Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Αυτόματη Κατάταξη Μηνυμάτων Ηλεκτρονικού Ταχυδρομείου σε Κατηγορίες

Γεώργιος Σάκκης

Υπεύθυνος Καθηγητής:
Π. Σταματόπουλος

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου Παναγιώτη Σταματόπουλο για τη γενική καθοδήγηση και βοήθεια καθόλη τη διάρκεια της εργασίας. Ιδιαίτερα ευχαριστώ τους ερευνητές του Ε.ΚΕ.Φ.Ε. "Δημόκριτος" Ίωνα Ανδρουτσόπουλο και Γεώργιο Παλιούρα για την ουσιαστικότατη συμβολή τους στην εκπόνηση της εργασίας και την άριστη συνεργασία που είχαμε. Τα όποια λάθη και παραλείψεις βαρύνουν, φυσικά, εμένα. Ο Ιωάννης Κούτσιας, επίσης από το Ε.ΚΕ.Φ.Ε. "Δημόκριτος", βοήθησε σημαντικά στο πειραματικό μέρος της εργασίας. Ευχαριστώ, τέλος, το συμφοιτητή και φίλο μου Ορέστη Τελέλη για τις εποικοδομητικές συζητήσεις που είχαμε πάνω στην εργασία και την περιοχή της μηχανικής μάθησης γενικότερα.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ.....1

1.Α. Αντικείμενο της πτυχιακής εργασίας.....1
1.Β. Στόχοι της πτυχιακής εργασίας.....2
1.C. Διάρθρωση της πτυχιακής εργασίας.....3

2. ΕΠΙΣΤΗΜΟΝΙΚΟ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΟ ΥΠΟΒΑΘΡΟ.....5

2.Α. Αυτόματη κατηγοριοποίηση κειμένου.....5
2.Α.Ι. Μοντελοποίηση του προβλήματος – Ορισμοί.....6
2.Α.ΙΙ. Εφαρμογές της αυτόματης κατηγοριοποίησης κειμένου.....7
Φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου.....9
2.Β. Μηχανική μάθηση.....10
Αλγόριθμοι μηχανικής μάθησης.....13
2.Β.Ι. Μπαιυζιανή μάθηση.....14
Απλοϊκός ταξινομητής Μπαίυζ (Naïve Bayes) 16
2.Β.ΙΙ. Μάθηση βασισμένη στα στιγμιότυπα.....18
Αλγόριθμος των *k* κοντινότερων γειτόνων (*k*-Nearest Neighbor).....19
2.C. Σχεδίαση συστήματος αυτόματης κατηγοριοποίησης κειμένου.....23
2.C.I. Αναπαράσταση κειμένου.....23
Μείωση διαστασιμότητας.....25
2.C.II. Επαγωγική κατασκευή του ταξινομητή.....28
2.C.III. Αξιολόγηση του ταξινομητή.....29
2.C.III.a. Μέτρα αξιολόγησης.....29
2.C.III.b. Εκτίμηση αποτελεσματικότητας και έλεγχος υποθέσεων..... 31
2.C.IV. Σύνοψη της σχεδίασης.....34

3. ΠΕΡΙΒΑΛΛΟΝ ΔΙΕΞΑΓΩΓΗΣ ΤΩΝ ΠΕΙΡΑΜΑΤΩΝ.....37

3.Α. Συλλογή μηνυμάτων.....37
3.Β. Προεπεξεργασία και αναπαράσταση μηνυμάτων.....38
3.C. Αξιολόγηση με βάση το κόστος.....39
3.D. Αποτελέσματα προηγούμενων πειραμάτων.....43

4. ΠΕΙΡΑΜΑΤΑ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ ΤΩΝ *k*-ΚΟΝΤΙΝΟΤΕΡΩΝ ΓΕΙΤΟΝΩΝ.....47

4.Α. Παράμετροι προς διερεύνηση.....48
4.Β. Αποτίμηση χαρακτηριστικών..... 49

4.B.I. Μέτρα αποτίμησης..... 49
4.B.II. Πειραματική σύγκριση μέτρων..... 51
4.B.III. Θεωρητική διερεύνηση.....54
4.B.III.a. Ισοβαρής αποτίμηση (*EW*) 54
4.B.III.b. Σύγκριση των μέτρων *IG – GR – EW*.....55
4.B.III.c. Επίδραση της διαστασιμότητας..... 57
4.B.IV. Επίδραση της παραμέτρου *k*.....58
4.C. Αποτίμηση γειτόνων με βάση την απόσταση 63
4.C.I. Συναρτήσεις αποτίμησης γειτόνων.....63
4.C.II. Επίδραση της παραμέτρου *k*.....66
4.D. Επίδραση του μεγέθους του σώματος εκπαίδευσης.....67

5. ΠΕΙΡΑΜΑΤΑ ΜΕ ΟΜΑΔΕΣ ΤΑΞΙΝΟΜΗΤΩΝ.....69

5.A. Ομάδες ταξινομητών.....69
Συσσωρευμένη γενίκευση.....72
5.B. Κίνητρο συνδυασμού NB με *k*-NN..... 73
5.C. Σχεδιαστικές επιλογές 74
5.C.I. Συσσώρευση διασταυρωμένης επικύρωσης.....76
5.C.II. Συσσώρευση δείγματος ελέγχου..... 78
5.D. Πειραματικά αποτελέσματα.....79
Στατιστικά στοιχεία των προβλέψεων..... 83
5.E. Σύγκριση καλύτερων επιδόσεων.....85
Έλεγχος στατιστικής σημαντικότητας..... 86

6. ΑΝΑΚΕΦΑΛΑΙΩΣΗ..... 89

Προοπτικές.....90

ΑΝΑΦΟΡΕΣ.....93

1) ΕΙΣΑΓΩΓΗ

1.Α) Αντικείμενο της πτυχιακής εργασίας

Το αντικείμενο της παρούσας εργασίας είναι η αυτόματη κατηγοριοποίηση μηνυμάτων ηλεκτρονικού ταχυδρομείου (e-mail) με χρήση τεχνικών μηχανικής μάθησης. Το ενδιαφέρον εστιάζεται στη σύνθεση δύο πεδίων γνώσης: Του τεχνολογικού πεδίου της αυτόματης κατηγοριοποίησης εγγράφων (υποπερίπτωση του οποίου αποτελεί η κατηγοριοποίηση μηνυμάτων ηλεκτρονικού ταχυδρομείου) και του επιστημονικού πεδίου της μηχανικής μάθησης. Και οι δύο τομείς αποτελούν σήμερα ενεργές ερευνητικές περιοχές, οι οποίες βρίσκονται σε συνεχή ανάπτυξη, ιδιαίτερα κατά τη διάρκεια της τελευταίας δεκαετίας. Τα αποτελέσματα αυτής της έρευνας έχουν ήδη αρχίσει να περνούν και στο στάδιο των εμπορικών εφαρμογών με αξιόλογη επιτυχία, χωρίς ωστόσο να είναι αρκετά διαδεδομένα ακόμα. Είναι σίγουρο πάντως πως η χρήση προϊόντων και ολοκληρωμένων συστημάτων αυτόματης κατηγοριοποίησης εγγράφων θα ενταθεί τα προσεχή χρόνια, καθώς η τεχνογνωσία στην περιοχή αυτή θα αυξάνεται, ενώ παράλληλα η ανάγκη διαχείρισης ενός όλο και περισσότερο διογκούμενου αριθμού εγγράφων διαθέσιμων σε ηλεκτρονική μορφή, κυρίως λόγω της αλματώδους ανάπτυξης και χρήσης του Διαδικτύου, θα καταστήσει ανέφικτη ή ασύμφορη τη χειρωνακτική (manual) κατηγοριοποίηση των ηλεκτρονικών εγγράφων.

Η *κατηγοριοποίηση κειμένου* (*text categorization - ΚΚ*), γνωστή και ως *κατάταξη κειμένου* (*text classification*), είναι η διαδικασία κατάταξης κειμένων φυσικής γλώσσας σε ένα προκαθορισμένο αριθμό θεματικών κατηγοριών γνωστών εκ των προτέρων. Η ιστορία της ΚΚ, ως πεδίου έρευνας στην περιοχή της βασισμένης στο περιεχόμενο (content-based) διαχείρισης εγγράφων, ξεκίνησε στις αρχές της δεκαετίας του ’60. Ωστόσο έγινε κύριο πεδίο ενασχόλησης ενός σημαντικού αριθμού ερευνητών κατά τις αρχές της δεκαετίας του ’90, λόγω του αυξημένου ενδιαφέροντος πρακτικής αξιοποίησής της και των ισχυρών υπολογιστικών μέσων που ήταν πλέον διαθέσιμα. Σήμερα, η ΚΚ χρησιμοποιείται σε διάφορα περιβάλλοντα εφαρμογής, όπως στην ευρετηριοποίηση εγγράφων με βάση ένα ελεγχόμενο λεξικό, στο φιλτράρισμα εγγράφων, στην αυτόματη δημιουργία μεταδεδομένων, στη δημιουργία ιεραρχικών καταλόγων για πόρους του Διαδικτύου, κ.α.

Το αντικείμενο της εργασίας αυτής είναι μία ειδική εφαρμογή κατηγοριοποίησης που προσπαθεί να αντιμετωπίσει ένα συνεχώς διογκούμενο πρόβλημα: πρόκειται για το μαζικό “βομβαρδισμό” των χρηστών του ηλεκτρονικού ταχυδρομείου με διαφημιστικά μηνύματα από εταιρείες που προσπαθούν μέσω αυτού του τρόπου να προωθήσουν με ελάχιστο κόστος και κόπο τα προϊόντα και τις υπηρεσίες τους. Για τα μηνύματα αυτά έχει επικρατήσει η ονομασία “spam” e-mail (ή junk e-mail – μηνύματα-“σκουπίδια”)[*]. Αν και οι περισσότεροι χρήστες τα βρίσκουν ενοχλητικά και τα διαγράφουν αμέσως, χάνουν πολύ χρόνο προσπαθώντας να εντοπίσουν τη χρήσιμη αλληλογραφία τους. Ένα ακόμα πρόβλημα είναι

[*] Μια πιο επίσημη ονομασία είναι “Μη αιτηθείσα εμπορική ηλεκτρονική αλληλογραφία” (Unsolicited Commercial E-mail – UCE)

πως οι ανήλικοι χρήστες βρίσκονται συχνά εκτεθειμένοι σε ακατάλληλο (π.χ. πορνογραφικό) υλικό μέσω τέτοιων μηνυμάτων.

Για την αντιμετώπιση της κατάστασης, οι εμπορικές λύσεις που διατίθενται μέχρι στιγμής δίνουν τη δυνατότητα στο χρήστη να ορίσει ο ίδιος λέξεις-κλειδιά και λογικούς κανόνες με στόχο το φιλτράρισμα των spam e-mails. Αυτή η προσέγγιση είναι προβληματική, γιατί πέραν του ότι απαιτεί εμπειρία στην κατασκευή κανόνων από τους χρήστες, οι τελευταίοι πρέπει να συντηρούν και να εκλεπτύνουν τους κανόνες με την πάροδο του χρόνου, καθώς η μορφή των spam mails δεν είναι σταθερή. Θα ήταν σαφώς προτιμότερη μια λύση που αυτόματα κατατάσσει τα μηνύματα ως "θεμιτά" (legitimate) ή "αθέμιτα" (spam) και η οποία θα προσαρμόζεται επίσης αυτόματα στις αλλαγές στα χαρακτηριστικά των μηνυμάτων με το χρόνο.

Μια πολλά υποσχόμενη λύση σε αυτό το πρόβλημα, όπως και σε πολλά άλλα προβλήματα αυτόματης κατηγοριοποίησης κειμένου, αλλά και οποιασδήποτε μορφής πληροφορίας, έρχεται από το χώρο της μηχανικής μάθησης. Η *μηχανική μάθηση (machine learning)* έχει ως σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν, κατά τον τρόπο που χρησιμοποιούμε τον όρο "μάθηση" για τον άνθρωπο, δηλαδή τη βελτίωση ικανοτήτων μέσω της αξιοποίησης της συσσωρευμένης γνώσης και εμπειρίας. Η πρόοδος που έχει συντελεστεί στη μηχανική μάθηση, ιδιαίτερα την τελευταία δεκαετία, είναι σημαντική και έχει δώσει τόσο αλγορίθμους και θεωρητικά αποτελέσματα, όσο και πρακτικές εφαρμογές με μεγάλη επιτυχία.

Μία από τις περιοχές στις οποίες διείσδυσε η εφαρμογή της μηχανικής μάθησης ήταν και η ΚΚ. Μέχρι τα τέλη της δεκαετίας του '80, η πιο αποτελεσματική προσέγγιση στην ΚΚ ήταν μέσω μεθόδων *γνωσιακής μηχανικής* (*knowledge-engineering*), δηλαδή το χειρωνακτικό ορισμό λογικών κανόνων που να κωδικοποιούν την γνώση των ανθρώπων-ειδικών (experts) ως προς την κατηγοριοποίηση κειμένων. Στην επόμενη δεκαετία, η προσέγγιση αυτή ξεπεράστηκε μέσω της επικράτησης του παραδείγματος της μηχανικής μάθησης (*machine learning paradigm*). Σύμφωνα με το παράδειγμα αυτό, μια γενική επαγωγική διαδικασία δημιουργεί έναν αυτόματο ταξινομητή, "μαθαίνοντας" τα χαρακτηριστικά κάθε κατηγορίας μέσω ενός συνόλου προκαταταγμένων κειμένων από ειδικούς. Τα πλεονεκτήματα αυτού του σχήματος είναι μια ακρίβεια κατάταξης συγκρίσιμη με αυτή των ανθρώπων-ειδικών και η εξοικονόμηση ανθρώπινου δυναμικού, καθώς δεν απαιτείται η επέμβαση γνωσιολόγων-μηχανικών και ειδικών.

1.Β) Στόχοι της πτυχιακής εργασίας

Για το πρόβλημα των spam e-mails, η λύση που παρέχεται σήμερα, όπως περιγράφηκε παραπάνω, είναι βασισμένη στη λογική της γνωσιακής μηχανικής, με την απαίτηση μάλιστα κάθε χρήστης του ηλεκτρονικού ταχυδρομείου να παίζει το ρόλο του "ειδικού" στην αναγνώριση των spam mails, ορίζοντας ο ίδιος κατάλληλους κανόνες και ανανεώνοντάς τους όποτε κρίνει ότι είναι απαραίτητο. Με δεδομένη την επιτυχία των αλγορίθμων μηχανικής μάθησης σε άλλες εφαρμογές κατάταξης κειμένου, ο πρώτος στόχος της εργασίας είναι να δείξει πως η χρήση τους και για το αυτόματο φιλτράρισμα των spam e-mails παρέχει ικανοποιητική ακρίβεια. Ακόλουθος στόχος ήταν η βελτιστοποίηση της απόδοσης του

τελικού συστήματος μέσω του συντονισμού κάποιων από τις παραμέτρους που υπάρχουν ως σχεδιαστικές επιλογές. Επιπλέον, κάποιες από τις παρατηρήσεις που έγιναν κατά τη διαδικασία της βελτιστοποίησης οδήγησαν σε γενικότερα συμπεράσματα, πέραν του συγκεκριμένου πεδίου εφαρμογής, στηριζόμενα τόσο στα πειραματικά αποτελέσματα, όσο και σε θεωρητικά και διαισθητικά επιχειρήματα.

Συνοπτικά, οι στόχοι της εργασίας είναι:

- Να μελετηθεί και να παρουσιαστεί η μέχρι σήμερα δραστηριότητα στους τομείς της αυτόματης κατηγοριοποίησης κειμένου, της μηχανικής μάθησης και της εφαρμογής της δεύτερης στην πρώτη, μέσω της εκτεταμένης βιβλιογραφίας που έχει δημιουργηθεί, κατά τη διάρκεια των τελευταίων κυρίως ετών.
- Να μοντελοποιηθεί το πρόβλημα του φιλτραρίσματος των spam e-mails στο πλαίσιο της κατηγοριοποίησης κειμένου.
- Να καταδειχθεί πειραματικά η υψηλή απόδοση που επιτυγχάνεται με τη χρήση τεχνικών μηχανικής μάθησης για την αντιμετώπιση του προβλήματος.
- Να βελτιστοποιηθεί η επίδοση της μεθόδου μέσω του συντονισμού κάποιων εκ των διαθέσιμων σχεδιαστικών επιλογών.
- Να ερμηνευθούν τα αποτελέσματα των πραγματοποιηθέντων πειραμάτων και να συγκριθούν με προηγούμενα αποτελέσματα.
- Να γενικευτούν όπου είναι δυνατόν τα συμπεράσματα που έχουν προκύψει.
- Να αναφερθούν και άλλες κατευθύνσεις που δεν διερευνήθηκαν στα πλαίσια της εργασίας, αλλά προβάλλουν ως πολλά υποσχόμενες από άλλες έρευνες.

1.C) Διάρθρωση της πτυχιακής εργασίας

Η παρουσίαση της εργασίας είναι οργανωμένη ως εξής:

Στο κεφάλαιο 2 σκιαγραφούνται τα γνωστικά πεδία (domains) που αποτελούν το επιστημονικό και τεχνολογικό υπόβαθρο πάνω στο οποίο στηρίζεται η εργασία. Στο κεφάλαιο 3 περιγράφεται το περιβάλλον εκτέλεσης των πειραμάτων που έγιναν, με αναφορά στη συλλογή των ηλεκτρονικών μηνυμάτων που χρησιμοποιήθηκαν και στον τρόπο προεπεξεργασίας και αναπαράστασής τους πριν τη χρήση τους από τους αλγορίθμους μάθησης, ορίζονται κατάλληλα μέτρα αξιολόγησης της αποτελεσματικότητας ενός φίλτρου και παρουσιάζονται με βάση τα μέτρα αυτά προηγούμενα αποτελέσματα πειραμάτων πάνω στην ίδια συλλογή. Στο κεφάλαιο 4 περιγράφονται τα πειράματα που έγιναν με διάφορες παραλλαγές του αλγορίθμου μάθησης των *k-κοντινότερων γειτόνων (k-Nearest Neighbor algorithm)*. Στο κεφάλαιο 5 αναλύεται η συνδυαστική χρήση αλγορίθμων μάθησης μέσω των *ομάδων ταξινομητών (classifier ensembles)* και παρουσιάζονται τα αποτελέσματα πειραμάτων με συνδυασμό δύο αλγορίθμων μάθησης. Τέλος, στο κεφάλαιο 6 ανακεφαλαιώνονται τα κύρια ζητήματα που θίχτηκαν στην εργασία, τα καλύτερα αποτελέσματα και τα συμπεράσματα που προέκυψαν, μνημονεύονται τα σημεία που παρέμειναν ανοιχτά και προτείνονται άλλες πειραματικές κατευθύνσεις που δε διερευνήθηκαν. Η εργασία κλείνει με αναφορές στη σχετική βιβλιογραφία.

