv(A) * ((mu * W * Y) + Yo);
   diff = dist(Y',Yold);
end;
end
```

Αλγόριθμος 3 **Διάδοση ετικέτας,** (Zhou et al., 2004)

Συνάρτηση υλοποίησης αλγορίθμου.

```
function [ Y ] = LabelSpreadingZhou0422( X, Yl, sigma, alpha, theta )
% Αλγόριθμος 3
```

```
%Συνολικός αριθμός δεδομένων
n = size(X,1);
%Αριθμός labeled δεδομένων
l = length(Yl);
%Αριθμός unlabeled δεδομένων
u = n - l;
% Υπολογισμός πίνακα W.
W = ConnectivityWeightMatrix(X,'full',sigma);
Idiag = [1:n+1:n*n];
W(Idiag) = 0;
% Υπολογισμός D.
D = diag(sum(W'));
%Υπολογισμές Laplacian.
L = inv(D)* W;
% Αρχικοποίηση πίνακα Y και αποθήκευση αρχικής μορφής στον Yo.
Y = [Yl;zeros(u,1)];
Yo = Y;
mu = alpha/(1-alpha);
I = ones (size(Yo,1),size(Yo,2));
% Αρχικοποίηση διαφοράς
diff  = 1;
% Επανάληψη μέχρι την σύγκλιση για απόδοση ετικετών
while(diff > theta)
   Yold  = Y;
   Y = alpha*L*Yold + (1 - alpha)* Yo;
   diff = dist(Y',Yold);
   end;
end
```

Αλγόριθμος 4 **Διάδοση ετικέτας,** (Belkin και Niyogi, 2003b)

```
function [ Y ] = LaplacianRegulation03b(X , Yl, sigma, alpha, epsilon
n = size(X,1);
l = length(Yl);
u = n - l;
W = ConnectivityWeightMatrix(X,'full',sigma);
D = diag(sum(W'));
L = D-W;
Yo = [Yl;zeros(u,1)];
mu = alpha / (1 - alpha);
Il = [ones(1,l),zeros(1,u)];
S = diag(Il);
Y = ((S+ mu*L + mu*epsilon*ones(n))^(-1))*S * Yo;
end
```

Διαμόρφωση δεδομένων πίνακα
*Παρατίθεται η εφαρμογή στον αλγόριθμο 3. Με αλλαγή στο όνομα της συνάρτησης στο εσωτερικό της δεύτερης επανάληψης***.**

```
function [Y] = ApplyinZhu04 (X)

N = importdata(X);

%Αρχικοποίηση πινάκων που χρησιμοποιούνται αργότερα. Οι πίνακες δηλώνονται κενοί, διότι αλλάζουν μέγεθος σε κάθε επανάληψη. Για διατήρηση της ανεξαρτησίας του αριθμού των κλάσεων δεν δηλώνεται το μέγεθός τους, αλλά προκύπτει από τις επαναληπτικές μεθόδους.
AllLabels = [];

%Χρήση της get_normalized_matrix για υπολογισμό του κανονιοκοποιημένου πίνακα δεδομένων
data = get_normalized_matrix(N) ;
%Αποθήκευση κανονικοποιημένων δεδομένων και σε δεύτερο πίνακα για μελλοντική χρήση.
dataoriginal = data;
% Υπολογισμός αριθμού γραμμών πίνακα.
datasize = size(data,1);

%Μεταβλητές που χρησιμοποιούνται για επιλογή ποσοστού δεδομένων για εκπαίδευση και δοκιμή. Ανεξαρτησία αριθμού κλάσεων. Όλες οι κλάσεις θεωρείται ότι αποτελούνται από 100 στοιχεία. Από αυτά, το 90% χρησιμοποιείται για εκπαίδευση και το 10% για δοκιμή.

 m = datasize*0.1;
 n = m * 0.1;
 v = m * 0.9;
 l = datasize*0.9;
 k = l*0.1;
 f = l * 0.9;

 %Δημιουργία label πινάκων
 Ylabel_only_estimated = [ones(k,1);(-1)*ones(f,1)];
 Ylabeled= [ones(k,1);(-1)*ones(f,1);ones(n,1);(-1)*ones(v,1)];

 num = datasize/100;

%Πίνακες που χρησιμοποιούνται στις επαναληπτικές μεθόδους για προσπέλαση όλων των δεδομένων του κανονικοποιημένου πίνακα δεδομένων
```

```
matrix_of_numbers = 1:num;
%Πίνακας για το k-problem
start_indices_k =[1:100:100*(num)];
stop_indices_k = 100 * matrix_of_numbers;
%Πίνακας για το f - fold
start_indices_f = [1:n:m];
stop_indices_f = [num:num:m];
%Πίνακας για την αρνητική κλάση.
start_indices_neg = [ 1 :k : l ];
stop_indices_neg = [ k: k : l];
```

%Η πρώτη επανάληψη αφορά την επιλογή της κάθε κλάσης ξεχωριστά. Οι κλάσεις θεωρούνται ίσες σε αριθμό γραμμών και στηλών. Οι γραμμές των κλάσεων είναι 100. Ο αριθμός των στηλών μπορεί να είναι οποιοσδήποτε, αρκεί οι κλάσεις να έχουν ίδιο αριθμό γραμμών, μέσα στον πίνακα δεδομένων.

```
 for i = 1 :1: num
    data = dataoriginal;
```

% Τα δεδομένα διαγράφονται και επαναφέρονται, διότι οι έτοιμες συναρτήσεις διαγράφουν γραμμές, οι οποίες είναι απαραίτητες.

```
    %Δεδομένα για την κ - κλάση
    pos_class_data = data(start_indices_k(i):stop_indices_k(i),:);
    pos_backup = pos_class_data;
    data(start_indices_k(i):stop_indices_k(i),:)=[];

    %Δεδομένα όλων των υπόλοιπων κλάσεων
    negclassdata = data;
```

% Η μεταβλητή S επιβεβαιώνει ότι τα δεδομένα της αρνητικής, δηλαδή της μεγάλης κλάσης, είναι 900. Η μεταβλητή δεν χρησιμοποιείται για άλλο λόγο, αλλά μόνο για επιβεβαίωση

```
    s = size(negclassdata,1);
```

%Η δεύτερη επανάληψη εναλλάσσει τα στοιχεία που χρησιμοποιούνται για εκπαίδευση και δοκιμή για την μικρή κλάση. Θα εναλλάσσονται επίσης ανά 100 και τα στοιχεία της μεγάλης κλάσης.

```
    for j = 1:1:num
       negclassdata = data;

       datafor_test_neg_class =
negclassdata(start_indices_neg(j):stop_indices_neg(j),:);
       negclassdata(start_indices_neg(j):stop_indices_neg(j),:) = [];
       datafor_train_neg_class= negclassdata;
```

```
        pos_class_data = pos_backup;

        % Επιλογή δεδομένων δοκιμής μικρής κλάσης.
        test_pos_class = pos_class_data(start_indices_f(j):stop_indices_f(j),:);
        pos_class_data(start_indices_f(j):stop_indices_f(j),:)=[];
        train_pos_tclass = pos_class_data;

        finalmatrixdata =
[train_pos_tclass;datafor_train_neg_class;test_pos_class;datafor_test_neg_class];



 %Επιλογή συνάρτησης. Εδώ LabelSpreadingZhou04.
  CompleteLabelMatrix =
LabelSpreadingZhou04(finalmatrixdata,Ylabel_only_estimated,1,0.00001,0.001);
  %Πίνακας με τα αποτελέσματα για όλες τις ετικέτες, απο όλα τις 100
  %δοκιμές,όλων των δεδομένων. Ο πίνακας ολοκληρώνεται μετά το τέλος των 100
επαναλήψεων.
  AllLabels  = [AllLabels, CompleteLabelMatrix];
    end
 end
 Y = AllLabels;
end
```

Υπολογισμός μαζών, βιβλιογραφική διόρθωση

%Υπολογίζονται οι πιθανότητες να ανήκει μια ετικέτα στην θετική ή την αρνητική κλάση. Υπολογίζονται το p (αρχική πιθανότητα κάθε κλάσεις) και το m (μάζα κλάσης στα unlabel δεδομένα).Υπολογίζετε το w κάθε κλάσης Πολλαπλασιασμός W1 και W2 με τον τελικό πίνακα ετικετών που προέκυψε από την χρήση του αλγορίθμου. Εύρεση του μεγαλύτερου αριθμού από τους πολλαπλασιασμούς των βαρών με τον αρχικό πίνακα ετικετών. Εφαρμόζεται ομοίως και για όλους τους υπόλοιπους αλγορίθμους με αλλαγή του ονόματος της συνάρτησης μετά το ίσον της πρώτης γραμμής.