2) ΕΠΙΣΤΗΜΟΝΙΚΟ ΚΑΙ ΤΕΧΝΟΛΟΓΙΚΟ ΥΠΟΒΑΘΡΟ

Στο κεφάλαιο αυτό δίνεται μια ευρεία εικόνα του πλαισίου στο οποίο τοποθετείται η εργασία. Στην πρώτη υποενότητα μοντελοποιείται το πρόβλημα της κατηγοριοποίησης κειμένου (ΤC) και δίνονται σχετικοί ορισμοί. Γίνεται αναφορά στις πιο σημαντικές εφαρμογές της ΚΚ που έχουν αναπτυχθεί, ενώ αναλύεται το πρόβλημα της τυφλής αποστολής διαφημιστικών e-mails από εταιρείες και η έκτασή του σήμερα.

Στη δεύτερη υποενότητα παρουσιάζεται ο τομέας της μηχανικής μάθησης, με κάποια γενικά στοιχεία στην αρχή και επικεντρώνοντας στη συνέχεια σε δύο κατηγορίες αλγορίθμων μάθησης, στους οποίους βασίσθηκαν τα πειράματα που έγιναν. Οι κατηγορίες αυτές είναι η *Μπαυζιανή μάθηση* (*Bayesian Learning*), στην οποία βασίζεται ο αρκετά διαδεδομένος *απλοϊκός ταξινομητής Μπαίυζ* (*Naive Bayes classifie*r), και η *μάθηση βασισμένη στα στιγμιότυπα*, ή αλλιώς *στη μνήμη* (*Instance-Based* ή *Memory-Based Learning*), στην οποία εντάσσεται ο επίσης διαδεδομένος αλγόριθμος των *k-κοντινότερων γειτόνων* (*k-Nearest Neighbor Algorithm*).

Στην τρίτη και τελευταία υποενότητα, δίνονται τα τυπικά βήματα από τα οποία περνάει η σχεδίαση ενός συστήματος αυτόματης κατηγοριοποίησης κειμένου, με αναφορά στις σημαντικότερες επιλογές που υπάρχουν σε κάθε βήμα. Περιγράφονται οι τρόποι αναπαράστασης των εγγράφων, με έμφαση στο ζήτημα της *μείωσης της διαστασιμότητας (dimensionality reduction)*, την οποία ακολουθεί η επαγωγική δημιουργία του ταξινομητή. Τέλος, θίγεται το σημαντικό θέμα της αξιολόγησης ενός παραχθέντος ταξινομητή, η οποία είναι απαραίτητη για τις συγκρίσεις μεταξύ διαφορετικών ταξινομητών.

2.Α) Αυτόματη κατηγοριοποίηση κειμένου

Ο όρος *αυτόματη κατηγοριοποίηση κειμένου* (*automated text categorization*) έχει καθιερωθεί στη σύγχρονη βιβλιογραφία να σημαίνει, όπως αναφέρθηκε και στην εισαγωγή, τη διαδικασία αυτόματης κατάταξης κειμένων φυσικής γλώσσας σε ένα προκαθορισμένο αριθμό θεματικών κατηγοριών γνωστών εκ των προτέρων. Ο όρος "αυτόματη κατάταξη κειμένου" (automatic text classification) εμφανίζεται στην παλαιότερη κυρίως βιβλιογραφία, όπου δεν είχε διαμορφωθεί αρκετά η ορολογία, με τρεις διαφορετικές σημασίες, προκαλώντας δικαιολογημένα σύγχυση: Πέρα από τον (i) παραπάνω ορισμό για την ΚΚ, ο ίδιος όρος έχει χρησιμοποιηθεί επίσης εννοώντας (ii) τον αυτόματο ορισμό ενός συνόλου θεματικών κατηγοριών για κείμενα (σήμερα αναφέρεται διεθνώς ως *ομαδοποίηση* (*clustering)* και (iii) την αυτόματη ανάθεση κειμένων σε ένα σύνολο θεματικών κατηγοριών *μη* προκαθορισμένων (που σήμερα αναφέρεται ως *ευρετηριοποίηση ((free text) indexing*). Ό,τι ακολουθεί αναφέρεται στην πρώτη σημασία του όρου, εκτός αν δηλωθεί ρητώς το αντίθετο.

2.Α.Ι) Μοντελοποίηση του προβλήματος – Ορισμοί

Η κατηγοριοποίηση κειμένου μπορεί να οριστεί φορμαλιστικά ως το έργο της ανάθεσης μιας εκ των δύο τιμών του συνόλου {0,1} σε κάθε κελί a_{ij} του m*n *πίνακα απόφασης (decision matrix)* M_{ij}

	d_1	**…**	**…**	**d_j**	**…**	**…**	**d_n**
c_1	a_{11}	…	…	a_{1j}	…	…	a_{1n}
…	…	…	…	…	…	…	…
c_i	a_{i1}	…	…	a_{ij}	…	…	a_{in}
…	…	…	…	…	…	…	…
c_m	a_{m1}	…	…	a_{mj}	…	…	a_{mn}

όπου $C = \{c_1, \ldots, c_m\}$ είναι ένα σύνολο από προκαθορισμένες κατηγορίες, και $D = \{d_1, \ldots, d_n\}$ είναι ένα σύνολο από έγγραφα κειμένου προς ταξινόμηση. Η τιμή ανάθεσης 1 για το κελί a_{ij} υποδηλώνει την απόφαση να ταξινομηθεί το έγγραφο d_j στην κατηγορία c_i, ενώ η τιμή 0 υποδηλώνει την απόφαση να μην ταξινομηθεί το έγγραφο d_j στην κατηγορία c_i.

Πιο αυστηρά, το πρόβλημα είναι να προσεγγιστεί η άγνωστη συνάρτηση $f : D \times C \to \{0,1\}$ (που περιγράφει πώς πρέπει πραγματικά να καταταχτούν τα έγγραφα) με μια συνάρτηση $f' : D \times C \to \{0,1\}$ (η οποία καλείται *ταξινομητής*, ή *μοντέλο*, ή *υπόθεση*), τέτοια ώστε οι f και η f' να συμπίπτουν όσο το δυνατόν περισσότερο. Το πώς ορίζεται ο βαθμός σύμπτωσης των δύο (τον οποίο λέμε *αποτελεσματικότητα – effectiveness*) θα συζητηθεί παρακάτω. Η προσέγγιση γίνεται με τη βοήθεια μιας *αρχικής συλλογής (initial corpus)* $D_0 = \{\bar{d}_1, \ldots, \bar{d}_s\}$ από έγγραφα γνωστής κατηγοριοποίησης, δηλαδή των οποίων ο πίνακας απόφασης είναι δεδομένος και θεωρείται γενικά σωστός (αν και πολλοί αλγόριθμοι κατηγοριοποίησης δεν υποθέτουν απόλυτη ακρίβεια, αλλά ανέχονται και λάθη στα αρχικά δεδομένα, ή αλλιώς "θόρυβο").

Για την ορθή κατανόηση του προβλήματος, είναι θεμελιώδεις δύο παρατηρήσεις:

- Οι κατηγορίες είναι απλά συμβολικές ετικέτες. Καμία επιπλέον γνώση ως προς τη "σημασία" τους δεν θεωρείται διαθέσιμη για την κατασκευή του ταξινομητή. Με άλλα λόγια, μια κατηγορία c_i συνίσταται σε ένα σύνολο εγγράφων που θεωρούμε ότι μπορούν να ομαδοποιηθούν μαζί. Ιδιαιτέρως τονίζεται πως το όνομα c_i της κατηγορίας θεωρείται αυθαίρετο και κατά συνέπεια ένας αλγόριθμος κατηγοριοποίησης δε θα πρέπει πρέπει να το λαμβάνει υπόψη (π.χ. τη λέξη "οικονομικά" στην περίπτωση κατηγοριοποιήσεως ειδήσεων).
- Εν γένει, η κατάταξη των εγγράφων σε κατηγορίες πρέπει να βασίζεται στο περιεχόμενο του εγγράφου και όχι στα τυχόν μεταδεδομένα που υπάρχουν γι' αυτό (π.χ. συγγραφέας, ημερομηνία δημοσίευσης, κ.α.). Δηλαδή, η κατηγοριοποίηση πρέπει να βασίζεται κυρίως σε *ενδογενή* γνώση (γνώση που μπορεί να εξαχθεί από το ίδιο το έγγραφο), παρά σε *εξωγενή* γνώση (δεδομένα που προέρχονται από κάποια εξωτερική πηγή).

Με δεδομένο πως η σημασιολογία ενός εγγράφου είναι από τη φύση της μια υποκειμενική έννοια, γίνεται φανερό πως η θεμελιώδης έννοια της ΚΚ, η συσχέτιση ενός κειμένου με μια κατηγορία, δε μπορεί να αποφασιστεί ντετερμινιστικά. Είναι, άλλωστε, πολύ συχνό το φαινόμενο δύο άνθρωποι να διαφωνούν στην κρίση τους σχετικά με την κατάταξη ενός κειμένου κάτω από μία κατηγορία. Για παράδειγμα, ένα κείμενο πάνω στο θέμα της υποχρεωτικής μη αναγραφής της θρησκείας στην αστυνομική ταυτότητα μπορεί να καταταχθεί στα πολιτικά, στα νομικά, στα θρησκευτικά, στα εθνικά ή σε οποιονδήποτε συνδυασμό από τις προηγούμενες κατηγορίες, ανάλογα με την κρίση του καθενός.

Όπως φαίνεται από τον παραπάνω ορισμό της ΚΚ, στη γενική περίπτωση δεν επιβάλλεται κανένας περιορισμός στον αριθμό των εγγράφων που μπορούν να καταταχθούν υπό μια κατηγορία, ούτε στον αριθμό των κατηγοριών στις οποίες επιτρέπεται να ανήκει ένα έγγραφο. Είναι συχνά όμως απαραίτητο ή επιθυμητό, ανάλογα με την εφαρμογή, να υπάρχουν περιορισμοί στους προαναφερθέντες αριθμούς. Για παράδειγμα, μπορεί να απαιτείται κάθε κατηγορία να περιλάμβάνει ακριβώς r (ή ≥ r ή ≤ r) έγγραφα, για κάποιο δεδομένο r. Αντίστοιχα, ένα έγγραφο μπορεί να πρέπει να καταταχθεί σε ακριβώς r (ή ≥ r ή ≤ r) κατηγορίες. Σχετικά με το τελευταίο, αρκετά συχνή είναι στην πράξη η περίπτωση r=1, η οποία λέγεται και κατηγοριοποίηση *μονής ετικέτας (single-label categorisation)* ή *μη επικαλυπτόμενης κατηγοριοποίησης (non-overlapping categorisation)*, ενώ η γενική περίπτωση κατά την οποία ένα έγγραφο μπορεί να ανήκει σε καμία έως m κατηγορίες χαρακτηρίζεται περίπτωση *πολλαπλών ετικετών* (*multi-label categorisation*).

Μια ακόμα διάκριση που γίνεται είναι το αν ο πίνακας απόφασης συμπληρώνεται κατά γραμμές (κατηγοριοποίηση με άξονα τις κατηγορίες – *category-pivoted categorisation*/CPC) ή κατά στήλες (κατηγοριοποίηση με άξονα τα έγγραφα – *document-pivoted categorisation*/DPC). Αν και αυτή η διάκριση φαίνεται να είναι περισσότερο θέμα υλοποίησης και λιγότερο εννοιολογική, είναι σημαντική από την άποψη πως το σύνολο *C* των κατηγοριών και το σύνολο *D* των εγγράφων συνήθως δεν είναι και τα δύο εξ ολοκλήρου διαθέσιμα από την αρχή. Επίσης, μερικοί επαγωγικοί αλγόριθμοι κατασκευής ταξινομητών (όπως ο *k*-NN που θα συζητηθεί παρακάτω) είναι πιο κατάλληλοι για τον ένα από τους δύο τρόπους κατηγοριοποίησης. Η DPC, που είναι και η πιο συχνά χρησιμοποιούμενη προσέγγιση, ταιριάζει περισσότερο σε εφαρμογές που τα έγγραφα γίνονται διαθέσιμα διαδοχικά, μέσα σε ένα εκτενές διάστημα χρόνου και όχι μαζικά, π.χ. αν προέρχονται από αιτήσεις χρηστών για ένα έγγραφο τη φορά. Η CPC είναι αντίθετα κατάλληλη αν κατά τη διάρκεια λειτουργίας του συστήματος προστίθενται δυναμικά νέες κατηγορίες. Σε αυτή την περίπτωση, όλα τα έγγραφα που έχουν ήδη καταταχθεί στις παλιές κατηγορίες, πρέπει να εξεταστούν για το αν πρέπει να καταταχθούν και στη νέα (π.χ. [Larkey 1999]).

2.Α.ΙΙ) Εφαρμογές της αυτόματης κατηγοριοποίησης κειμένου

Η κατηγοριοποίηση κειμένου έχει ιστορία τεσσάρων τουλάχιστον δεκαετιών, κατά τη διάρκεια των οποίων έχει δώσει ένα αριθμό από διαφορετικές εφαρμογές. Ακολούθως αναφέρονται οι σημαντικότερες από αυτές:

- Αυτόματη ευρετηριοποίηση (indexing) για συστήματα ανάκτησης πληροφοριών (Information Retrieval systems – IR systems). Σε αυτά τα συστήματα, σε κάθε έγγραφο ανατίθενται μία ή περισσότερες λέξεις ή φράσεις κλειδιά (keywords ή keyphrases), οι οποίες ανήκουν σε ένα πεπερασμένο σύνολο λέξεων που καλείται *ελεγχόμενο λεξικό (controlled dictionary)* και συχνά σχηματίζει ένα ιεραρχικό θησαυρό (π.χ. ο θησαυρός της NASA για την αεροδιαστημική επιστήμη ή ο θησαυρός MeSH (Medical Subject Headings) που καλύπτει το πεδίο της ιατρικής). Στη βιβλιογραφία περιγράφονται διάφοροι αυτόματοι ταξινομητές ειδικοί για εφαρμογές ευρετηριοποίησης εγγράφων (π.χ. [Fuhr 1985], [Robertson & Harding 1984], [Tzeras & Hartmann 1993]).
- Στενά σχετιζόμενο με το παραπάνω είναι το αντικείμενο της αυτόματης δημιουργίας μεταδεδομένων (*automated metadata generation*). Πολλά μετα-δεδομένα που χαρακτηρίζουν ένα έγγραφο είναι θεματικά, δηλαδή ο ρόλος τους είναι να περιγράψουν τη σημασιολογία του εγγράφου μέσω βιβλιογραφικών κωδικών, λέξεων-κλειδιών ή φράσεων-κλειδιών. Η δημιουργία τέτοιων μεταδεδομένων μπορεί να αντιμετωπισθεί ως πρόβλημα ευρετηριοποίησης κειμένων με ελεγχόμενο λεξικό. Ένα παράδειγμα συστήματος για αυτό το σκοπό είναι το σύστημα KLARITY (http://www.topic.com.au/products/klarity.html).
- Οργάνωση εγγράφων σε κατηγορίες (*document organization*), όπως για παράδειγμα η κατηγοριοποίηση των μικρών αγγελιών που λαμβάνονται από μια εφημερίδα (π.χ. "πώληση αυτοκινήτων", "αγορά ακινήτων", κ.τ.λ.).
- Φιλτράρισμα εγγράφων (*document filtering*), το οποίο αναφέρεται στη *δυναμική* συλλογή και κατάταξη εγγράφων, τα οποία περνούν ασύγχρονα από ένα *παραγωγό πληροφορίας* σε έναν *καταναλωτή πληροφορίας*, π.χ. το φιλτράρισμα των ειδήσεων που έρχεται από ένα πρακτορείο ειδήσεων (π.χ. Reuters) σε θεματικές κατηγορίες από μία εφημερίδα. Εδώ εντάσσεται και η εφαρμογή που αποτελεί το αντικείμενο μελέτης αυτής της εργασίας, δηλαδή η on-line κατάταξη μηνυμάτων ηλεκτρονικού ταχυδρομείου σε κατηγορίες καθώς αυτά παραλαμβάνονται από τον εξυπηρέτη ταχυδρομείου (mail server) ([Sahami et al. 1998], [Drucker et al. 1999], [Hidalgo & López 2000]). Η κατασκευή συστημάτων φιλτραρίσματος της πληροφορίας μέσω τεχνικών μηχανικής μάθησης έχει μελετηθεί ευρέως (π.χ. [Hull et al. 1996], [Schapire et al. 1998], [Schütze et al. 1995]).
- Η αντιμετώπιση ζητημάτων επεξεργασίας φυσικής γλώσσας (*natural language processing* – NLP), μερικά από τα οποία είναι:
  - Η αποσαφήνιση της έννοιας των λέξεων (*word sense disambiguation*), δηλαδή η κατάλληλη αντιστοίχιση λέξεων σε έννοιες σύμφωνα με τα συμφραζόμενα (context). Η πρόκληση εδώ είναι ο σωστός χειρισμός πολύσημων και συνώνυμων λέξεων (π.χ. [Gale et al. 1993], [Hearst 1991])
  - Ο συντακτικός προσδιορισμός των λέξεων μέσα σε μια πρόταση (*part of speech tagging*).
  - Ο αυτόματος συλλαβισμός λέξεων (*hyphenation*), χρήσιμος για τη διόρθωση λαθών συλλαβισμού.

- Η δημιουργία ιεραρχικών καταλόγων ιστοσελίδων (webpages) ή/και δικτυακών τόπων (websites) για χρήση στο Διαδίκτυο, π.χ. αυτοί που έχουν ενσωματωθεί στο YAHOO! και στο INFOSEEK. Οι ιδιαιτερότητες αυτής της εφαρμογής είναι πως, αφ'ενός αποτελεί πρόβλημα κατηγοριοποίησης πολλαπλών ετικετών (multi-label), ενώ οι προηγούμενες συνήθως είναι μονής ετικέτας (single-label), και αφ'ετέρου είναι πιο κατάλληλη η χρήση CPC αντί της DPC κατηγοριοποίησης, μιας και οι κατηγορίες δημιουργούνται και καταργούνται δυναμικά. Δείτε σχετικά π.χ. τα [Mladenić 1998b], [McCallum et al. 1998].
- Τέλος, η ΚΚ έχει χρησιμοποιηθεί σε συνδυασμό με άλλες τεχνολογίες σε εφαρμογές όπως:
  - Η αναγνώριση ομιλίας ([Schapire & Singer 2000])
  - Η κατηγοριοποίηση πολυμεσικών (multimedia) εγγράφων με βάση τις λεζάντες που αναφέρονται στις εικόνες ή με βάση την πληροφορία που φέρει η ίδια η εικόνα (image processing) ([Sable & Hatzivassiloglou 1999]).
  - Η απόδοση κειμένων άγνωστης ή αμφισβητούμενης πατρότητας σε συγκεκριμένο συγγραφέα ([Forsyth 1999]).

Φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου

Είναι γεγονός πως το ηλεκτρονικό ταχυδρομείο αποτελεί σήμερα μία από τις πιο γρήγορες, οικονομικές και εύχρηστες μορφές επικοινωνίας. Τα σαφή πλεονεκτήματα που παρουσιάζει το έχουν κάνει ιδιαίτερα δημοφιλές, όχι μόνο για τους απλούς χρήστες που θέλουν να επικοινωνούν με φίλους και συναδέλφους τους, αλλά και για εταιρείες, οι οποίες βρήκαν δελεαστική την προοπτική να διαφημίζουν τα προϊόντα ή τις υπηρεσίες τους μέσω ηλεκτρονικών μηνυμάτων.

Η ύπαρξη λογισμικού μαζικής αποστολής e-mails, η αυξανόμενη διαθεσιμότητα τεράστιων λιστών από ηλεκτρονικές διευθύνσεις – οι οποίες έχουν συλλεχθεί κυρίως από ιστοσελίδες και αρχεία ομάδων συζήτησης (newsgroups) – και ο συνεχής πολλαπλασιασμός των εταιρειών που επιλέγουν να δραστηριοποιηθούν στο Διαδίκτυο έχουν διογκώσει υπερβολικά το πλήθος των διαφημιστικών e-mails, τα οποία στέλνονται "τυφλά" σε χιλιάδες υποψήφιους πελάτες ταυτόγχρονα, με ελάχιστο κόστος και κόπο. Το περιεχόμενο των μηνυμάτων αυτών ποικίλει, από διαφημίσεις τουριστικών πακέτων, μέχρι σχήματα γρήγορου πλουτισμού ("get-rich-quick") και πληροφορίες πρόσβασης σε πορνογραφικούς δικτυακούς τόπους (websites). Δεν αποτελεί έκπληξη, λοιπόν, το γεγονός πως αποκαλούνται ευρέως πλέον από τους δυσαρεστημένους χρήστες "spam" e-mails (ή junk e-mails – μηνύματα-"σκουπίδια").

Μία έρευνα το 1997 [Cranor & LaMacchia 1998] έδειξε πως περίπου το 10% των εισερχομένων e-mails σε ένα επιχειρησιακό (corporate) δίκτυο είναι spam. Κατά συνέπεια, πολλοί χρήστες του ηλεκτρονικού ταχυδρομείου αναλώνουν ένα μη αμελητέο ποσοστό του χρόνου τους προσπαθώντας να εντοπίσουν τη χρήσιμη αλληλογραφία τους. Επιπλέον, το προσβλητικό ή ακατάλληλο περιεχόμενο τους είναι ένας σημαντικός λόγος δυσανασχέτησης, ειδικά όταν σε αυτό έχουν πρόσβαση και ανήλικοι. Τέλος, σπαταλούνται πόροι όπως το εύρος ζώνης του δικτύου και ο αποθηκευτικός χώρος στον εξυπηρέτη ταχυδρομείου, ο οποίος

είναι εύκολο να γεμίσει σε ένα μεγάλο σύστημα με χιλιάδες χρήστες που λαμβάνουν συχνά αντίγραφα των ίδιων spam.

Λογισμικό που προσπαθεί να αντιμετωπίσει το πρόβλημα είναι ήδη διαθέσιμο, κυρίως στη μορφή shareware*. Πέραν από τη δυνατότητα δημιουργίας "μαύρων λιστών" από (γνωστούς) ανεπιθύμητους αποστολείς και λιστών από έμπιστους αποστολείς, αυτού του είδους το λογισμικό βασίζεται κυρίως στο χειρωνακτικό ορισμό κανόνων ταιριάσματος προτύπων, συνήθως λέξεις ή φράσεις κλειδιά, που κατά την κρίση του χρήστη μπορούν να διαχωρίσουν τα επιθυμητά από τα ανεπιθύμητα mails. Αυτή η λύση απέχει πολύ από το να είναι ικανοποιητική. Απαιτεί από τους χρήστες την ικανότητα και την εμπειρία αναγνώρισης των spam μηνυμάτων, που και αν ακόμα υπάρχει, θα πρέπει να εκφραστεί μέσω της σύνταξης σωστών κανόνων. Επιπλέον, τα χαρακτηριστικά των spam (π.χ. προϊόντα που διαφημίζονται, συχνοί όροι) μεταβάλλονται με τον καιρό, και κατά συνέπεια απαιτείται συνεχής συντήρηση και προσαρμογή των κανόνων από τους χρήστες. Αυτή η διαδικασία είναι χρονοβόρα, κουραστική και επιρρεπής σε λάθη και παραλείψεις.

Τα προβλήματα που παρουσιάζει ο χειρωνακτικός ορισμός συνόλων από κανόνες αναδεικνύουν την ανάγκη για αυτόματα προσαρμοζόμενες μεθόδους. Ένα σύστημα φιλτραρίσματος spam που χρησιμοποιεί τέτοιες μεθόδους θα πρέπει να είναι ικανό να προσαρμόζεται αυτόματα στις αλλαγές στα χαρακτηριστικά των mails. Επιπλέον, ένα σύστημα το οποίο θα εκπαιδεύεται κατ'ευθείαν από τα mails στο mailbox του χρήστη, θα δημιουργεί φίλτρα ειδικά προσαρμοσμένα στις αντιλήψεις του τελευταίου σχετικά με το ποια mails είναι επιθυμητά και ποια όχι (π.χ. μπορεί να ενδιαφέρεται για διαφημιστικά μηνύματα προϊόντων μιας συγκεκριμένης κατηγορίας). Αυτό με τη σειρά του μπορεί να οδηγήσει σε εξατομικευμένα φίλτρα μεγάλης ακρίβειας.

2.Β) Μηχανική μάθηση

Στην ενότητα αυτή σκιαγραφείται η επιστημονική περιοχή της *μηχανικής μάθησης (machine learning)*, η οποία αποτελεί πλέον την κυρίαρχη προσέγγιση στην αυτόματη κατηγοριοποίηση κειμένου, όπως και σε πλήθος άλλες εφαρμογές. Η διεξοδική περιγραφή του χώρου δεν αποτελεί στόχο της ακόλουθης παρουσίασης (για μία πολύ καλή εισαγωγή δείτε το [Mitchell 1996]). Ο στόχος εδώ είναι να δοθεί το αναγκαίο υπόβαθρο για την κατανόηση των αλγορίθμων που χρησιμοποιήθηκαν για τη διενέργεια των πειραμάτων της εργασίας. Γι' αυτό και μετά από λίγα γενικά στοιχεία, η παρουσίαση θα εστιαστεί στους συγκεκριμένους αλγορίθμους.

Όπως αναφέρθηκε και στην εισαγωγή, η μηχανική μάθηση έχει ως σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν, δηλαδή ικανών να βελτιώνουν την απόδοση τους σε κάποιους τομείς μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας. Αν και απέχουμε πάρα πολύ από τη δημιουργία μηχανών που να μαθαίνουν τόσο καλά και τόσο μεγάλη ποικιλία πραγμάτων όσο ο άνθρωπος, έχουν αναπτυχθεί αλγόριθμοι για συγκεκριμένες περιοχές

* Επισκεφθείτε, για παράδειγμα, τη διεύθυνση http:/www.tucows.com. Επίσης, σχετικές πληροφορίες μπορείτε να βρείτε στις διευθύνσεις http://www.cause.org, http://www.junkermail.org, http://spam.abuse.net και http://www.esi.uem.es/vjmgomez/spam.

μάθησης, οι οποίοι έχουν επιτρέψει την εμφάνιση εμπορικών εφαρμογών με σημαντική επιτυχία. Για προβλήματα όπως η αναγνώριση φωνής (speech recognition) και η εξόρυξη γνώσης (data mining) από μεγάλες βάσεις δεδομένων, η χρήση αλγορίθμων μηχανικής μάθησης αποτελεί πλέον ρουτίνα, ενώ έχουν σχεδιαστεί προγράμματα ικανά από το να μαθαίνουν να παίζουν τάβλι σε επίπεδο ανάλογο με των παγκόσμιων πρωταθλητών [Tesauro 1995] μέχρι να μαθαίνουν να οδηγούν αυτόνομα οχήματα σε δημόσιες λεωφόρους [Pomerlau 1989]. Επίσης, έχουν δημοσιευθεί θεωρητικά αποτελέσματα σχετικά με τις θεμελιώδεις σχέσεις μεταξύ του όγκου της εμπειρίας που είναι διαθέσιμος, του αριθμού των υπό θεώρηση υποθέσεων και του προβλεπόμενου λάθους στην επιλεχθείσα υπόθεση, ενώ έχουν αρχίσει να εμφανίζονται μοντέλα μάθησης για τον άνθρωπο και τα ζώα και να συσχετίζονται με τους αλγόριθμους που έχουν αναπτυχθεί για υπολογιστές. Μερικές σύγχρονες κατευθύνσεις της μηχανικής μάθησης δίνονται στο [Dietterich 1997].

Ένας αρκετά γενικός ορισμός που θα μπορούσε να δοθεί για τη μηχανική μάθηση δίνεται στο [Mitchell 1996]:

" Ένα πρόγραμμα υπολογιστή λέμε ότι *μαθαίνει* από την εμπειρία Ε ως προς κάποια κλάση εργασιών Τ και μέτρο απόδοσης Ρ, αν η απόδοση του σε εργασίες από το Τ, όπως μετριέται από το Ρ, βελτιώνεται μέσω της εμπειρίας Ε."

Για παράδειγμα, το πρόβλημα της αυτόματης κατηγοριοποίησης κειμένου θα μπορούσε να προσδιοριστεί σύμφωνα με τον παραπάνω ορισμό ως εξής:

- Έργο Τ: Η κατάταξη κειμένων φυσικής γλώσσας σε ένα προκαθορισμένο σύνολο θεματικών κατηγοριών.
- Μέτρο απόδοσης Ρ: Το ποσοστό των κειμένων που ταξινομήθηκαν σωστά.
- Εμπειρία Ε: Ένα σύνολο από κείμενα με γνωστή κατηγοριοποίηση.

Πολλές φορές, το πρόβλημα της βελτίωσης της απόδοσης Ρ στην εργασία Τ μπορεί να αναχθεί στο πρόβλημα της προσέγγισης μιας *συνάρτησης-στόχου* (*target function*) ή *αντικειμενικής συνάρτησης* (*object function*), γεγονός που απλοποιεί τους περαιτέρω συλλογισμούς. Σε κάποια προβλήματα η συνάρτηση-στόχος είναι προφανής, ενώ σε άλλα δεν είναι και η επιλογή της αποτελεί καίρια σχεδιαστική επιλογή.

Πεδίο ορισμού αυτής της συνάρτησης είναι ένα σύνολο οντοτήτων σε κάποια δεδομένη αναπαράσταση, η οποία αποτελεί το *χώρο στιγμιοτύπων (instance space)* του προβλήματος. Η πλέον συνηθισμένη αναπαράσταση είναι αυτή που παρέχει το *μοντέλο του διανυσματικού χώρου* (*vector space model*, [Salton & McGill,1983]). Σύμφωνα με αυτό το μοντέλο, οι οντότητες αναπαρίστανται ως διανύσματα, τα στοιχεία των οποίων αναπαριστούν τα *χαρακτηριστικά* (*features* ή *attributes*) της οντότητας που έχουν επιλεγεί ως σχετικά για το συγκεκριμένο πρόβλημα. Τα χαρακτηριστικά μπορούν να παίρνουν συμβολικές ή αριθμητικές τιμές. Για παράδειγμα, αν οι οντότητες αντιπροσωπεύουν μανιτάρια και το ζητούμενο είναι το αν αυτά είναι δηλητηριώδη, το διάνυσμα που αντιστοιχεί σε κάθε μανιτάρι είναι δυνατόν να περιλαμβάνει χαρακτηριστικά όπως την οσμή του, την προέλευσή του, το βάρος του κ.α.

Οι τιμές της συνάρτησης-στόχου μπορεί να είναι πρακτικά οτιδήποτε: Αριθμητικές ή συμβολικές, διακριτές ή συνεχείς, βαθμωτές ή διανυσματικές, κ.ο.κ. Ακόμα είναι δυνατόν να

έχουν φυσική σημασία (π.χ. θεματικές κατηγορίες στο πρόβλημα της ΚΚ) ή να μην έχουν (π.χ. ένας αριθμός που εκτιμά πόσο καλή είναι η κατάσταση σε μια σκακιέρα για κάθε παίκτη).

Ο παραπάνω ορισμός που δόθηκε για τη μηχανική μάθηση αναφέρεται στην πραγματικότητα στην περίπτωση της *μάθησης υπό επίβλεψη (supervised learning)*, όπως λέγεται, υπό την έννοια πως η διαδικασία της μάθησης μπορεί να θεωρηθεί πως επιβλέπεται από ειδικούς που γνωρίζουν την τιμή της συνάρτησης-στόχου για τα στιγμιότυπα που ανήκουν στην Ε. Δεν είναι όλα τα προβλήματα μάθησης επιβλεπόμενα – ένα παράδειγμα *μη επιβλεπόμενης (unsupervised)* μάθησης είναι αυτό της *ομαδοποίησης εγγράφων (document clustering),* κατά το οποίο το ζητούμενο είναι να ομαδοποιηθούν τα έγγραφα σε κατηγορίες, άγνωστες εκ των προτέρων. Στη μη επιτηρούμενη μάθηση, δεν παρέχεται κάποια εμπειρία Ε για να καθοδηγήσει τη μάθηση, αλλά ο στόχος είναι να αναδειχθεί η δομή οργάνωσης των δεδομένων μέσω κάποιου ή κάποιων κατάλληλα επιλεγμένων κριτηρίων "ομοιότητας".

Γενικά, το ζητούμενο στην περίπτωση της μάθησης υπό επίβλεψη είναι να κατασκευαστεί ένα *μοντέλο* (ή αλλιώς *υπόθεση*) που να αναπαριστά τη γνώση που παρέχεται μέσω της εμπειρίας Ε και το οποίο στη συνέχεια πρόκειται να χρησιμοποιηθεί για την αξιολόγηση νέων (μη παρατηρηθέντων) στιγμιοτύπων. Κατά κανόνα, οι προβλέψεις του προκύπτοντος μοντέλου (οι τιμές της συνάρτησης που προσεγγίζει τη συνάρτηση-στόχο) θα επαληθεύονται (θα ισούνται με την τιμή της συνάρτησης-στόχου) για την πλειοψηφία από τα στοιχεία που περιλαμβάνονται στην Ε, τα οποία λέγονται *στιγμιότυπα εκπαίδευσης (training instances)*. Μία θεμελιώδης υπόθεση στην οποία στηρίζονται οι περισσότεροι αλγόριθμοι και η θεωρία στη μηχανική μάθηση είναι πως η κατανομή των στιγμιότυπων εκπαίδευσης είναι αντιπροσωπευτική της γενικής κατανομής των στιγμιοτύπων στον υπό μοντελοποίηση χώρο. Οι προβλέψεις ενός μοντέλου για μελλοντικά (άγνωστα) στιγμιότυπα είναι περισσότερο αξιόπιστες αν τα στιγμιότυπα εκπαίδευσης ακολουθούν παρόμοια κατανομή με αυτή των μελλοντικών. Αν και αυτή η υπόθεση είναι αναγκαία για να εξάγουμε θεωρητικά αποτελέσματα, στην πράξη συχνά παραβιάζεται.

Σε μια πρώτη προσέγγιση ακούγεται αρκετά λογικό πως κάθε υποψήφιο προς επιλογή μοντέλο θα πρέπει να επαληθεύεται από όλα τα στιγμιότυπα εκπαίδευσης, ή όπως λέγεται, το μοντέλο θα πρέπει να είναι *συνεπές* (*consistent*). Στην πράξη, πέρα από το γεγονός πως δεν είναι βέβαιο ότι υπάρχει ακριβώς ένα τέτοιο μοντέλο, ακόμα κι αν υπάρχει και βρεθεί, δεν είναι σίγουρα κι η καλύτερη λύση. Η αιτία είναι το φαινόμενο του $overfitting$, το οποίο θα μπορούσε να αποδοθεί ως το υπερβολικό ταίριασμα με τα δεδομένα εκπαίδευσης. Μία υπόθεση h λέγεται πως *υπερταιριάζει (overfits)* με τα δεδομένα εκπαίδευσης αν υπάρχει μια άλλη υπόθεση h' τέτοια ώστε η h να έχει μικρότερο σφάλμα από την h' για τα δεδομένα εκπαίδευσης, αλλά η h' να έχει μικρότερο σφάλμα από την h για τη συνολική κατανομή των στιγμιοτύπων. Η h' δηλαδή είναι καλύτερη προσέγγιση του πραγματικού μοντέλου από την h. Οι κύριοι λόγοι εμφάνισης του overfitting είναι οι εξής:

- Ο μεγάλος αριθμός παραμέτρων του μοντέλου, ή πιο γενικά η ικανότητα του αλγορίθμου μάθησης να κατασκευάζει ιδιαίτερα πολύπλοκα μοντέλα.
- Η μη κατάλληλη επιλογή των χαρακτηριστικών αναπαράστασης.