```
X = ApplyinZhu04 ('features30.mat');

%Για αριθμό γραμμών
datasize1= size(X,1);
%Για αριθμό στηλών
datasize2=size(X,2);
%Εύρεση αριθμού κλάσεων. Θεωρώ ότι οι κλάσεις έχουν μέγεθος 100 γραμμές.
num = datasize1/10;
```

```
%Πόσοτητα στοιχείων για εκπαίδευση
l  = datasize1 * 0.9;
%Πόσοτητα στοιχείων για δοκιμή
u  = datasize1 * 0.1;

% Επαναληπτική μέθοδος σε κάθε στήλη
for i = 1:datasize2

%Εύρεση δύο στηλών, μία για την πιθανότητα της θετικής κλάσης και μία για την
πιθανότητα της αρνητικής κλάσης
      Posprop(:,i) = (1/2 * (1+X(:,i)))';
      Negprop(:,i) = (1/2 * (1-X(:,1)))';

      %Υπολογισμός Pk1 και θετική κλάση και Pk2 για αρνητική κλάση
      Pk1(i) = 1/l *(sum(Posprop(1:l,i)));
      Pk2(i) = 1/l *(sum(Negprop(1:l,i)));

      %Υπολογισμός M1 για θετική κλάση και M2 για αρνητική κλάση
      M1(i) = 1/u *(sum(Posprop((l+1):datasize1,i)));
      M2(i) = 1/u *(sum(Negprop((l+1):datasize1,i)));

      %Υπολογισμός W1 και W2.
      W1(i) = Pk1/M1;
      W2(i) = Pk2/M2;

%Πολλαπλασισμός W1 και W2 με τον τελικό πίνακα ετικετών που προέκυψε απο την
χρήση του αλγορίθμου
      Product_yest_w1(:,i) = W1(i).* X(:,i);
      Product_yest_w2(:,i) = W2(i) .* X(:,i);

end

%Επανάληψη σε κάθε στήλη και σε όλες τις γραμμές κάθε στήλης
for i = 1:datasize2
   for j = 1:datasize1

%Εύρεση του μεγαλύτερου αριθμού από τα γινόμενα των βαρών με τον αρχικό
πίνακα ετικετών
       Max(j,i) = max (Product_yest_w1(j,i),Product_yest_w2(j,i));
       %  Πρόσημο
       signmax (j, i) = sign(Max(j,i));

   end
end
```

Κατώφλι, προτεινόμενη διόρθωση

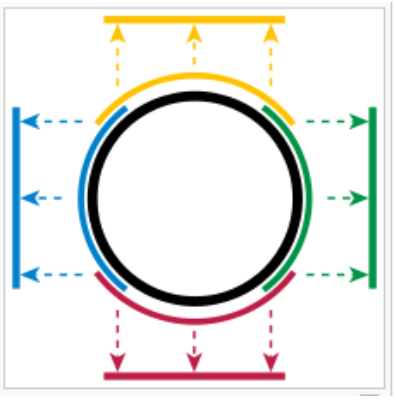
```
dist_for_m1_all=[];
Y = Posprop;

%Για αριθμό γραμμών
datasize1= size(Y,1);
%Για αριθμό στηλών
datasize2=size(Y,2);
%Εύρεση αριθμού κλάσεων. Θεωρώ ότι οι κλάσεις έχουν μέγεθος 100 γραμμές.
num = datasize1/10;

%Πόσοτητα στοιχείων για εκπαίδευση
l  = datasize1 * 0.9;
%Πόσοτητα στοιχείων για δοκιμή
u  = datasize1 * 0.1;
%Ποσότητα στοιχείων θετικής κλάσης
p = u * 0.1;
%Ποσότητα στοιχείων αρνητικής κλάσης
n = u * 0.9;

%Επιλογή μόνο των στοιχείων χώρις ετικέτα
Y_est_unlabel = Y(l+1:datasize1,:);

%Εύρεση μεγαλύτερης τιμής του πίνακα
Maxpos = max(Y_est_unlabel);
%Εύρεση μικρότερης τιμής το πίνακα
Minpos = min(Y_est_unlabel);

%Για κάθε στήλη
for i = 1:datasize2
   %Κενός πίνακας για την αποθήκευση των ποσοστών των στοιχείων που κατατάσσονται στην θετική κλάση
   dist_for_m1_all = [];
   K =  Minpos(i):0.001:Maxpos(i);
for k =1: size(K,2)
  %Κάθε στήλη εξετάζεται για κάθε k. Το k μεταβάλλεται μεταξύ της ελάχιστης και της μέγιστης τιμής του πίνακα των θετικών πιθανοτήτων.

   %Σε κάθε γραμμή θα μελετηθούν όλα τα στοιχεία
for j = 1:u
        if (Y_est_unlabel(j,i)> K(k))
            Y_est_sign(j,i)=1;
```

```
        else
            Y_est_sign(j,i)=-1;
        end
end
    % Ποσοστό στοιχείων που έχουν καταταχθεί στην θετική κλάση
     Percentage_of_pos_class = (length(find(Y_est_sign(:,i)==1)))/u;
    %Απόσταση απο το υπολογισμένο m για την θετική κλάση
     dist_from_m1 = dist(M1(i),Percentage_of_pos_class);
    % Αποθήκευση ποσοστών που δίνει κάθε k.
     dist_for_m1_all = [dist_for_m1_all;dist_from_m1];
end
     %Εύρεση κατάληλου κατωφλιού.
     [value,position] = min (dist_for_m1_all);
     %Αποθήκευση σε πίνακα όλων των κατάλληλων κατωφλιών
     Threshold(i) = K(position);
end
```

ΠΑΡΑΡΤΗΜΑ Α: Πολλαπλότητα (Manifold)

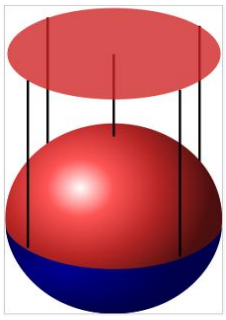
Πολλαπλότητα (manifold) διάστασης n ονομάζεται ο τοπολογικός χώρος, όπου κάθε στοιχείο του έχει μια γειτονία που είναι ομομορφική (ή τοπολογικά ισομορφική) με τον Ευκλείδειο χώρο διάστασης n. Οι γραμμές και οι κύκλοι αντιστοιχούν σε μονοδιάστατες πολλαπλότητες. Οι πολλαπλότητες δύο διαστάσεων ονομάζονται και επιφάνειες.

Κοντά σε κάθε σημείο (τοπικά) η πολλαπλότητα μοιάζει με τον Ευκλείδιο χώρο, δεν συμβαίνει όμως το ίδιο και ολικά. Η χρήση τους επιτρέπει την περιγραφή και κατανόηση πιο σύνθετων δομών σε σχέση με εκείνες που περιγράφονται επαρκώς από τον Ευκλείδειο χώρο. Τα manifold προκύπτουν από σύνολο λύσεων συστημάτων εξισώσεων ή ως γραφικές λύσεις. Ένα σημαντικό είδος manifold είναι τα διαφορίσιμα manifold, που επιτρέπουν το διαφορισμό. Μια Riemannian μετρική σε ένα manifold επιτρέπει την μέτρηση αποστάσεων και γωνιών. *Συμπλεκτικά (Symplectic)* manifolds χρησιμεύουν ως χώροι φάσης στον χαμηλτονιανό φορμαλισμό της κλασικής μηχανικής. Τεσσάρων διαστάσεων Lorentzian manifolds αναπαριστούν το χωροχρόνο στη ανάλυση της θεωρίας της σχετικότητας.



Ο κύκλος είναι το απλούστερο παράδειγμα τοπολογικής πολλαπλότητας. Η τοπολογία αγνοεί την κάμψη και ένα μικρό κομμάτι κύκλου αντιμετωπίζεται όπως και ένα μικρό κομμάτι γραμμής. Θεωρούμε, για παράδειγμα, το πάνω μισό του μοναδιαίου κύκλου $x^2 + y^2 = 1$, όπου το y είναι θετικό (κίτρινο εικόνας). Κάθε σημείο αυτού του ημικυκλίου μπορεί να περιγραφεί μοναδικά από την x μεταβλητή του. Η προβολή επί της πρώτης συντεταγμένη είναι μια συνεχής και αντιστρέψιμη χαρτογράφηση από το άνω ημικύκλιο στο ανοικτό διάστημα (-1,1).

$$X_{top}(x, y) = x.$$



Τέτοιες συναρτήσεις με τις ανοιχτές περιοχές που σχηματίζουν ονομάζονται διαγράμματα. Όμοια διαγράμματα μπορούν να προκύψουν για το κάτω ημικύκλιο (κόκκινο εικόνας), το αριστερό (μπλε εικόνας) και το δεξί (πράσινο εικόνας). Όλα μαζί καλύπτουν όλο τον κύκλο και τα τέσσερα διαγράμματα σχηματίζουν ένα άτλαντα για τον κύκλο. Στη διπλανή εικόνα φαίνεται το γράφημα της σφαίρας που προκύπτει με αναπαράσταση του θετικού ημισφαιρίου.