- Ο θόρυβος των δεδομένων εκπαίδευσης, δηλαδή τα τυχαία λάθη που είναι δυνατόν να περιέχονται στα δεδομένα. Αν και θα θέλαμε να είχαμε απολύτως αξιόπιστα δεδομένα τα οποία να χρησιμοποιούσαμε για την κατασκευή του ταξινομητή, στην πράξη αυτό δεν είναι πάντα εφικτό. Για παράδειγμα, μπορεί τα δεδομένα να είναι σήματα από βιντεοκάμερες ή μικρόφωνα αλλοιωμένα από τυχαίο ηλεκτρομαγνητικό θόρυβο, ή να προέρχονται από ανακριβείς πειραματικές μετρήσεις σε μη ελεγχόμενο περιβάλλον, όπως αυτές που γίνονται στο διάστημα. Αξίζει να σημειωθεί πως η πιο κοινή πηγή θορύβου είναι ο "ανθρώπινος παράγοντας", π.χ. στην εισαγωγή των δεδομένων. Είναι επομένως λογικό πως ένας ταξινομητής προσαρμοσμένος απόλυτα ή πολύ κοντά στα (θορυβώδη) δεδομένα εκπαίδευσης, δεν αναμένεται να διατηρήσει την υψηλή του απόδοση σε νέα μη παρατηρηθέντα δεδομένα, ή όπως λέγεται δε θα έχει μεγάλη *ακρίβεια γενίκευσης (generalization accuracy)*.
- Οι τυχαίες κανονικότητες που είναι δυνατόν να εμφανιστούν, σε μικρά κυρίως σύνολα εκπαίδευσης, και οι οποίες μπορούν να οδηγήσουν στη δημιουργία ταξινομητών που έχουν κάνει λανθασμένες, στην πραγματικότητα, γενικεύσεις.

Το overfitting είναι μια σημαντική πρακτική δυσκολία για πολλούς αλγορίθμους μάθησης. Για τη μετρίασή του έχουν επινοηθεί μέθοδοι, τόσο προσαρμοσμένες σε καθέναν από αυτούς, όσο και ανεξάρτητες αλγορίθμου. Βασικός οδηγός στην αποφυγή του είναι η "αρχή" του *ξυραφιού του Occam* (*Occam's Razor)*: "μεταξύ όλων των ικανοποιητικών λύσεων, προτιμήστε την απλούστερη".

Αλγόριθμοι μηχανικής μάθησης

Μία οπτική γωνία απ' την οποία μπορεί κανείς να δει τη μηχανική μάθηση είναι αυτή της αναζήτησης, σε ένα πολύ μεγάλο χώρο δυνατών υποθέσεων, μιας υπόθεσης που ταιριάζει "αρκετά καλά" με τα δεδομένα εκπαίδευσης και την τυχόν εκ των προτέρων (a priori) γνώση. Οι διάφοροι επαγωγικοί αλγόριθμοι μάθησης (*inducers,* για συντομία ΕΑΜ) που έχουν αναπτυχθεί έως σήμερα διαφέρουν ως προς την υποκείμενη αναπαράσταση του χώρου των δυνατών υποθέσεων, και κατά συνέπεια και του τρόπου που οργανώνουν την αναζήτηση σε αυτό το χώρο. Μερικά παραδείγματα αναπαραστάσεων είναι οι γραμμικοί συνδυασμοί, οι λογικές περιγραφές (λογικοί τύποι), τα δέντρα απόφασης (decision trees), τα τεχνητά νευρωνικά δίκτυα (artificial neural networks), κ.α. Διαφορετικές αναπαραστάσεις είναι κατάλληλες για τη μάθηση διαφορετικών ειδών συναρτήσεων-στόχων. Για κάθε μια από αυτές τις αναπαραστάσεις, ο αντίστοιχος ΕΑΜ εκμεταλλεύεται τη διαφορετική υποκείμενη δομή για να οργανώσει την αναζήτηση στο χώρο των υποθέσεων.

Μία θεμελιώδης ιδιότητα που χαρακτηρίζει κάθε ΕΑΜ είναι η *επαγωγική προδιάθεση* ή *κλίση* του (*inductive bias*). Κάθε ΕΑΜ απαιτεί κάποιου είδους a priori υποθέσεις για να μπορέσει να γενικεύσει πέρα από τα παρατηρηθέντα δεδομένα. Διαφορετικά, ένας πλήρως αμερόληπτος αλγόριθμος (*bias-free learner*) ο οποίος δεν κάνει καμιά υπόθεση σχετικά με την ταυτότητα της συνάρτησης-στόχου, δεν έχει κανένα λογικό έρεισμα για να αποφασίσει την τιμή κάποιου άγνωστου στιγμιοτύπου. Με διαφορετική διατύπωση, η επαγωγική κλίση ενός ΕΑΜ L είναι ένα ελάχιστο σύνολο από υποθέσεις, οι οποίες σε συνδυασμό με ένα δοθέν

σύνολο εκπαίδευσης και ένα άγνωστο στιγμιότυπο προς κατάταξη, μπορούν να οδηγήσουν *παραγωγικά (deductively)* στην πρόβλεψη που δίνει ο L για αυτό το στιγμιότυπο. Οι υποθέσεις αυτές άλλοτε περιορίζουν το χώρο των υπό θεώρηση μοντέλων (*restriction bias* ή *language bias*), άλλοτε επιβάλλουν μια συγκεκριμένη στρατηγική αναζήτησης στο χώρο αυτό, επιβάλλοντας έτσι την προτίμηση κάποιων υποθέσεων ως προς άλλες (*search* ή *preference bias*) και άλλοτε συνδυάζουν και τα δύο. Κατά κανόνα, δεν είναι ρητά διατυπωμένες και είναι έμφυτες στον αλγόριθμο (αν και υπάρχουν και εξαιρέσεις, π.χ. η βασισμένη στις εξηγήσεις μάθηση / Explanation-Based Learning-EBL). Η αξία της έννοιας της επαγωγικής κλίσης είναι πως δίνει ένα μη διαδικαστικό τρόπο χαρακτηρισμού της πολιτικής γενίκευσης που χαρακτηρίζει έναν ΕΑΜ.

Παρακάτω παρουσιάζονται δύο θεωρίες μηχανικής μάθησης, η Μπαυζιανή (Bayesian) και η βασισμένη στα στιγμιότυπα (instance-based), μαζί με έναν αλγόριθμο για την κάθε μία. Αυτοί είναι ακριβώς οι αλγόριθμοι που χρησιμοποιήθηκαν για τη σχεδίαση του συστήματος φιλτραρίσματος των spam e-mails και για τα πειράματα που έγιναν για διάφορες δυνατές σχεδιαστικές επιλογές.

2.Β.Ι) Μπαυζιανή μάθηση

Η *Μπαυζιανή συλλογιστική (Bayesian reasoning)* παρέχει μια πιθανοτική προσέγγιση στο πρόβλημα του επαγωγικού συμπερασμού. Στηρίζεται στην υπόθεση πως οι υπό μελέτη ποσότητες ακολουθούν πιθανοτικές κατανομές και πως οι βέλτιστες αποφάσεις μπορούν να παρθούν βάσει αυτών των κατανομών και των παρατηρούμενων δεδομένων. Στα πλεονεκτήματα της συγκαταλέγεται η δυνατότητα συνδυασμού της προϋπάρχουσας γνώσης με τα παρατηρούμενα δεδομένα, η θεώρηση πιθανοτικών (μη ντετερμινιστικών) μοντέλων και η εκτίμηση της καταλληλότητας για κάθε μοντέλο, επιτρέποντας έτσι την εξέταση και εναλλακτικών μοντέλων πέραν του εκτιμώμενου βέλτιστου.

Εκτός από την αξία της ως βάση για κάθε πιθανοτική μέθοδο, η επιρροή της Μπαυζιανής συλλογιστική είναι ευρύτερη. Πολλοί αλγόριθμοι που δε χειρίζονται άμεσα πιθανότητες μπορούν να κατανοηθούν καλύτερα ως προς τις δυνατότητες και τους περιορισμούς τους αν εξετασθούν από μία Μπαυζιανή προοπτική. Για παράδειγμα, το κριτήριο της ελαχιστοποίησης του αθροίσματος των τετραγώνων των λαθών που χρησιμοποιείται συχνά από μεθόδους παλινδρόμησης (regression), μπορεί να δειχθεί με Μπαυζιανή συλλογιστική ότι υπό ορισμένες συνθήκες δίνει την πιθανότερη υπόθεση με βάση τα δεδομένα. Μέσα στο ίδιο πλαίσιο μπορεί να διατυπωθεί και η γενική *αρχή του ελαχίστου μήκους περιγραφής (minimum description length principle – MDL principle)* [Mitchell 1996].

Στη μηχανική μάθηση, συχνά μας ενδιαφέρει να βρούμε την καλύτερη υπόθεση σε ένα χώρο Η με βάση τα γνωστά δεδομένα *D*. Ένας τρόπος να καθορίσουμε τι εννοούμε λέγοντας καλύτερη είναι να απαιτήσουμε την *πιθανότερη* υπόθεση με βάση τα δεδομένα *D* και την τυχόν προηγούμενη γνώση για τις πιθανότητες των υποθέσεων στο Η. Το θεώρημα του Μπαίυζ (Bayes), το οποίο είναι ο ακρογωνιαίος λίθος της ομώνυμης συλλογιστικής, παρέχει ένα άμεσο τρόπο υπολογισμού της πιθανότητας για μια υπόθεση h. Η έκφρασή του είναι η εξής:

$$P(h \mid D) = \frac{P(D \mid h) \cdot P(h)}{P(D)}, \qquad (2.1)$$

όπου:

- $P(h \mid D)$ είναι η πιθανότητα να ισχύει η υπόθεση h με βάση τα παρατηρηθέντα δεδομένα D και καλείται *εκ των υστέρων πιθανότητα (posterior probability)* της h, γιατί εκφράζει την εμπιστοσύνη στην h αφού έχουμε δει τα δεδομένα D.
- $P(D \mid h)$ είναι η πιθανότητα να παρατηρηθούν τα δεδομένα D σε κάποιο κόσμο που η υπόθεση h ισχύει και λέγεται *πιθανοφάνεια (likelihood)* των δεδομένων D δοθείσας της h.
- $P(h)$ είναι η πιθανότητα να ισχύει η υπόθεση h πριν την παρατήρηση των δεδομένων και λέγεται *εκ των προτέρων πιθανότητα (prior probability)* της h. Εκφράζει την προηγούμενη γνώση που τυχόν έχουμε για την ισχύ της h.
- $P(D)$ είναι η πιθανότητα να παρατηρηθούν τα δεδομένα D ανεξαρτήτως της υπόθεσης που ισχύει και λέγεται εκ των προτέρων πιθανότητα των δεδομένων D.

Σε πολλές περιπτώσεις, ο αλγόριθμος μάθησης θεωρεί ένα σύνολο υποψήφιων υποθέσεων H και αναζητεί την πιο πιθανή από αυτές δοθέντων των δεδομένων εκπαίδευσης. Μια τέτοια υπόθεση h λέγεται *μέγιστη εκ των υστέρων (maximum a posteriori* – MAP) υπόθεση. Ένας ευθύς τρόπος εύρεσης των MAP υποθέσεων είναι η εφαρμογή του θεωρήματος του Bayes για κάθε υπόθεση στο H και η επιλογή των μέγιστων από αυτές, δηλαδή:

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h \mid D) = \underset{h \in H}{\operatorname{argmax}} \frac{P(D \mid h) \cdot P(h)}{P(D)} = \underset{h \in H}{\operatorname{argmax}} P(D \mid h) \cdot P(h) \qquad (2.2)$$

Στο τελευταίο βήμα, το $P(D)$ παραλήφθηκε γιατί είναι σταθερά ως προς τις υποθέσεις. Μερικές φορές δεν έχουμε καμιά εκ των προτέρων γνώση για τις υποθέσεις h και δεν έχουμε λόγο να πιστεύουμε πως είναι ανισοπίθανες. Τότε μπορούμε να θεωρήσουμε πως και ο όρος P(h) είναι σταθερός για όλες τις υποθέσεις και να τον απαλείψουμε και αυτόν από τον τύπο (2.2). Έτσι, η MAP υπόθεση θα είναι αυτή που μεγιστοποιεί την πιθανοφάνεια $P(D|\text{h})$ και η οποία λέγεται υπόθεση *μέγιστης πιθανοφάνειας (maximum likelihood – ML)* h_{ML}.

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}} P(D \mid h) \qquad (2.3)$$

Στην πράξη, περισσότερο από το ποια είναι η πιο πιθανή υπόθεση δοθέντων των δεδομένων μας ενδιαφέρει συνήθως το ποια είναι η πιο πιθανή τιμή της συνάρτησης-στόχου ενός νέου στιγμιοτύπου δοθέντων των δεδομένων. Αν και μια απλή προσέγγιση είναι να θεωρήσουμε την τιμή της MAP υπόθεσης ως πιθανότερη τιμή, υπάρχει και καλύτερη λύση. Αυτή προκύπτει αν λάβουμε υπόψη τις προβλέψεις *όλων* των υποθέσεων, ζυγισμένες κατά την εκ των υστέρων πιθανότητά τους. Έτσι, αν η συνάρτηση-στόχος παίρνει τιμές σε ένα πεπερασμένο σύνολο V, τότε η πιθανότητα $P(\text{v}_\text{j}|\text{x},D)$ πως η σωστή τιμή για το στιγμιότυπο x είναι η v_j δίνεται από τη σχέση:

$$P(\text{v}_j \mid \text{x}, D) = \sum_{h \in H} P_h(\text{v}_j \mid \text{x}) \cdot P(h \mid D), \qquad (2.4)$$

όπου $P_h(\mathrm{v}_j \mid \mathrm{x})$ είναι η πιθανότητα να έχει το στιγμιότυπο x την τιμή v_j σύμφωνα με την υπόθεση *h*. Η σχέση (2.4), όπως φαίνεται, μπορεί να εφαρμοστεί και για μη ντετερμινιστικές υποθέσεις, δηλαδή υποθέσεις h που για ένα δεδομένο στιγμιότυπο x δεν ισχύει απαραίτητα

$$P_h(\mathrm{v}_j \mid \mathrm{x}) = \begin{cases} 1, \mathrm{v}_j = \mathrm{v}_\mathrm{x} \\ 0, \mathrm{v}_j \neq \mathrm{v}_\mathrm{x} \end{cases}, \text{για κάποιο } \mathrm{v}_\mathrm{x} \in \mathrm{V} \qquad (2.5)$$

Η βέλτιστη απόφαση είναι η τιμή v_j για την οποία ο τύπος (2.4) μεγιστοποιείται:

$$\mathrm{v}_{\mathrm{opt}}(\mathrm{x}, D) = \underset{\mathrm{v}_j \in V}{\operatorname{argmax}} \sum_{h \in H} P_h(\mathrm{v}_j \mid \mathrm{x}) \cdot P(h \mid D) \qquad (2.6)$$

Ένα σύστημα που ταξινομεί τα στιγμιότυπα χρησιμοποιώντας την εξίσωση (2.6) καλείται *βέλτιστος ταξινομητής Μπαίυζ (Bayes optimal classifier)*. Καμιά άλλη μέθοδος που θεωρεί τον ίδιο χώρο υποθέσεων, την ίδια a priori γνώση και τα ίδια δεδομένα δεν μπορεί να τον ξεπεράσει κατά μέσο όρο [Mitchell 1996].

Απλοϊκός ταξινομητής Μπαίυζ (Naïve Bayes)

Δύο πρακτικά προβλήματα εμφανίζονται στη χρήση του βέλτιστου ταξινομητή Μπαίυζ. Το ένα είναι πως έχει γραμμική πολυπλοκότητα ως προς τον πληθικό αριθμό |H| του χώρου υποθέσεων, γεγονός που καθιστά την εφαρμογή του αδύνατη για απειροδιάστατους χώρους και μη αποδοτική για μεγάλους πεπερασμένους χώρους. Το άλλο είναι πως απαιτεί τη γνώση ή την εκτίμηση πάρα πολλών πιθανοτήτων: την πιθανοφάνεια *P*(*D*|h) των δεδομένων *D* και την εκ των προτέρων πιθανότητα *P*(h) για κάθε υπόθεση h. Μία Μπαιυζιανή μέθοδος που αντιμετωπίζει σε μεγάλο βαθμό αυτές τις δυσκολίες είναι ο *απλοϊκός ταξινομητής Μπαίυζ (naive Bayes classifier* – NB για συντομία. Βλ. και [Lewis 1998]).

Ο NB εφαρμόζεται σε προβλήματα μάθησης όπου τα στιγμιότυπα αναπαρίστανται μέσω του μοντέλου του διανυσματικού χώρου, τα χαρακτηριστικά παίρνουν διακριτές τιμές (αν κάποια είναι συνεχή, πρέπει να κβαντιστούν) και η συνάρτηση-στόχος παίρνει τιμές (*ετικέτες – labels*) σε ένα πεπερασμένο σύνολο V. Παρέχεται ένα σύνολο από διανύσματα εκπαίδευσης, βάσει του οποίου ο ταξινομητής πρέπει να προβλέψει την ετικέτα ενός νέου στιγμιοτύπου αναπαριστώμενου από το διάνυσμα $\langle a_1, a_2, ..., a_n \rangle$.

Η Μπαιυζιανή προσέγγιση στην κατάταξη του νέου στιγμιοτύπου είναι η ανάθεση σε αυτό της πιο πιθανής τιμής $\mathrm{v}_{\mathrm{opt}}$, δεδομένων των τιμών των χαρακτηριστικών του, $\alpha_1, \alpha_2, \ldots \alpha_n$:

$$\mathrm{v}_{\mathrm{opt}} = \underset{\mathrm{v}_j \in V}{\operatorname{argmax}} P(\mathrm{v}_j \mid \alpha_1, \alpha_2, ..., \alpha_\mathrm{n}),$$

η οποία μέσω του θεωρήματος του Μπαίυζ εκφράζεται ως:

$$\mathrm{v}_{\mathrm{opt}} = \underset{\mathrm{v}_j \in V}{\operatorname{argmax}} \frac{P(\alpha_1, \alpha_2, ..., \alpha_\mathrm{n} \mid \mathrm{v}_j) \cdot P(\mathrm{v}_j)}{P(\alpha_1, \alpha_2, ..., \alpha_\mathrm{n})} = \underset{\mathrm{v}_j \in V}{\operatorname{argmax}} P(\alpha_1, \alpha_2, ..., \alpha_\mathrm{n} \mid \mathrm{v}_j) \cdot P(\mathrm{v}_j) \quad (2.7)$$

Η εκτίμηση των πιθανοτήτων που εμφανίζονται στην εξίσωση (2.7) πρέπει να γίνει μέσω των δεδομένων εκπαίδευσης. Οι $P(\mathrm{v}_j)$ μπορούν να εκτιμηθούν εύκολα ως η συχνότητα εμφάνισης κάθε ετικέτας v_j στα δεδομένα. Το ίδιο όμως δε μπορεί να γίνει για τις $P(\alpha_1, \alpha_2, \ldots, \alpha_n \mid \mathrm{v}_j)$, δηλαδή τις πιθανότητες εμφάνισης κάθε δυνατού στιγμιοτύπου δεδομένης μιας ετικέτας,

αφού για συνηθισμένα μεγέθη συνόλων εκπαίδευσης, τα περισσότερα στιγμιότυπα δε θα έχουν εμφανιστεί, και επομένως η συχνότητα εμφάνισής τους θα είναι μηδέν, που προφανώς δεν είναι αξιόπιστη εκτίμηση της πραγματικής πιθανότητας εμφάνισης τους.