Το πάνω και το δεξί διάγραμμα επικαλύπτονται. Η τομή τους βρίσκεται στην περιοχή του κύκλου, όπου τόσο οι x και τα y-συντεταγμένες είναι θετικές Τα δύο διαγράμματα X_{top} και X_{right} τοποθετούν στο χάρτη αυτή την περιοχή στο διάστημα (0,1). Έτσι, μια συνάρτηση T από το (0,1) στον εαυτό της μπορεί να κατασκευασθεί, με χρήση πρώτα της αντιστροφής του άνω γραφήματος για να φθάσει τον κύκλο και στη συνέχεια το δεξί διάγραμμα πίσω στο διάστημα. Αν α ένας οποιοσδήποτε αριθμός στο (0, 1), τότε

$$T(\alpha) = X_{right}\left(X_{top}^{-1}[a]\right) = X_{right}\left(a, \sqrt{1-a^2}\right) = \sqrt{1-a^2}\,.$$

Μια τέτοια συνάρτηση καλείται χάρτης μετάβασης.

Το καθένα από τα τέσσερα διαγράμματα δείχνουν ότι ο κύκλος είναι ένα manifold, αλλά δεν σχηματίζουν το μοναδικό πιθανό άτλαντα. Τα διαγράμματα χρειάζεται αν είναι γεωμετρικά. Θεωρούμε τα διαγράμματα

$$X_{minus}(x,y) = s = \frac{y}{1+x} \;\; και \;\; X_{plus}(x,y) = t = \frac{y}{1-x},$$

με s την κλίση της γραμμής από το σημείο με συντεταγμένες x και y και το σταθερό σημείο περιστροφής (-1,0) και t το είδωλο με σημείο περιστροφής (+1,0). Η αντίστροφη χαρτογράφηση από το s στο (x,y) δίνεται από

$$x = \frac{1-s^2}{1+s^2} \;\; και \;\; y = \frac{2s}{1+s^2}\,.$$

Εύκολα αποδεικνύεται ότι $x^2 + y^2 = 1$ για κάθε τιμή της κλίσης s. Τα δυο τελευταία διαγράμματα δίνουν ένα δεύτερο άτλαντα του κύκλου με

$$t = \frac{1}{s}.$$

Είναι αδύνατον να καλυφθεί ολόκληρος ο κύκλος με ένα μόνο διάγραμμα. Το manifold δεν χρειάζεται να είναι ενιαίο ή κλειστό. Έτσι, άλλα παραδείγματα manifolds είναι δύο κύκλοι, η παραβολή, η υπερβολή και άλλα.

Σημαντικές μελέτες της πολλαπλότητας έγιναν από τους N.H. Abel και C.G. Jacobi. Αρκετά σημαντική είναι και η συνεισφορά του Riemann, του W.K. Clifford, και επίσης των S.Poisson, Hamilton, Lagrange, Euler, Poincare καθώς και αρκετών άλλων. Με την μελέτη τους έγινε εφικτός ο φορμαλισμός εννοιών της φυσικής και της κλασικής μηχανικής.

Ένα τοπολογικό manifold είναι ένας πλήρως διαχωρίσιμος Hausdorff χώρος[13] που είναι τοπικά ομομορφικός στον Ευκλείδειο χώρο. Η πλήρης διαχωρισιμότητα αποκλείει χώρους πολύ μεγάλους όπως η *μακριά γραμμή (long line)* και ο Hausdorff χώρος αποκλείει χώρους όπως η γραμμή με δύο προελεύσεις. Τοπική ομομορφία στον Ευκλείδιο χώρο σημαίνει ότι κάθε σημείο έχει μια γειτονιά ομομορφική σε μια ανοιχτή Ευκλείδεια n – μπάλα[14],

$$B^n = \{(x1, \dots, xn) \in \mathbb{R}^n | \, x1^2 + x2^2 + \cdots + xn^2 < 1 \}.$$