Ο απλοϊκός ταξινομητής Μπαίυζ βασίζεται στην απλουστευτική υπόθεση πως οι τιμές των χαρακτηριστικών είναι ανεξάρτητες δοθείσας της ετικέτας. Τότε, η πιθανότητα της κοινής εμφάνισης των $\alpha_1,\alpha_2,\ldots,\alpha_n$, δεδομένης μιας ετικέτας είναι το γινόμενο των πιθανοτήτων εμφάνισης για καθένα από αυτά: $P(\alpha_1,\alpha_2,...,\alpha_n \mid v_j) = \prod_{i=1}^{n} P(\alpha_i \mid v_j)$. Αντικαθιστώντας αυτή την έκφραση στην εξίσωση (2.7) έχουμε την έκφραση του NB:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \cdot \prod_{i=1}^{n} P(\alpha_i \mid v_j) \qquad (2.8)$$

Από την εξίσωση (2.8) φαίνεται πως το πλήθος των πιθανοτήτων $P(\alpha_i|v_j)$ που πρέπει να εκτιμηθούν επιπλέον των $P(v_j)$ ισούται με το πλήθος των διαφορετικών τιμών των features επί το πλήθος των ετικετών, σημαντικά μικρότερο από αυτό που θα απαιτούνταν για όλες τις $P(\alpha_1,\alpha_2,\ldots,\alpha_n|v_j)$, ακόμα κι αν οι εκτιμήσεις τους ήταν αξιόπιστες. Έτσι, ο NB στη φάση εκπαίδευσής του εκτιμά με βάση τα δεδομένα τις $P(v_j)$ και $P(\alpha_i|v_j)$, το σύνολο των οποίων αποτελούν το μοντέλο ταξινόμησης που μαθαίνει, και στη φάση εξέτασης χρησιμοποιεί την εξίσωση (2.8) για να κατατάξει κάθε νέο στιγμιότυπο. Ένα ενδιαφέρον χαρακτηριστικό του είναι πως δεν ερευνά το χώρο υποθέσεων για την εντοπισμό της καλύτερης υπόθεσης, όπως κάνουν πολλοί αλγόριθμοι μάθησης, αλλά σχηματίζει άμεσα ένα μοντέλο, απλά μετρώντας τη συχνότητα των συνδυασμών των τιμών των features και των ετικετών μέσα στο σύνολο εκπαίδευσης.

Αν και μια εύλογη εκτίμηση των πιθανοτήτων $P(\alpha_i|v_j)$ είναι το ποσοστό των στιγμιοτύπων εκπαίδευσης με ετικέτα v_j τα οποία έχουν τιμή α_i στο αντίστοιχο feature, δεν ενδείκνυται σε περιπτώσεις που η πραγματική τιμή της πιθανότητας είναι αρκετά μικρή. Η αιτία είναι ουσιαστικά η ίδια που έκανε προβληματική την εκτίμηση των πιθανοτήτων $P(\alpha_1,\alpha_2,\ldots,\alpha_n|v_j)$, αν και στην περίπτωση των τελευταίων η εμφάνιση του προβλήματος είναι ο κανόνας, ενώ εδώ η εξαίρεση. Ωστόσο, για ετικέτες με λίγα στιγμιότυπα εκπαίδευσης είναι πιθανό να μην υπάρχει κανένα από αυτά με τιμή α_i στο αντίστοιχο feature, και επομένως η $P(\alpha_i|v_j)$ να εκτιμηθεί ως μηδέν. Πέρα από την υποεκτίμηση της πραγματικής πιθανότητας που συμβαίνει, το χειρότερο είναι πως ο όρος αυτός θα κυριαρχήσει για όλα τα μελλοντικά στιγμιότυπα με τιμή α_i στο feature, καθώς η ποσότητα που υπολογίζεται στην εξίσωση (2.8) απαιτεί τον πολλαπλασιασμό των $P(\alpha_i|v_j)$ για αυτό το στιγμιότυπο, και επομένως η εκ των υστέρων πιθανότητα του για την κλάση v_j θα είναι μηδέν, ανεξαρτήτως των τιμών των άλλων features.

Μια προσέγγιση για την αποφυγή αυτής της δυσκολίας είναι η χρήση της *m-εκτίμησης (m-estimate)* της πιθανότητας, η οποία ορίζεται ως:

$$\hat{P}_m(\alpha_i \mid v_j) = \frac{\#(a_i, v_j) + m \cdot p}{\#(v_j) + m}, \qquad (2.9)$$

όπου $\#(a_i, \mathrm{v}_j)$ είναι το πλήθος των δεδομένων με τιμή α_i στο αντίστοιχο feature και ετικέτα v_j, $\#(\mathrm{v}_j)$ το πλήθος των δεδομένων με ετικέτα v_j, p είναι η εκ των προτέρων εκτίμηση της πιθανότητας που θέλουμε να προσδιορίσουμε και m είναι μια σταθερά που λέγεται *ισοδύναμο μέγεθος δείγματος (equivalent sample size)*, η οποία καθορίζει πόσο ισχυρό θεωρείται το p σε σχέση με τα παρατηρούμενα δεδομένα. Απουσία άλλης πληροφορίας, η συνήθης μέθοδος επιλογής του p είναι να υποθέσουμε πως οι τιμές του feature δοθείσας της v_j είναι ισοπίθανες – έτσι αν αυτό έχει k δυνατές τιμές, το p τίθεται $1/k$. Για $m = 0$, η m-εκτίμηση αντιστοιχεί απλά στο ποσοστό των δεδομένων με ετικέτα v_j τα οποία έχουν τιμή α_i στο feature. Ο λόγος που το m καλείται ισοδύναμο μέγεθος δείγματος είναι πως η εξίσωση (2.9) μπορεί να ερμηνευθεί ωσάν τα αρχικά δεδομένα να έχουν αυξηθεί κατά m εικονικά στιγμιότυπα με ετικέτα v_j, από τα οποία τα $m \cdot p$ να έχουν τιμή α_i στο feature.

Ο απλοϊκός ταξινομητής Μπαίυζ, παρά την αρκετά δεσμευτική υπόθεση της υπό συνθήκη ανεξαρτησίας των χαρακτηριστικών, έχει να επιδείξει αναπάντεχα μεγάλη ακρίβεια και σε εφαρμογές που η υπόθεση της ανεξαρτησίας εμφανώς παραβιάζεται. Στο [Domingos & Pazzani 1996] παρέχεται μια ενδιαφέρουσα ανάλυση για το ευτυχές αυτό φαινόμενο. Ένα ακόμα πλεονέκτημα του NB είναι η σχετική απλότητα των μοντέλων που κατασκευάζει, τα οποία μπορούν να γίνουν εύκολα κατανοητά από τον άνθρωπο, ιδιαίτερα μέσω *οπτικοποίησης (visualization)* [Becker et al. 1997]. Η κατανόηση του υποκείμενου μοντέλου αυξάνει γενικά την εμπιστοσύνη των χρηστών σε ένα σύστημα σε σχέση με τη θεώρηση του τελευταίου ως "μαύρου κουτιού" που δέχεται στην είσοδο στιγμιότυπα και επιστρέφει στην έξοδο προβλέψεις για αυτά. Παράλληλα, οι χρήστες μπορούν να απορρίψουν προβλέψεις του μοντέλου ή και ολόκληρο το μοντέλο αν κρίνουν πως αυτό βασίζεται σε ασήμαντους ή άσχετους παράγοντες ή αγνοεί άλλους περισσότερο κρίσιμους.

2.Β.ΙΙ) Μάθηση βασισμένη στα στιγμιότυπα

Οι *βασισμένες στα στιγμιότυπα (instance-based,* για συντομία IB*)* μέθοδοι μάθησης έχουν μια θεμελιώδη διαφορά από τις άλλες μεθόδους μάθησης που έχουν αναπτυχθεί: δεν κατασκευάζουν ένα γενικό ρητά διατυπωμένο μοντέλο που προσεγγίζει τη συνάρτηση-στόχο καθολικά. Το μόνο που κάνουν στη φάση της μάθησης είναι να αποθηκεύουν τα δεδομένα εκπαίδευσης, γι' αυτό είναι γνωστές και ως μέθοδοι *βασισμένες στη μνήμη (memory-based)*. Η γενίκευση πέρα από τα παρατηρηθέντα δεδομένα γίνεται κάθε φορά που εμφανίζεται ένα νέο στιγμιότυπο προς κατάταξη. Τότε, ένα σύνολο από σχετιζόμενα με αυτό γνωστά στιγμιότυπα ανακαλείται από τη μνήμη και χρησιμοποιείται για την κατάταξη του νέου στιγμιοτύπου. Έτσι, αυτό που συμβαίνει ουσιαστικά είναι να παρέχεται μια τοπική προσέγγιση στη συνάρτηση-στόχο αντί μίας καθολικής [Aha et al. 1991].

Το κύριο πλεονέκτημα των IB μεθόδων είναι πως μπορούν να προσεγγίσουν πολύ καλύτερα από άλλες μεθόδους τη συνάρτηση-στόχο αν αυτή είναι πολύπλοκη καθολικά, αλλά μπορεί να περιγραφεί ως μια συλλογή λιγότερο σύνθετων τοπικών προσεγγίσεων. Το κύριο μειονέκτημα τους είναι πως το υπολογιστικό κόστος κατά την ταξινόμηση νέων στιγμιοτύπων μπορεί να είναι πολύ υψηλό. Ο λόγος είναι πως σχεδόν όλοι οι υπολογισμοί λαμβάνουν χώρα

τότε και όχι κατά τη φάση εκπαίδευσης. Η IB μάθηση αναφέρεται και ως *οκνηρή μάθηση (lazy learning)*, ακριβώς για το λόγο ότι αναβάλει τους υπολογισμούς μέχρι την αίτηση για κατάταξη ενός νέου στιγμιοτύπου (query). Έτσι, ένα σημαντικό πρακτικό ζήτημα είναι η ανάπτυξη τεχνικών αποδοτικής ευρετηριοποίησης των στιγμιοτύπων εκπαίδευσης, για να μειωθεί ο χρόνος ανάκτησης τους κατά τη φάση κατάταξης.

Παρακάτω περιγράφεται ο αρκετά διαδεδομένος αλγόριθμος των *k κοντινότερων γειτόνων (k-NN)*, με τις διάφορες παραλλαγές του. Άλλες IB μέθοδοι περιλαμβάνουν την *τοπική παλινδρόμηση με βάρη (local weighted regression*), που αποτελεί μια γενίκευση του *k*-NN, τα *δίκτυα συναρτήσεων ακτινικής βάσης (radial basis function networks),* που σχετίζονται επίσης με τα τεχνητά νευρωνικά δίκτυα, και τη *συλλογιστική βασισμένη σε περιπτώσεις (case-based reasoning),* μια IB προσέγγιση στην οποία τα στιγμιότυπα αναπαρίστανται από πλούσιες συμβολικές περιγραφές.

Αλγόριθμος των *k* κοντινότερων γειτόνων (*k*-Nearest Neighbor)

Ο αλγόριθμος ταξινόμησης με βάση τους *k κοντινότερους γείτονες* (*k-Nearest Neighbor Algorithm* – *k*-NN) είναι η πιο βασική IB μέθοδος μάθησης. Η κεντρική ιδέα είναι πως η τιμή της συνάρτησης-στόχου για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των *k* πιο "κοντινών" του στιγμιοτύπων εκπαίδευσης, τα οποία αποτελούν τους "γείτονές" του. Τρία ζητήματα πρέπει να αποφασιστούν προκειμένου να καθοριστεί πλήρως ο αλγόριθμος:

- Ο ορισμός της απόστασης μεταξύ δύο στιγμιοτύπων, δηλαδή μιας μετρικής πάνω στο χώρο των στιγμιοτύπων (instance space), που θα εκφράζει την εγγύτητα, ή αλλιώς την "ομοιότητα" μεταξύ των στιγμιοτύπων.
- Ο τρόπος συνδυασμού των τιμών των *k* κοντινότερων γειτόνων.
- Η τιμή του *k.*

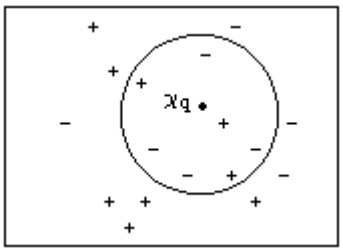
Για το πρώτο ζήτημα, υπάρχουν πολλές εναλλακτικές επιλογές. Η απόφαση εξαρτάται από τα ειδικά χαρακτηριστικά του χώρου στιγμιοτύπων του προβλήματος. Ιδιαίτερη σημασία έχει το αν στην αναπαράσταση των στιγμιοτύπων περιλαμβάνονται αριθμητικά ή συμβολικά χαρακτηριστικά. Στον "παραδοσιακό" *k*-NN αλγόριθμο, στον οποίο τα στιγμιότυπα θεωρούνται πως ανήκουν στον n-διάστατο χώρο $\Re^n$, μια μετρική που υιοθετείται συχνά είναι η γνωστή Ευκλείδια απόσταση. Πιο συγκεκριμένα, αν τα στιγμιότυπα αναπαρίστανται ως διανύσματα από χαρακτηριστικά που παίρνουν τιμές πραγματικούς αριθμούς, δηλαδή το στιγμιότυπο x αναπαρίσταται από το διάνυσμα:

$\langle a_1(x), a_2(x), ..., a_n(x) \rangle$,

όπου $a_r(x)$ δηλώνει την τιμή του r-οστού feature του x, τότε η απόσταση $d(x_i, x_j)$ μεταξύ δύο στιγμιοτύπων x_i και x_j ορίζεται ως:

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \qquad (2.10)$$

Φυσικά, κάθε συνάρτηση που πληροί τα κριτήρια μετρικής είναι δυνατόν να επιλεχθεί αντί της Ευκλείδιας.



Στην εικόνα δεξιά φαίνεται η λειτουργία του *k*-NN στην περίπτωση δυαδικής συνάρτησης-στόχου και δισδιάστατων στιγμιοτύπων. Τα "+" και τα "-" δείχνουν τα στιγμιότυπα εκπαίδευσης της κάθε κλάσης και το x_q ένα στιγμιότυπο προς κατάταξη. Φαίνεται πως ο 1-NN κατατάσσει το x_q ως "+", ενώ ο 7-NN το κατατάσσει ως "-".

Στην περίπτωση που τα χαρακτηριστικά είναι συμβολικά, η Ευκλείδια απόσταση δεν μπορεί να χρησιμοποιηθεί, αφού δεν έχει νόημα η αφαίρεση συμβολικών ποσοτήτων. Το πιο βασικό μέτρο για αυτήν την περίπτωση είναι το *μέτρο επικάλυψης (overlap metric)*, το οποίο αναφέρεται και ως *απόσταση Hamming* ή *απόσταση Manhattan*, και ορίζεται ως εξής:

$$d(x_i, x_j) \equiv \sum_{r=1}^{n} \delta(a_r(x_i), a_r(x_j)), \qquad (2.11)$$

$$\text{όπου } \delta(x, y) \equiv \begin{cases} 0, & \text{εάν } x = y \\ 1, & \text{εάν } x \neq y \end{cases}$$

Το μέτρο αυτό απλά ισούται με τον αριθμό των features στα οποία διαφέρουν τα στιγμιότυπα. Πάνω σε αυτό, μπορούν να οριστούν και άλλα πιο εξελιγμένα μέτρα.

Ένα μειονέκτημα που παρουσιάζουν τα δύο προηγούμενα παραδείγματα μετρικών είναι πως όλα τα features θεωρούνται ισοδύναμα κατά τον υπολογισμό της απόστασης. Αυτό είναι ιδιαίτερα προβληματικό αν στην πραγματικότητα δεν είναι όλα τα features σχετικά με τη συγκεκριμένη συνάρτηση-στόχο που επιδιώκεται να προσεγγιστεί, αλλά και γενικότερα, οποτεδήποτε υπάρχουν σημαντικές διαφορές μεταξύ των features ως προς την αξία τους στον προσδιορισμό της συνάρτησης. Σε μια τέτοια περίπτωση, οι παραπάνω μετρικές είναι παραπλανητικές, από την άποψη πως στιγμιότυπα που πραγματικά σχετίζονται μεταξύ τους, είναι δυνατόν να θεωρούνται απομακρυσμένα λόγω των διαφορών τους σε άσχετα ή ασήμαντα features.

Μια λύση σε αυτό το πρόβλημα είναι κάθε feature να αποτιμάται διαφορετικά στον υπολογισμό της απόστασης, ανάλογα με την αξία του. Αυτό αντιστοιχεί στο να επιμηκυνθούν οι άξονες στον Ευκλείδιο χώρο για τα σχετικά features και να συρρικνώθούν για τα λιγότερο σχετικά. Η μέθοδος αυτή λέγεται *αποτίμηση των χαρακτηριστικών (feature weighting)* και είναι χρήσιμη και σε άλλες περιπτώσεις, πέραν της χρήσης της στη διαμόρφωση της μετρικής για τον *k*-NN. Με βάση αυτήν, ο τύπος (2.11), για παράδειγμα, θα μπορούσε να γίνει:

$$d(x_i, x_j) \equiv \sum_{r=1}^{n} w_r \cdot \delta(a_r(x_i), a_r(x_j)), \qquad (2.12)$$

όπου w_r είναι το βάρος του feature a_r.

Σχετικά με το πώς υπολογίζονται τα βάρη, υπάρχουν δύο κύριες προσεγγίσεις. Η πιο απλή και άμεση προσέγγιση είναι να προσδιοριστούν τα βάρη μέσω της βελτιστοποίησης της αποτελεσματικότητας του ταξινομητή. Ένας τρόπος να γίνει αυτό είναι ο εξής:

Επιλέγονται τυχαία κάποια από τα δεδομένα εκπαίδευσης για να εκπαιδεύσουν τον αλγόριθμο μάθησης (training set) και τα υπόλοιπα χρησιμοποιούνται για έλεγχο (test set). Πάνω σε αυτά τα δεδομένα, τα βάρη $w_1, w_2, \ldots, w_n$ των features επιλέγονται έτσι ώστε να ελαχιστοποιήσουν το ποσοστό λαθών (ή κάποιο άλλο μέτρο αξιολόγησης). Στη συνέχεια, η όλη διαδικασία μπορεί να επαναληφθεί αρκετές φορές για διαφορετικούς διαμερισμούς των δεδομένων σε σύνολα εκπαίδευσης και ελέγχου, με σκοπό τη μεγαλύτερη προσέγγιση της πραγματικής κατανομής των στιγμιοτύπων. Τέλος, υπολογίζονται οι μέσοι όροι των βαρών για το σύνολο των επαναλήψεων, οι οποίοι αποτελούν και τα τελικά βάρη.

Η παραπάνω διαδικασία του επαναληπτικού χωρισμού σε σύνολα εκπαίδευσης και ελέγχου εφαρμόζεται με διάφορες παραλλαγές σε αρκετές περιπτώσεις εκτίμησης παραμέτρων και μέτρων αποτελεσματικότητας, πέραν από αυτή που αναφέρθηκε. Επίσης γενική είναι η προσέγγιση της βελτιστοποίησης παραμέτρων μέσω της χρησιμοποίησης του ίδιου του αλγορίθμου μάθησης ως μέσο για την αξιολόγηση της απόδοσης, η οποία αναφέρεται ως *προσέγγιση περιτυλίγματος (wrapper approach)*. Περισσότερα για αυτά τα θέματα θα ακολουθήσουν στην επόμενη ενότητα.

Η προσέγγιση περιτυλίγματος, αν και άμεση, δεν είναι αποδοτική, για το λόγο ότι απαιτεί κατά την αναζήτηση σε ένα μεγάλο χώρο παραμέτρων την κλήση του αλγορίθμου μάθησης τόσες φορές, όσα είναι τα βήματα της αναζήτησης. Η άλλη διαδεδομένη προσέγγιση στην εκτίμηση παραμέτρων είναι αυτή του *φίλτρου (filter approach)*. Σύμφωνα με αυτήν, η εκτίμηση γίνεται χωρίς τη χρήση του αλγορίθμου μάθησης, με τη βοήθεια των δεδομένων μόνο και μιας αριθμητικής συνάρτησης που εκτιμά τη "σημαντικότητα" της παραμέτρου. Περισσότερα και για αυτή την προσέγγιση, επίσης στην επόμενη ενότητα.

Αφού προσδιορισθούν μέσω κάποιας μετρικής οι k κοντινότεροι γείτονες ενός νέου στιγμιοτύπου x_q, οι τιμές της συνάρτησης-στόχου που έχει ο καθένας από αυτούς πρέπει να συνδυαστούν για να δώσουν την εκτιμώμενη τιμή για το νέο. Και εδώ είναι δυνατές διάφορες επιλογές. Στην περίπτωση που η συνάρτηση-στόχος παίρνει διακριτές τιμές, η πιο συνηθισμένη τακτική είναι να επιλέγεται η πιο συχνή από τις τιμές των γειτόνων, ή υπό τη μορφή τύπου:

$$\hat{f}(x_q) = \underset{v \in V}{\operatorname{argmin}} \sum_{i=1}^{k} \delta(v, f(x_i)), \qquad (2.13)$$

όπου $V = \{v_1, \ldots, v_s\}$ είναι το σύνολο των τιμών της συνάρτησης-στόχου και $f : A \to V$ η συνάρτηση-στόχος. Σε περίπτωση ισοβαθμιών επιλέγεται εκ των ισοβαθμούντων μια τιμή, είτε τυχαία ή η καθολικά πιο συχνή τιμή (για το σύνολο των στιγμιοτύπων εκπαίδευσης).

Εξίσου απλή είναι η προσέγγιση συνεχούς συνάρτησης-στόχου. Η συνήθης πρακτική είναι να υπολογίζεται ο μέσος όρος των τιμών των γειτόνων. Έτσι, η συνεχής συνάρτηση $f : \Re^n \to \Re$ προσεγγίζεται στο σημείο x_q της αίτησης από το:

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} f(x_i)}{k}. \qquad (2.14)$$

Μία βελτιωμένη παραλλαγή του *k*-NN όσον αφορά το συνδυασμό των τιμών των γειτόνων είναι η αποτίμηση της συνεισφοράς καθενός από τους *k* γείτονες με βάση την απόσταση από το προς κατάταξη στιγμιότυπο, δίνοντας μεγαλύτερο βάρος στους κοντινότερους γείτονες. Αυτή αποτελεί τη *με βάση την απόσταση (distance-weighted)* εκδοχή του αλγορίθμου. Έτσι, ο τύπος (2.13) για την περίπτωση συμβολικής συνάρτησης-στόχου γίνεται:

$$\hat{f}(x_q) \leftarrow \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} w(x_q, x_i) \cdot (1 - \delta(v, f(x_i))), \qquad (2.15)$$

όπου $w(x_q, x_i)$ μία γνησίως φθίνουσα ως προς την απόσταση $d(x_q, x_i)$ συνάρτηση και η οποία αποτελεί τη συνάρτηση αποτίμησης των γειτόνων με βάση την απόσταση. Μία κλάση συναρτήσεων που χρησιμοποιείται συχνά περιλαμβάνει συναρτήσεις της μορφής

$$w(x_q, x_i) = \frac{1}{d(x_q, x_i)^n + c},$$

όπου n και c είναι μη αρνητικές σταθερές. Αν $c = 0$ και το x_q ταυτίζεται με κάποιο στιγμιότυπο εκπαίδευσης x_i, και επομένως η απόσταση $d(x_q, x_i)$, άρα και ο παρονομαστής του κλάσματος είναι 0, ορίζουμε το $\hat{f}(x_q)$ να είναι ίσο με $f(x_i)$. Για την περίπτωση συνεχούς συνάρτησης-στόχου, ο αντίστοιχος τύπος (2.14) γίνεται

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} w(x_q, x_i) \cdot f(x_i)}{\sum_{i=1}^{k} w(x_q, x_i)}. \qquad (2.16)$$

Ο παρονομαστής χρησιμοποιείται για την κανονικοποίηση των συνεισφορών των διαφόρων βαρών.

Ο *k*-NN είναι ένας πολύ αποτελεσματικός αλγόριθμος μάθησης, τόσο για αριθμητικά όσο και για συμβολικά δεδομένα, ιδιαίτερα όταν γίνεται με αποτίμηση χαρακτηριστικών και γειτόνων. Είναι ανθεκτικός σε θορυβώδη στιγμιότυπα εκπαίδευσης, ειδικά για μεγαλύτερες τιμές του *k*, καθώς τα απομονωμένα λανθασμένα δεδομένα "απορροφώνται" κατά τον υπολογισμό του μέσου όρου. Η επαγωγική κλίση του *k*-NN είναι η υπόθεση πως η τιμή της συνάρτησης-στόχου ενός στιγμιοτύπου είναι παρόμοια με αυτή των γειτονικών του.

Ένα πρακτικό θέμα κατά την εφαρμογή του *k*-NN, όπως αναφέρθηκε και παραπάνω για τις IB μεθόδους γενικότερα, είναι η αποδοτική ευρετηριοποίηση των στιγμιοτύπων στη μνήμη. Σε μια απλή υλοποίηση, η υπολογιστική πολυπλοκότητα για την κατάταξη ενός νέου στιγμιοτύπου είναι ανάλογη του αριθμού των στιγμιοτύπων εκπαίδευσης, αφού χρειάζεται να υπολογιστεί η απόσταση του νέου με κάθε στιγμιότυπο εκπαίδευσης, για να επιλεχθούν στη συνέχεια τα *k* κοντινότερα. Κάτι τέτοιο έχει υψηλότατο κόστος για μεγάλα σύνολα δεδομένων. Για το λόγο αυτό έχουν αναπτυχθεί διάφορες μέθοδοι ευρετηριοποίησης, όπως τα *k-d δέντρα (k-d trees)* [Friedman et al. 1977], που σκοπό έχουν τον πιο γρήγορο εντοπισμό των κοντινότερων γειτόνων με κάποιο επιπλέον κόστος στη μνήμη.

2.C) Σχεδίαση συστήματος αυτόματης κατηγοριοποίησης κειμένου

Στην ενότητα αυτή δίνεται μια εικόνα της πορείας που ακολουθείται κατά τη σχεδίαση ενός συστήματος αυτόματης κατηγοριοποίησης κειμένου με χρήση τεχνικών μηχανικής μάθησης. Ο σκοπός είναι αφ' ενός μεν να καταδειχθεί η θέση της μηχανικής μάθησης στο γενικότερο πεδίο της κατηγοριοποίησης κειμένου και αφ'ετέρου να παρουσιαστούν οι σημαντικότερες επιλογές που υπάρχουν σε κάθε βήμα της διαδικασίας σχεδίασης, πέραν από αυτές που τελικά χρησιμοποιήθηκαν για τη διεξαγωγή των πειραμάτων αυτόματου φιλτραρίσματος των spam e-mails.

Η προσέγγιση της μηχανικής μάθησης στη δημιουργία ταξινομητών κειμένου σχετίζεται στενά με την περιοχή της *ανάκτησης πληροφοριών (information retrieval, ΑΠ)*. Ο λόγος είναι πως τόσο η ΚΚ, όσο και η ΑΠ είναι εργασίες διαχείρισης εγγράφων βασισμένες στο περιεχόμενο (content-based). Η ΑΠ, έχοντας προσελκύσει το ενδιαφέρον σε ερευνητικό και τεχνολογικό επίπεδο από τη δεκαετία του '40, ασχολείται με το πρόβλημα του εντοπισμού σε μια συλλογή εγγράφων, εκείνων τα οποία *σχετίζονται* με την *πληροφοριακή ανάγκη (information need)* ενός ενδιαφερόμενου, διατυπωμένης μέσω μιας ερώτησης-αίτησης (query). Κεντρική στην ΑΠ είναι η υποκειμενική έννοια της *σχετικότητας (relevance)* ενός εγγράφου με μια ερώτηση, όπως στην ΚΚ είναι η σχετικότητα ενός εγγράφου με μία ή περισσότερες θεματικές κατηγορίες. Ο στόχος ενός συστήματος ΑΠ είναι να ανακτηθούν όλα τα σχετικά έγγραφα και παράλληλα να μην ανακτηθεί κανένα άσχετο. Τεχνικές ΑΠ εφαρμόζονται σε τρία στάδια στον κύκλο ζωής ενός ταξινομητή κειμένου:

(1) Στην αναπαράσταση των εγγράφων, τόσο αυτών που χρησιμοποιούνται για τη δημιουργία (εκπαίδευση) του ταξινομητή, όσο και αυτών που εισάγονται για κατηγοριοποίηση στο σύστημα κατά το στάδιο λειτουργίας του.
(2) Συχνά στην επαγωγική κατασκευή του ταξινομητή (π.χ. ταίριασμα εγγράφου με αίτηση).
(3) Στην αξιολόγηση της αποτελεσματικότητας του ταξινομητή.

Ακολούθως παρουσιάζονται οι διαφορετικές προσεγγίσεις που έχουν χρησιμοποιηθεί σε καθένα από τα τρία αυτά στάδια. Η παρουσίαση θα είναι, για λόγους έκτασης περιληπτική. Για περισσότερες λεπτομέρειες ο αναγνώστης παραπέμπεται στη βιβλιογραφία (π.χ. [Sebastiani 1999])

2.C.I) Αναπαράσταση κειμένου

Τα έγγραφα κειμένου δεν μπορούν να χρησιμοποιηθούν απευθείας από έναν αλγόριθμο μάθησης στη φυσική τους μορφή. Γι' αυτό είναι απαραίτητη η αναπαράσταση τους σε μια κατάλληλη μορφή, η οποία παράλληλα πρέπει να διατηρεί τα ουσιώδη χαρακτηριστικά του περιεχομένου του εγγράφου. Η επιλογή της αναπαράστασης εξαρτάται από το ποιά θεωρούνται τα πληροφοριακά δομικά στοιχεία του κειμένου (πρόβλημα της *λεκτικής σημασιολογίας – lexical semantics*) και ποιοι οι γλωσσικοί κανόνες οι οποίοι διαμορφώνουν τη σημασία των συνδυασμών αυτών των δομικών στοιχείων (πρόβλημα της *συνθετικής σημασιολογίας – compositional semantics*). Το μοντέλο του διανυσματικού χώρου (vector space model) που χρησιμοποιείται στην "παραδοσιακή" IR είναι η συνήθης επιλογή, κατά την

οποία n *αποτιμημένοι όροι (weighted terms)* που εμφανίζονται στο έγγραφο σχηματίζουν το διάνυσμα αναπαράστασης του εγγράφου. Οι διαφορές ανάμεσα στις διάφορες προσεγγίσεις οφείλονται:

(1) στις διαφορετικές θεωρήσεις για το τι μπορεί να είναι ένας όρος

(2) στους διαφορετικούς τρόπους αποτίμησης των όρων

Για το πρώτο θέμα, η τυπική επιλογή είναι κάθε όρος να αντιστοιχεί σε μία λέξη, γνωστή και ως προσέγγιση του *σάκου λέξεων (bag of words)*. Μια άλλη πιο εξελιγμένη προσέγγιση είναι η θεώρηση *φράσεων* ως όρων, όπου η φράση νοείται:

- είτε *συντακτικά*, δηλαδή αποτελεί φράση σύμφωνα με το συντακτικό της γλώσσας
- είτε *στατιστικά*, δηλαδή δεν είναι απαραίτητα φράση σύμφωνα με το συντακτικό της γλώσσας, αλλά αποτελείται από ένα σύνολο ή ακολουθία λέξεων που εμφανίζονται γειτονικά με μεγάλη συχνότητα στη συλλογή των κειμένων.

Η χρήση φράσεων αντί λέξεων ως όρων έχει ως κίνητρο την αποφυγή προβλημάτων όπως η συνωνυμία και η πολυσημία των λέξεων, ιδιότητες που τις καθιστούν μη ιδανικές για την αναπαράσταση του περιεχομένου ενός κειμένου. Ωστόσο, σε πολλά πειράματα, η απλούστερη αναπαράσταση με τη χρήση λέξεων έχει αποδειχθεί αποτελεσματικότερη. Στο [Lewis 1992] δίνονται κάποια επιχειρήματα για τα αποθαρρυντικά αποτελέσματα με τη χρήση φράσεων.

Σχετικά με το θέμα της αποτίμησης των όρων, τα αποδιδόμενα σε αυτούς βάρη συνήθως κυμαίνονται μεταξύ του 0 και του 1. Μια ειδική περίπτωση είναι η χρήση δυαδικών βαρών, με το 1 να δηλώνει την παρουσία και το 0 την απουσία του όρου από το έγγραφο. Συχνά χρησιμοποιούνται τεχνικές αριθμητικής αναπαράστασης από το χώρο της ΑΠ, με κυριότερη τη συνάρτηση αποτίμησης *tfidf* (από τα αρχικά των λέξεων term frequency – inverse document frequency) [π.χ. Salton & Buckley 1988], η οποία ορίζεται ως:

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|T_r|}{\#T_r(t_k)}, \qquad (2.17)$$

όπου $\#(t_k, d_j)$ δηλώνει τον αριθμό των εμφανίσεων του όρου t_k στο έγγραφο d_j και $\#T_r(t_k)$ το πλήθος των εγγράφων στη συλλογή T_r τα οποία περιέχουν τον t_k τουλάχιστον μία φορά (γνωστό και ως *συχνότητα εγγράφων* του όρου t_k). Αυτή η συνάρτηση εκφράζει τις υποθέσεις πως

(i) όσο πιο συχνά εμφανίζεται ένας όρος σε ένα έγγραφο, τόσο πιο αντιπροσωπευτικός είναι για αυτό το έγγραφο, και

(ii) σε όσο περισσότερα έγγραφα εμφανίζεται ο όρος, τόσο λιγότερο σημαντικός είναι για την κατηγοριοποίηση των κειμένων.

Τα βάρη που προκύπτουν από την *tfidf* συχνά κανονικοποιούνται μέσω της *κανονικοποίησης συνημιτόνου (cosine normalisation)* για να ανήκουν στο διάστημα [0,1] και να σχηματίζουν διανύσματα ίσου μήκους:

$$w(t_k, d_j) = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{r} (tfidf(t_s, d_j))^2}}, \qquad (2.18)$$

όπου r η διάσταση του διανύσματος αναπαράστασης.

Στον παραπάνω τύπο, όπως και στους περισσότερους που έχουν χρησιμοποιηθεί, τα βάρη των όρων βασίζονται στη συχνότητα παρουσίας τους και μόνο στα έγγραφα, ενώ αγνοείται η σειρά των όρων, καθώς και ο συντακτικός τους ρόλος μέσα στις προτάσεις. Με άλλα λόγια, η σημασιολογία του εγγράφου περιορίζεται στη λεκτική σημασιολογία των όρων που συμμετέχουν σε αυτά, χωρίς αναφορά στη συνθετική σημασιολογία (εξαιρέσεις είναι τα συστήματα FOIL [Cohen 1995] και SLEEPING EXPERTS [Cohen & Singer 1999]).

Αν και η *tfidf* είναι η πιο συχνά εφαρμοζόμενη μέθοδος αποτίμησης, έχουν χρησιμοποιηθεί και άλλες όπως προσαρμοσμένες για δομημένα έγγραφα [Larkey & Croft 1996] και εμπειρικά υποκατάστατα της *tfidf* [Dagan et al. 1997], στην περίπτωση που η συλλογή των εγγράφων δεν είναι διαθέσιμη εξ ολοκλήρου από την αρχή, π.χ. στο προσαρμοστικό φιλτράρισμα (*adaptive filtering)*.