Ο πιο ευρέως διαδεδομένος ορισμός του manifold είναι ότι αποτελεί έναν τοπολογικό χώρο τοπικά ομομορφικό σε ένα τοπολογικό διανυσματικό χώρο πραγματικών τιμών. Ο ορισμός παραλείπει τα δύο παραπάνω αξιώματα, καθιστώντας τα manifold περισσότερο επεκτάσιμα και επιτρέπει τη μοντελοποίηση πολυπλοκότερων δομών.

Riemannian manifolds

Για να είναι δυνατή η μέτρηση αποστάσεων και γωνιών στο manifold πρέπει να είναι Riemannian. Ένα Riemannian manifolds είναι ένα διαφορίσιμο manifold στο οποίο κάθε εφαπτόμενος χώρος είναι εξοπλισμένος με εσωτερικό γινόμενο με τέτοιο τρόπο ώστε να ποικίλλει ομαλά από σημείο σε σημείο. Το αποτέλεσμα ενός εσωτερικού γινομένου είναι πραγματικός αριθμός, έτσι είναι δυνατός ο υπολογισμός διαφόρων μεγεθών όπως το μήκος, η γωνία, ο όγκος, η κλίση και άλλα μεγέθη που αποτελούν αριθμό.

[13] Στην τοπολογία ένας πλήρως διαχωρίσιμος χώρος είναι ένας τοπολογικός χώρος που ικανοποιεί το αξίωμα της μετρησιμότητας. Ένας χώρος θεωρείται ότι είναι πλήρως διαχωρίσιμος εάν η τοπολογία του έχει μετρήσιμη βάση.

[14] Μία μπάλα είναι ένας χώρος μέσα σε μια σφαίρα. Μπορεί να είναι κλειστή, να περιλαμβάνει δηλαδή και τα συνοριακά σημεία, ή ανοιχτή.

ΠΑΡΑΡΤΗΜΑ Β: Απόσταση Mahalanobis, Απόσταση Hellinger

Απόσταση Mahalanobis

Η απόσταση Mahalanobis είναι ένας τρόπος μέτρησης απόστασης, που παρουσιάστηκε από τον ινδό στατιστικολόγο P.C. Mahalanobis το 1936. Βασίζεται σε συσχετίσεις μεταξύ μεταβλητών, μέσω των οποίων διαφορετικά πρότυπα μπορούν να προσδιοριστούν και να αναλυθούν. Μετρά την ομοιότητα μεταξύ ενός άγνωστου συνόλου δεδομένων και ενός γνωστού. Διαφέρει από την ευκλείδεια απόσταση στο ότι λαμβάνει υπόψη συσχετίσεις του συνόλου των δεδομένων και είναι *scale-invariante (αναλλοίωτης κλίμακας),* είναι, δηλαδή, ανεξάρτητη από την κλίμακα των παρατηρήσεων.

Αλγεβρικά ορίζεται η απόσταση Mahalanobis για πολυδιάστατο διάνυσμα x = (x1, x2, …, xN)^T με μέσο διάνυσμα μ = (μ1, μ2, …, μN)^T και πίνακα συνδιακύμανσης S ως

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)}.$$

Η απόσταση Mahalanobis μπορεί, επίσης, να οριστεί ως μέγεθος ανομοιότητας μεταξύ δύο τυχαίων διανυσμάτων x και y της ίδιας κατανομής με πίνακα συνδιακύμανσης S ως

$$\mathrm{d}(\vec{x},\vec{y}) = \sqrt{(\mathrm{x}-\mathrm{y})^{\mathrm{T}}\, \mathrm{S}^{-1}(\mathrm{x}-\mathrm{y})}.$$

Εάν ο πίνακας συνδιακύμανσης είναι ο μοναδιαίος πίνακας, η απόσταση Mahalanobis εκφυλίζεται στην ευκλείδεια απόσταση. Εάν ο πίνακας συνδιακύμανσης είναι διαγώνιος, η προκύπτουσα απόσταση καλείται κανονικοποιημένη ευκλείδεια απόσταση

$$\mathrm{d}(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(\mathrm{xi}-\mathrm{yi})^2}{s_i^2}},$$

με si την τυπική απόκλιση των xi και yi στο σύνολο των δεδομένων.

Απόσταση Hellinger

Στην θεωρία πιθανοτήτων και την στατιστική, η απόσταση Hellinger χρησιμοποιείται για την μέτρηση της ομοιότητας μεταξύ δύο κατανομών