Συχνά τα έγγραφα υφίστανται κάποιας μορφής προεπεξεργασία πριν από την αναπαράσταση τους. Σε αυτήν περιλαμβάνεται η απομάκρυνση *λειτουργικών λέξεων (function words)*, όπως άρθρα, προθέσεις, συχνές λέξεις "ουδέτερης" σημασίας, κ.α., και η *λημματοποίηση (stemming)*. Η πρώτη γίνεται με τη βοήθεια έτοιμων λιστών με τέτοιες λέξεις *(stop-lists)* για λόγους *μείωσης της διαστασιμότητας (dimensionality reduction)*, η οποία θα αναλυθεί αμέσως παρακάτω. Η δεύτερη, που απομακρύνει τις λέξεις που έχουν κοινό λήμμα και προσθέτει αντί αυτών το λήμμα τους, γίνεται τόσο για τη μείωση της διαστασιμότητας, όσο και για την ελάττωση της στοχαστικής εξάρτησης μεταξύ των όρων που θα χρησιμοποιηθούν για την αναπαράσταση.

Όσον αφορά το ποιο κομμάτι του εγγράφου αναπαρίσταται, πέραν από τον κανόνα που είναι η αναπαράσταση όλου του εγγράφου, υπάρχουν και εξαιρέσεις, οι οποίες στηρίζονται στην εκ των προτέρων γνωστή δομή των εγγράφων. Σε αυτή την περίπτωση, μόνο τα μέρη του εγγράφου που θεωρούνται σχετικά αναπαρίστανται, όπως π.χ. ο τίτλος και η περίληψη που έχει κάθε επιστημονική δημοσίευση. Ακόμα είναι δυνατόν να δοθεί μεγαλύτερο βάρος σε κάποια μέρη του κειμένου, όπως π.χ. στον τίτλο.

Μείωση διαστασιμότητας

Ένα πρόβλημα που εμφανίζεται σε αρκετές εφαρμογές μηχανικής μάθησης είναι ο μεγάλος αριθμός χαρακτηριστικών που θεωρούνται, κατ'αρχήν τουλάχιστον, σχετικά με τη συνάρτηση-στόχο που επιδιώκεται να προσεγγιστεί. Στην ΚΚ, το πρόβλημα είναι ιδιαίτερα έντονο, καθώς όπως είναι φανερό από τα παραπάνω, το σύνολο των θεωρούμενων όρων είναι το σύνολο των λέξεων που εμφανίζονται μία τουλάχιστον φορά σε κάποιο έγγραφο, αν η αναπαράσταση θεωρεί μεμονωμένες λέξεις ως όρους. Ακόμα κι όταν χρησιμοποιείται stop-list και γίνεται λημματοποίηση, για μια τυπική συλλογή εγγράφων προκύπτουν διανύσματα αποτελούμενα από χιλιάδες features.

Δύο είναι οι βασικοί λόγοι για τους οποίους η μεγάλη *διαστασιμότητα (dimensionality)* αποτελεί πρόβλημα, και μάλιστα τόσο σημαντικό που πολλές φορές αναφέρεται ως *κατάρα της διαστασιμότητας (curse of dimensionality)*. Ένας λόγος είναι η αναπόφευκτη αύξηση της υπολογιστικής πολυπλοκότητας για την εκπαίδευση και τη λειτουργία του ταξινομητή. Ωστόσο, το επιπλέον υπολογιστικό κόστος ενδεχομένως να γινόταν αποδεκτό σε διάφορες

εφαρμογές, αν μπορούσε να εγγυηθεί την παράλληλη αύξηση της ακρίβειας του προκύπτοντος ταξινομητή. Στην πράξη, η μεγάλη διαστασιμότητα συχνά όχι μόνο δεν αυξάνει την ακρίβεια, αλλά τη μειώνει. Η αιτία είναι η ισχυρότερη εμφάνιση του φαινομένου του υπερταιριάσματος, που συζητήθηκε στην ενότητα (2.Β). Περισσότερα διαθέσιμα features σημαίνουν μεγαλύτερη πιθανότητα να αναγνωριστούν λανθασμένα πολλά από αυτά ως ουσιώδη για τον προσδιορισμό της τιμής της συνάρτησης-στόχου, ενώ στην πραγματικότητα αποτελούν συμπτωματικά καλούς προσδιοριστές, για το συγκεκριμένο διαθέσιμο σύνολο στιγμιοτύπων εκπαίδευσης. Έτσι, ο ταξινομητής που προκύπτει έχει υψηλή ακρίβεια για το σύνολο εκπαίδευσης, αλλά σημαντικά χαμηλότερη για άγνωστα στιγμιότυπα.

Δύο ακόμα λόγοι για τους οποίους πολλά features δεν συνεισφέρουν στην ακρίβεια, ειδικά στην περιοχή της ΚΚ, είναι η *συνωνυμία* και η *πολυσημία,* που θίχτηκαν και πιο πάνω. Η συνωνυμία αναφέρεται στην ύπαρξη όρων, οι οποίοι αν και είναι δυνατόν να παρέχουν καλή διαχωριστική ικανότητα από μόνοι τους, δεν παρέχουν επιπλέον πληροφορία σε συνδυασμό, λόγω μεγάλης συσχέτισης. Σε τέτοιες περιπτώσεις, θα μπορούσε να επιλεχθεί μόνο ο ένας από τους όρους, χωρίς ιδιαίτερη επίπτωση στην ακρίβεια. Η πολυσημία είναι το αντίστροφο φαινόμενο, δηλαδή ένας όρος χρησιμοποιείται για εκφράζει παραπάνω από μία έννοιες. Αν ήταν δυνατόν να αναγνωριστούν αυτές οι διαφορετικές έννοιες, ο αρχικός όρος θα μπορούσε να διασπαστεί στους αντίστοιχους όρους-έννοιες, με αποτέλεσμα να έχουμε, αντί ενός πολύσημου όρου χαμηλής διαχωριστικής ικανότητας, περισσότερους (μονόσημους) όρους υψηλής διαχωριστικής ικανότητας. Στην τελευταία περίπτωση υφίσταται βέβαια αύξηση αντί για μείωση της διαστασιμότητας, οπότε θα ήταν ίσως σωστότερος γενικά ο όρος "αναπαραμετροποίηση" ("reparameterisation") του προβλήματος.

Οι περισσότερες τεχνικές μείωσης της διαστασιμότητας που έχουν προταθεί μπορούν να εφαρμοστούν είτε *καθολικά* για όλες τις κατηγορίες, είτε *τοπικά* (δηλ. για κάθε κατηγορία χωριστά από όλες τις υπόλοιπες). Για την καθολική μείωση της διαστασιμότητας, $r' << r$ όροι επιλέγονται από τους r συνολικά όρους για την αναπαράσταση του εγγράφου, ανεξαρτήτως της κατηγορίας στην οποία είναι υποψήφιο να κατηγοριοποιηθεί. Αντίθετα για την τοπική μείωση, για κάθε κατηγορία c_i, $r_i' << r$ επιλέγονται για την υποστήριξη της διαδικασίας ελέγχου του εγγράφου για κατάταξη υπό την κατηγορία c_i. Εννοιολογικά δηλαδή, κάθε έγγραφο έχει διαφορετική αναπαράσταση για κάθε κατηγορία, αν και στην πράξη αυτό που συμβαίνει είναι πως διαφορετικά υποσύνολα της αρχικής αναπαράστασης χρησιμοποιούνται κατά την κατάταξη σε διαφορετικές κατηγορίες.

Μια δεύτερη και σημαντικότερη διάκριση αφορά τη φύση των όρων μετά τη μείωση της διαστασιμότητας. Δύο είναι οι γενικές προσεγγίσεις:

- Μέσω *επιλογής όρων (term selection)*: Οι r' επιλεγμένοι όροι είναι υποσύνολο των r αρχικών.
- Μέσω *εξαγωγής όρων (term extraction)*: Οι r' επιλεγμένοι όροι δεν είναι υποσύνολο των r αρχικών, αλλά προέρχονται από συνδυασμούς ή μετασχηματισμούς των αρχικών.

Η *επιλογή όρων*, ή αλλιώς *ελάττωση του χώρου των όρων (term space reduction* – TSR) σκοπεύει στην επιλογή ενός υποσυνόλου όρων, το οποίο οδηγεί στη μέγιστη

αποτελεσματικότητα κατάταξης. Όπως αναφέρθηκε και σε προηγούμενη ενότητα, οι δύο μέθοδοι που εφαρμόζονται για το σκοπό αυτό είναι του *περιτυλίγματος (wrapper)* και του *φίλτρου (filter)*.

Η μέθοδος του περιτυλίγματος χρησιμοποιεί τον ίδιο τον αλγορίθμο μάθησης ως μέσο για την καθοδήγηση της αναζήτησης στο χώρο των όρων. Ξεκινώντας από ένα αρχικό σύνολο όρων, δημιουργούνται διαδοχικά νέα σύνολα μέσω κατάλληλα ορισμένων τελεστών που προσθέτουν ή διαγράφουν όρους. Καθένα απ' αυτά χρησιμοποιείται για την εκπαίδευση ενός ταξινομητή, ο οποίος στη συνέχεια ελέγχεται πάνω σε ένα *σύνολο δεδομένων επικύρωσης (validation set)*. Όταν εκπληρωθεί η συνθήκη τερματισμού της αναζήτησης, το καλύτερο σύνολο ως προς την απόδοση στο validation set επιστρέφεται. Το πλεονέκτημα της μεθόδου είναι πως συντονίζει το σύνολο των όρων της αναπαράστασης στο συγκεκριμένο επαγωγικό αλγόριθμο μάθησης. Το μειονέκτημα της είναι η απαγορευτική πολυπλοκότητά της αν επιχειρηθεί εξαντλητική αναζήτηση του χώρου αναζήτησης, γεγονός που αντιμετωπίζεται μερικώς μέσω τεχνικών ευριστικής αναζήτησης ([Kohavi & John 1998]).

Η μέθοδος του φίλτρου είναι περισσότερο εφικτή υπολογιστικά και συνίσταται στην επιλογή των $r' << r$ όρων που έχουν υψηλότερη τιμή σε μια προκαθορισμένη αριθμητική συνάρτηση που εκτιμά τη βαρύτητα του όρου για το έργο της κατηγοριοποίησης. Η πιο απλή και παραδόξως αποτελεσματική συνάρτηση είναι η συχνότητα $\#T_r(t_k)$ του όρου, εκτιμώντας πως οι σημαντικότεροι όροι είναι οι συχνότεροι [Yang & Pedersen 1997]. Φυσικά, απαραίτητη είναι η απομάκρυνση των λειτουργικών λέξεων πριν τη χρήση αυτού του μέτρου. Παρόμοιας λογικής εμπειρικές μορφές επιλογής έχουν υιοθετηθεί από αρκετούς ερευνητές πριν την εφαρμογή πιο πολύπλοκων μέτρων. Για παράδειγμα, συχνά αφαιρούνται όλοι οι όροι που εμφανίζονται σε λιγότερα από k έγγραφα εκπαίδευσης, με το k να επιλέγεται μεταξύ του 2 και 5.

Αρκετές πιο πολύπλοκες συναρτήσεις έχουν χρησιμοποιηθεί, προερχόμενες από τη θεωρία πληροφορίας, και οι οποίες γενικά προσπαθούν να εκφράσουν την πεποίθηση πως οι καλύτεροι όροι για μια κατηγορία είναι αυτοί που έχουν τη μεγαλύτερη διαχωριστική ως προς την κατηγορία ικανότητα, δηλαδή οι τιμές τους κατανέμονται όσο γίνεται πιο διαφορετικά για τα θετικά και τα αρνητικά στιγμιότυπα της κατηγορίας. Στις συναρτήσεις αυτές περιλαμβάνονται το *πληροφοριακό κέρδος (information gain),* η *αμοιβαία πληροφορία (mutual information),* η χ^2 *(chi-square),* ο *λόγος πιθανοτήτων (odds ratio),* ο *βαθμός σχετικότητας (relevancy score),* ο *συντελεστής συσχέτισης (correlation coefficient),* κ.α. (για λεπτομέρειες δείτε π.χ. τα [Yang & Pedersen 1997], [Mladenić 1998a], [Ruiz & Srinivasan 1999], [Li & Jain 1998]). Μέσω αυτών είναι εφικτή η ελάττωση του χώρου των όρων μέχρι και σε επίπεδα της τάξης του 98% (δηλ. να απομακρυνθεί το 98% των λέξεων που εμφανίζεται τουλάχιστον μία φορά στη συλλογή) χωρίς αισθητή πτώση της ακρίβειας κατηγοριοποίησης, ή ακόμα και με μικρή αύξηση [Yang & Pedersen 1997]. Οι ίδιες συναρτήσεις μπορούν να χρησιμοποιηθούν και για αποτίμηση των όρων (feature weighting), μετά από κανονικοποίηση ή αυτούσιες, κατά την κατασκευή του ταξινομητή από αλγόριθμο με δυνατότητα αξιοποίησης αυτών των βαρών (π.χ. *k*-NN).

Η *εξαγωγή όρων* αποσκοπεί στη σύνθεση $r' << r$ όρων προερχόμενων από τους αρχικούς που μεγιστοποιούν την αποτελεσματικότητα του ταξινομητή. Το κίνητρο αυτής της προσέγγισης είναι η αντιμετώπιση των προαναφερθέντων προβλημάτων της συνωνυμίας και της πολυσημίας των λέξεων. Οι δύο τεχνικές που έχουν προταθεί για την εξαγωγή όρων είναι η *ομαδοποίηση όρων (term clustering)* και η *λανθάνουσα σημασιολογική ευρετηριοποίηση (latent semantic indexing)*.

Η *ομαδοποίηση όρων (term clustering)* [Lewis 1992] στοχεύει στον εντοπισμό όρων με υψηλό βαθμό συσχέτισης μεταξύ τους, έτσι ώστε να αντικατασταθούν από ένα νέο ενιαίο αντιπροσωπευτικό τους όρο. Το πώς ομαδοποιούνται οι όροι και το πώς μετατρέπεται η αρχική αναπαράσταση στη νέα με τη συμμετοχή των εντοπισμένων ομάδων εξαρτώνται από τη συγκεκριμένη μέθοδο που υιοθετείται. Μια διαφοροποίηση που υπάρχει ως προς τη μέθοδο ομαδοποίησης είναι το αν αυτή λαμβάνει υπόψη τις κατηγορίες των στιγμιοτύπων ή όχι *(supervised/unsupervised clustering)*.

Η *λανθάνουσα σημασιολογική ευρετηριοποίηση* (*latent semantic indexing* – LSI) [Deerwester et al. 1990] προέρχεται από το χώρο της ΑΠ και προσπαθεί να συλλάβει την υποκείμενη σημασιολογική δομή των δεδομένων. Συνίσταται στην εφαρμογή μιας τεχνικής της γραμμικής άλγεβρας, της *αποσύνθεσης ιδιαζουσών τιμών* (*singular value decomposition*), πάνω στον πίνακα εμφανίσεων των όρων ανά έγγραφο. Έτσι προκύπτουν νέες διαστάσεις, μαζί με τη βαρύτητα της καθεμίας, και τόσο οι αρχικοί όροι όσο και τα έγγραφα αναπαριστώνται ενιαία ως ορθοκανονικά διανύσματα στον νέο διανυσματικό χώρο. Αντίθετα με την επιλογή και την ομαδοποίηση όρων, οι νέες διαστάσεις δεν είναι ερμηνεύσιμες διαισθητικά. Από αυτές επιλέγονται τελικά οι k σημαντικότερες για την αναπαράσταση, με το k να κυμαίνεται τυπικά μεταξύ 50 και 150. Η LSI πλεονεκτεί έναντι άλλων προσεγγίσεων σε περιπτώσεις που ένας πλήθος από όρους συνεισφέρει από λίγο ο καθένας στη διαμόρφωση σημαντικής πληροφορίας συνολικά.

2.C.II) Επαγωγική κατασκευή του ταξινομητή

Η επαγωγική κατασκευή ενός ταξινομητή για την κατηγορία $c_i \in C$ μπορεί γενικά να χωριστεί σε δύο επιμέρους βήματα:

(1) Τον ορισμό μιας συνάρτησης $CSV_i : D \to [0,1]$ η οποία για ένα δεδομένο έγγραφο d_j να επιστρέφει ένα μέτρο της εκτίμησης πως αυτό πρέπει να καταταχθεί στην κατηγορία c_i (*categorisation status value*).

(2) Τον ορισμό μιας πολιτικής κατηγοριοποίησης με βάση τις τιμές της συνάρτησης. Αυτή μπορεί π.χ. να είναι ο ορισμός ενός κατωφλιού τ_i τέτοιου ώστε αν $CSV_i(d_j) \geq \tau_i$, αυτό να κατατάσσεται στην κατηγορία, διαφορετικά να μην κατατάσσεται, ή μπορεί να ορίζει πως το έγγραφο κατατάσσεται στις k κατηγορίες με τις υψηλότερες τιμές. Σε μερικές περιπτώσεις (π.χ. στα δέντρα απόφασης), το βήμα (1) παρέχει ήδη δυαδική απόφαση, δηλαδή η συνάρτηση ορίζεται ως $CSV_i : D \to \{0,1\}$, οπότε το βήμα (2) είναι περιττό.

Για την επαγωγική κατασκευή του ταξινομητή έχουν εφαρμοστεί οι περισσότεροι αλγόριθμοι από το χώρο της μηχανικής μάθησης και από την ΑΠ. Ανάμεσα στους δημοφιλέστερους περιλαμβάνονται οι Μπαιυζιανοί (ή πιθανοτικοί) ταξινομητές ([Larkey & Croft 1996]), τα δέντρα απόφασης (decision trees) ([Lewis & Ringuette 1994]), η επαγωγική μάθηση λογικών κανόνων σε κανονική διαζευκτική μορφή (disjunctive normal form learning –DNF learning) ([Cohen & Singer 1999]), τα μοντέλα παλινδρόμησης (regression models) ([Schütze et al. 1995]), οι γραμμικοί ταξινομητές (π.χ. Rocchio ([Joachims 1997])), τα νευρωνικά δίκτυα ([Ruiz & Srinivasan 1999]), οι μηχανές διανυσμάτων υποστήριξης (support vector machines) ([Joachims 1998]), οι βασισμένοι στα στιγμιότυπα ταξινομητές ([Li & Jain 1998]) και οι ομάδες ταξινομητών (classifier ensembles) ([Schapire & Singer 2000]). Η περιγραφή των μεθόδων αυτών, πλην των δύο που παρουσιάστηκαν ήδη στην προηγούμενη ενότητα (Μπαιυζιανή και βασισμένη στα στιγμιότυπα μάθηση) και μερικών στοιχείων για τις ομάδες ταξινομητών που θα δοθούν στο κεφάλαιο 5, δεν καλύπτεται στην παρούσα εργασία.