πιθανότητας. Είναι ένα είδος f-απόκλισης. Η απόσταση Hellinger ορίζεται σύμφωνα με το Hellinger ολοκλήρωμα, το οποίο διατυπώθηκε από τον Ernst Hellinger το 1909.

Για τον ακριβή ορισμό της απόστασης Hellinger σε όρους της απλής θεωρίας πιθανοτήτων, θεωρείται αρχικά το λ ως η μετρική Lebesgue, έτσι ώστε dP/dλ και dQ/dλ να είναι συναρτήσεις πυκνότητας πιθανότητας. Αν θεωρήσουμε τις συχνότητες f και g αντίστοιχα, το τετράγωνο της απόστασης Hellinger μπορεί να εκφραστεί ως

$$\frac{1}{2}\left(\sqrt{f(x)}-\sqrt{g(x)}\right)^2 dx = 1 - \int \sqrt{f(x)g(x)}dx,$$

όπου το δεξί μέλος προκύπτει με ανάπτυξη του τετράγωνου και αντικατάσταση των ολοκληρωμάτων των πυκνοτήτων πιθανότητας με τη μονάδα. Η απόσταση Hellinger ικανοποιεί την ιδιότητα

$$0 \leq H(P,Q) \leq 1,$$

η οποία προκύπτει από την ανισότητα Cauchy-Schwartz.

Βιβλιογραφία

[1]. Chapelle, Olivier; Schölkopf, Bernhard; Zien, Alexander (2006). *Semi-supervised learning*. Cambridge, Mass, MIT Press. ISBN 978-0-262-03358-9.

[2]. Zhu, Xiaojin (2008). Semi-supervised learning literature survey. Computer Sciences, University of Wisconsin-Madison

[3]. Zhu, Xiaojin. Semi-Supervised Learning University of Wisconsin-Madison.

[4]. Zhu, Xiaojin, (2005), *Semi-Supervised Learning with Graphs,* Doctoral Thesis, Language Technologies Institute School of Computer Science

[5]. Αναγνώριση Προτύπων και Μηχανική Μάθηση, σημειώσεις μαθήματος, Τσιχριντζης Γ.

[6]. W. E. Boyce, R.C. Diprima, *Στοιχειώδεις διαφορικές εξισώσεις και προβλήματα συνοριακών τιμών*, μετάφραση, Πανεπιστημιακές εκδόσεις Ε.Μ.Π.

[7]. Rosenberg, S. (1997). The Laplacian on a riemmannian manifold.Cambridge University Press

[8]. Chung, F. R. K., Grigor'yan, A., & Yau, S.-T. (2000). Higher eigenvalues and isoperimetric inequalities on Riemannian manifolds and graphs. Communications on Analysis and Geometry, 8, 969–1026

[9]. Belkin M., Niyogi P., (2004), *Semi-Supervised Learning on Riemannian Manifolds*, Machine Learning, 56, 209–239.

[10]. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), 2003

[11]. Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7, p.2

[12]. X. Zhu, A. B. Goldberg, T. Khor (2009). Some new directions in graph-based semi-supervised learning, Proceeding, ICME'09 Proceedings of the 2009 IEEE international conference on Multimedia and Expo, Pages 1504-1507

[13]. Partha Pratim Talukdar, Fernado Pereira, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1473–1481Uppsala, Sweden, 11-16 July 2010

[14]. WolframMathworld, http://mathworld.wolfram.com/

[15]. http://www.encyclopediaofmath.org/index.php/Manifold

[16]. http://www.holehouse.org/mlclass/01_02_Introduction_regression_analysis_and_gr.html

[17]. http://en.wikipedia.org/wiki