2.C.III) Αξιολόγηση του ταξινομητή

Τυπικά, μετά και από την επαγωγική κατασκευή του ταξινομητή, ένα σύστημα σε λειτουργικό στάδιο έχει ολοκληρώσει το έργο της εκπαίδευσής του και είναι έτοιμο να χρησιμοποιηθεί για την κατηγοριοποίηση νέων κειμένων. Κατά την πειραματική εξέλιξη του συστήματος, ωστόσο, μένει το πολύ σημαντικό βήμα της αξιολόγησης. Αυτό έχει τρεις κυρίως σκοπούς:

1. Την εκτίμηση της χρησιμότητας του συστήματος για τους υποψήφιους χρήστες του, με τη βοήθεια κατάλληλα ορισμένων μέτρων αποτελεσματικότητας και τεχνικών εκτίμησης των μέτρων αυτών.
2. Τη ρύθμιση διαφόρων παραμέτρων του συστήματος έτσι ώστε να βελτιστοποιηθεί η απόδοσή του ως προς τα επιλεχθέντα μέτρα αποτελεσματικότητας.
3. Τη σύγκριση με άλλα υπάρχοντα συστήματα.

Ακολούθως περιγράφονται τα σημαντικότερα ζητήματα σχετικά με την αξιολόγηση.

2.C.III.a) Μέτρα αξιολόγησης

Το πιο σημαντικό κριτήριο για την αξία ενός συστήματος κατηγοριοποίησης είναι φυσικά η αποτελεσματικότητά του (effectiveness) στο έργο της κατηγοριοποίησης, η οποία αντικατοπτρίζει την ακρίβεια των προβλέψεων. Άλλα κριτήρια, όπως π.χ. η αποδοτικότητα (efficiency) η οποία αναφέρεται στην χρονική και χωρική πολυπλοκότητα των αλγορίθμων, θεωρούνται δευτερεύουσας σημασίας. Ωστόσο είναι χρήσιμα για συγκρίσεις μεταξύ συστημάτων με παρόμοια αποτελεσματικότητα, καθώς και για εφαρμογές με μεγάλους όγκους δεδομένων ή/και πραγματικού χρόνου που απαιτούν γρήγορη απόκριση.

Τα πιο συχνά χρησιμοποιούμενα μέτρα αποτελεσματικότητας προέρχονται από την ΑΠ και είναι η *ορθότητα*[*] (*precision* – Pr) και η *ανάκληση* (*recall* – Re). Η ορθότητα ως προς μια

[*] Ο όρος *precision* μεταφράζεται ως *ορθότητα* αντί για *ακρίβεια*, διότι η τελευταία χρησιμοποιείται ως μετάφραση του όρου *accuracy*.

κατηγορία c_i ορίζεται ως η υπό συνθήκη πιθανότητα $P(a_{ij} = 1 \mid \hat{a}_{ij} = 1)$, δηλαδή η πιθανότητα ένα τυχαίο έγγραφο d_j που έχει καταταχθεί στην κατηγορία c_i να έχει καταταχθεί σωστά. Αντίστοιχα, η ανάκληση ως προς μια κατηγορία c_i ορίζεται ως η υπό συνθήκη πιθανότητα $P(\hat{a}_{ij} = 1 \mid a_{ij} = 1)$, δηλαδή η πιθανότητα να καταταχθεί στην κατηγορία c_i ένα τυχαίο έγγραφο d_j που ανήκει σε αυτήν. Με όρους λογικής, η Pr μπορεί να θεωρηθεί ως ο "βαθμός ορθότητας" ("degree of soundness") του ταξινομητή ως προς την κατηγορία και η Re ο "βαθμός πληρότητάς" του ("degree of completeness"). Ένα άλλο μέτρο που χρησιμοποιείται μερικές φορές είναι η *αστοχία* (*fallout* ή miss-rate – Fl), η οποία ορίζεται ως η υπό συνθήκη πιθανότητα $P(\hat{a}_{ij} = 1 \mid a_{ij} = 0)$, δηλαδή η πιθανότητα να καταταχθεί στην κατηγορία c_i ένα τυχαίο έγγραφο d_j που δεν ανήκει σε αυτή.

Τα παραπάνω μέτρα, αλλά και κάθε μέτρο που ορίζεται ως προς μία κατηγορία ("τοπικά"), μπορεί να οριστεί και για το σύνολο των κατηγοριών ("καθολικά"). Δύο τρόποι υπάρχουν για το σκοπό αυτό:

- *Μακροεκτίμηση (macroaveraging)*: Το καθολικό μέτρο ορίζεται ως ο μέσος όρος των τοπικών μέτρων για κάθε κατηγορία.
- *Μικροεκτίμηση (microaveraging)*: Το καθολικό μέτρο ορίζεται όπως το τοπικό, με τη διαφορά πως αναφέρεται στην τυχαία κατηγορία αντί σε μία συγκεκριμένη. Για παράδειγμα, ο καθολικός ορισμός της ανάκλησης είναι η πιθανότητα ένα τυχαίο έγγραφο d_j που ανήκει σε μια (οποιαδήποτε) κατηγορία να καταταχθεί στην κατηγορία αυτή.

Οι δύο μέθοδοι είναι δυνατόν να δώσουν αρκετά διαφορετικά αποτελέσματα, ειδικά στην περίπτωση μεγάλων ανισοκατανομών στο πλήθος των πραγματικών στιγμιοτύπων των κλάσεων (ή απλά συχνότητα των κλάσεων). Η μακροεκτίμηση δίνει το ίδιο βάρος σε κάθε κλάση ανεξαρτήτως της συχνότητάς της, ενώ η μικροεκτίμηση ευνοεί τις συχνότερες κλάσεις. Το τι πραγματικά είναι σωστό δεν είναι κοινά αποδεκτό, αλλά η πλειονότητα των ερευνητών δείχνει να προτιμάει τη δεύτερη μέθοδο.

Ούτε η ορθότητα ούτε η ανάκληση αρκούν από μόνες τους η κάθε μια για να αξιολογήσουν ένα σύστημα. Για παράδειγμα, ένας ταξινομητής που κατατάσσει όλα τα έγγραφα σε όλες τις κατηγορίες *(τετριμμένος αποδέκτης – trivial acceptor)* έχει 100% Re, αλλά πιθανότατα πολύ χαμηλή Pr. Γενικότερα, ένα κοινό φαινόμενο που παρατηρείται στην πράξη είναι πως αυξάνοντας την Re, μειώνεται η Pr, και αντίστροφα. Προκύπτει έτσι η ανάγκη για μέτρα "συνδυασμένης" αποτελεσματικότητας, με την έννοια πως πρέπει να συνυπολογίζονται τόσο η Pr όσο και η Re.

Το πιο απλό συνδυαστικό μέτρο είναι η *ακρίβεια (accuracy)*, η οποία ορίζεται ως η πιθανότητα σωστής κατάταξης ενός εγγράφου $P(\hat{a}_{ij} = a_{ij})$. Παρά την απλότητά του, το μέτρο αυτό δε χρησιμοποιείται συχνά στην ΚΚ. Οι λόγοι γίνονται φανεροί αν αναλυθεί ως $P(\hat{a}_{ij} = 0 \wedge a_{ij} = 0) + P(\hat{a}_{ij} = 1 \wedge a_{ij} = 1)$. Τυπικά, ο πρώτος όρος είναι πολύ υψηλότερος από τον δεύτερο, αφού για τις συνήθεις εφαρμογές ένα έγγραφο ανήκει σε λίγες (συχνά μία μόνο) κατηγορίες. Έτσι οι μεταβολές του δεύτερου όρου, που σχετίζονται με τις αποφάσεις για κατάταξη υπό μία κατηγορία, έχουν μικρή επίδραση στο τελικό άθροισμα, μια μη επιθυμητή ιδιότητα. Επίσης φαίνεται πως ο ταξινομητής που δεν κατατάσσει κανένα έγγραφο σε καμιά

κατηγορία (*τετριμμένος απορρίπτης – trivial rejector)* τείνει να υπερτερεί άλλων μη τετριμμένων, καθώς για την περίπτωσή του μεγιστοποιείται ο πρώτος όρος.

Πιο συχνή είναι η χρήση του λεγόμενου *σημείου ισοζυγίου (breakeven point)* ως συνδυαστικού μέτρου, δηλαδή του σημείου στο οποίο η Pr ισούται με τη Re. Αυτό προυποθέτει πως το κατώφλι για τη συνάρτηση CSV_i του ταξινομητή (βλ. ενότητα (2.C.II)) μπορεί να τεθεί αυθαίρετα. Καθώς το κατώφλι αυξάνεται από 0 σε 1, η Pr αυξάνεται συνήθως μονότονα από τη μέση *γενικότητα*[*] του συνόλου ελέγχου σε μια τιμή κοντά στο 1, ενώ η Re μειώνεται πάντα μονότονα από 1 σε 0. Έτσι, υπάρχει συνήθως μια τιμή για την οποία οι Pr και Re είναι (περίπου) ίσες.

Ένα τρίτο μέτρο είναι το F-score, το οποίο υπολογίζει την τιμή της συνάρτησης F_β, με το β να είναι μία μη αρνητική παράμετρος:

$$F_\beta \equiv \frac{\left(\beta^2+1\right)\cdot \mathrm{Pr}\cdot \mathrm{Re}}{\beta^2\cdot \mathrm{Pr}+\mathrm{Re}}.$$

Στον τύπο αυτό, το β είναι ο αποδιδόμενος σχετικός βαθμός σημαντικότητας μεταξύ των Pr και Re. Για β=0, το F_β ταυτίζεται με την Pr, ενώ για $\beta \to +\infty$, $F_\beta \to \mathrm{Re}$. Συνήθως τίθεται β=1, το οποίο αποδίδει ίση σημασία στις Pr και Re.

Αν και η αποτελεσματικότητα κυριαρχεί ως κριτήριο για την αξιολόγηση του ταξινομητή, μία άλλη προσέγγιση που προέρχεται από τη θεωρία αποφάσεων και ταιριάζει σε αρκετές εφαρμογές είναι η έννοια της *χρησιμότητας (utility)*, η οποία επεκτείνει την αποτελεσματικότητα με "οικονομικά" κριτήρια, όπως *κέρδος* και *ζημία*. Η έννοια αυτή επιτρέπει την μοντελοποίηση καταστάσεων κατά τις οποίες μια σωστή απόφαση κατάταξης δεν είναι το ίδιο "κερδοφόρα" με μια σωστή απόφαση *μη* κατάταξης, όπως επίσης είναι δυνατόν μια λανθασμένη απόφαση κατάταξης να μην είναι το ίδιο "επιζήμια" με μια λανθασμένη απόφαση *μη* κατάταξης. Ένα μέτρο βασισμένο στην χρησιμότητα, κατάλληλο για ένα σύστημα φιλτραρίσματος των spam μηνυμάτων, θα προταθεί στο επόμενο κεφάλαιο.

2.C.III.b) Εκτίμηση αποτελεσματικότητας και έλεγχος υποθέσεων

Στην περίπτωση συστημάτων ΚΚ (όπως και ΑΠ), η εκτίμηση της αποτελεσματικότητας μέσω των παραπάνω μέτρων αξιολόγησης διενεργείται κυρίως πειραματικά, παρά αναλυτικά. Ο λόγος είναι πως η κεντρική έννοια της ΚΚ, η σχετικότητα ενός εγγράφου με μια κατηγορία, είναι υποκειμενική και επομένως δε μπορεί να μοντελοποιηθεί φορμαλιστικά, έτσι ώστε να είναι δυνατή η θεωρητική απόδειξη της ορθότητας και της πληρότητας του συστήματος. Για να είναι αξιόπιστα όμως τα πειραματικά δεδομένα είναι απαραίτητη η χρήση στατιστικών τεχνικών, οι οποίες στόχο έχουν να μεγιστοποιήσουν την πιθανότητα τα παρατηρούμενα αποτελέσματα να αντιπροσωπεύουν την πραγματική συμπεριφορά του συστήματος, δηλαδή τη μέση αποτελεσματικότητά του για κάθε δυνατό σύνολο δεδομένων

[*] Γ*ενικότητα (generality)* μιας κατηγορίας c_i ως προς ένα σύνολο εγγράφων S ορίζεται το ποσοστό των εγγράφων της κατηγορίας επί του συνόλου, δηλ: $g_S(c_i) \equiv \frac{\left|\left\{d_j \in S \mid a_{ij}=1\right\}\right|}{\left|\left\{d_j \in S\right\}\right|}$

εκπαίδευσης. Επίσης είναι χρήσιμες για την επιλογή του εκτιμώμενου καλύτερου ταξινομητή από ένα σύνολο *(model selection)*. Ακολούθως περιγράφονται οι τρεις πιο συνήθεις μέθοδοι εκτίμησης ακρίβειας και επιλογής μοντέλου.

Η πιο απλή μέθοδος είναι η εκτίμηση *μέσω δείγματος ελέγχου* (*test sample estimation* ή *holdout method* – [Kohavi 1995]). Σε αυτήν, η αρχική συλλογή των δεδομένων χωρίζεται σε δύο ξένα υποσύνολα, ένα *σύνολο εκπαίδευσης (training set)* και ένα *σύνολο ελέγχου (test set)*. Ο αλγόριθμος μάθησης δέχεται το πρώτο για την εκπαίδευσή του και "εξετάζεται" στο δεύτερο, χωρίς να του αποκαλύπτονται φυσικά οι γνωστές (και θεωρούμενες σωστές) απαντήσεις. Στην περίπτωση που υπάρχουν παράμετροι του αλγορίθμου που πρέπει να συντονισθούν, έτσι ώστε να επιλεχθούν αυτές που δίνουν την υψηλότερη αποτελεσματικότητα, το σύνολο εκπαίδευσης πρέπει να χωριστεί επίπλεον σε ένα "πραγματικό" σύνολο εκπαίδευσης και ένα *σύνολο επικύρωσης (validation set)*, το οποίο χρησιμοποιείται "εσωτερικά" από τον αλγόριθμο για την εκτίμηση των βέλτιστων παραμέτρων. Μόνο όταν αυτές σταθεροποιηθούν, εξετάζεται ο αλγόριθμος πάνω στο σύνολο ελέγχου.

Μετά την εξέταση, τα αποτελέσματα μπορούν να συνοψισθούν στον *πίνακα ενδεχομένων (contingency table)*. Στο περιβάλλον της KK, ο πίνακας ενδεχομένων για κάθε κατηγορία c_i έχει την ακόλουθη μορφή:

Κατηγορία c_i		*Ανήκουν στην κατηγορία*	
		ΝΑΙ	**ΟΧΙ**
Κατατάχθηκαν στην κατηγορία	**ΝΑΙ**	TP_i	FP_i
	ΟΧΙ	FN_i	TN_i

Εδώ, TP_i είναι το πλήθος των εγγράφων που ανήκουν στην κατηγορία c_i (θετικά ως προς την c_i) και κατατάχθηκαν σωστά *(true positives)*, FP_i εκείνων που δεν ανήκουν στη c_i (αρνητικά για την c_i) και κατατάχθηκαν λάθος (*false positives)*, TN_i εκείνων που δεν ανήκουν στη c_i και κατατάχθηκαν σωστά (*true negatives)* και FN_i εκείνων που ανήκουν στη c_i και κατατάχθηκαν λάθος (*false negatives)*.

Με βάση αυτό τον πίνακα μπορούν να εκτιμηθούν όλα τα μέτρα που αναφέρθηκαν στην προηγούμενη υποενότητα. Για παράδειγμα, τα Pr και Re εκτιμώνται ως:

$$\hat{\Pr}_i = \frac{TP_i}{TP_i + FP_i}, \qquad \hat{\Re e}_i = \frac{TP_i}{TP_i + FN_i}$$

και τα μικροεκτιμώμενα αντίστοιχα τους ως:

$$\hat{\Pr}^{\mu} = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FP_i)}, \qquad \hat{\Re e}^{\mu} = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FN_i)}$$

όπου ο δείκτης μ δηλώνει πως τα μεγέθη είναι μικροεκτιμώμενα, m το πλήθος των κατηγοριών και το καρέ ("^") τονίζει πως αυτά είναι μόνο εκτιμήσεις των μέτρων βασισμένες σε ένα συγκεκριμένο χωρισμό της συλλογής σε σύνολα εκπαίδευσης και ελέγχου και όχι θεωρητικές τιμές.

Το πρόβλημα με τη μέθοδο holdout είναι πως εξαρτάται έντονα από τον τυχαίο διαχωρισμό σε δύο σύνολα. Αν τύχει το δείγμα που θα επιλεγεί για σύνολο εκπαίδευσης να είναι αρκετά αντιπροσωπευτικό του συνόλου ελέγχου, τα αποτελέσματα θα είναι

υπερεκτιμημένα, ενώ το αντίθετο θα συμβεί αν δεν είναι αντιπροσωπευτικό. Μια μέθοδος που μειώνει το πρόβλημα είναι *η διασταυρωμένη επικύρωση (cross-validation – CV)*. Σε αυτήν, η συλλογή χωρίζεται τυχαία σε k ξένα μεταξύ τους τμήματα ίσου (περίπου) μεγέθους (*folds*). Στην συνέχεια γίνονται k επαναλήψεις. Σε κάθε επανάληψη, ένα από τα τμήματα είναι το σύνολο ελέγχου και τα υπόλοιπα k-1 αποτελούν το σύνολο εκπαίδευσης. Αφού κληθεί ο αλγόριθμος μάθησης για κάθε τμήμα, η εκτιμώμενη τιμή του μέτρου είναι ο μέσος όρος του για όλα τα τμήματα. Με τον τρόπο αυτό, ακόμα κι αν σε κάποια τμήματα γίνεται υπερεκτίμηση, για μεγάλο k αναμένεται ισόποση σχεδόν υποεκτίμηση, με αποτέλεσμα ο μέσος όρος να είναι πιο κοντά στην πραγματική τιμή. Ειδική περίπτωση CV είναι η *leave-one-out* μέθοδος, για την οποία τα σύνολα ελέγχου αποτελούνται από ένα μόνο στιγμιότυπο. Μια άλλη παραλλαγή είναι η *στρωματοποιημένη (stratified CV)*, στην οποία τα τμήματα δε δημιουργούνται τελείως τυχαία, αλλά διατηρούν την ίδια σχεδόν κατανομή στιγμιοτύπων για κάθε κλάση με αυτή της αρχικής συλλογής.

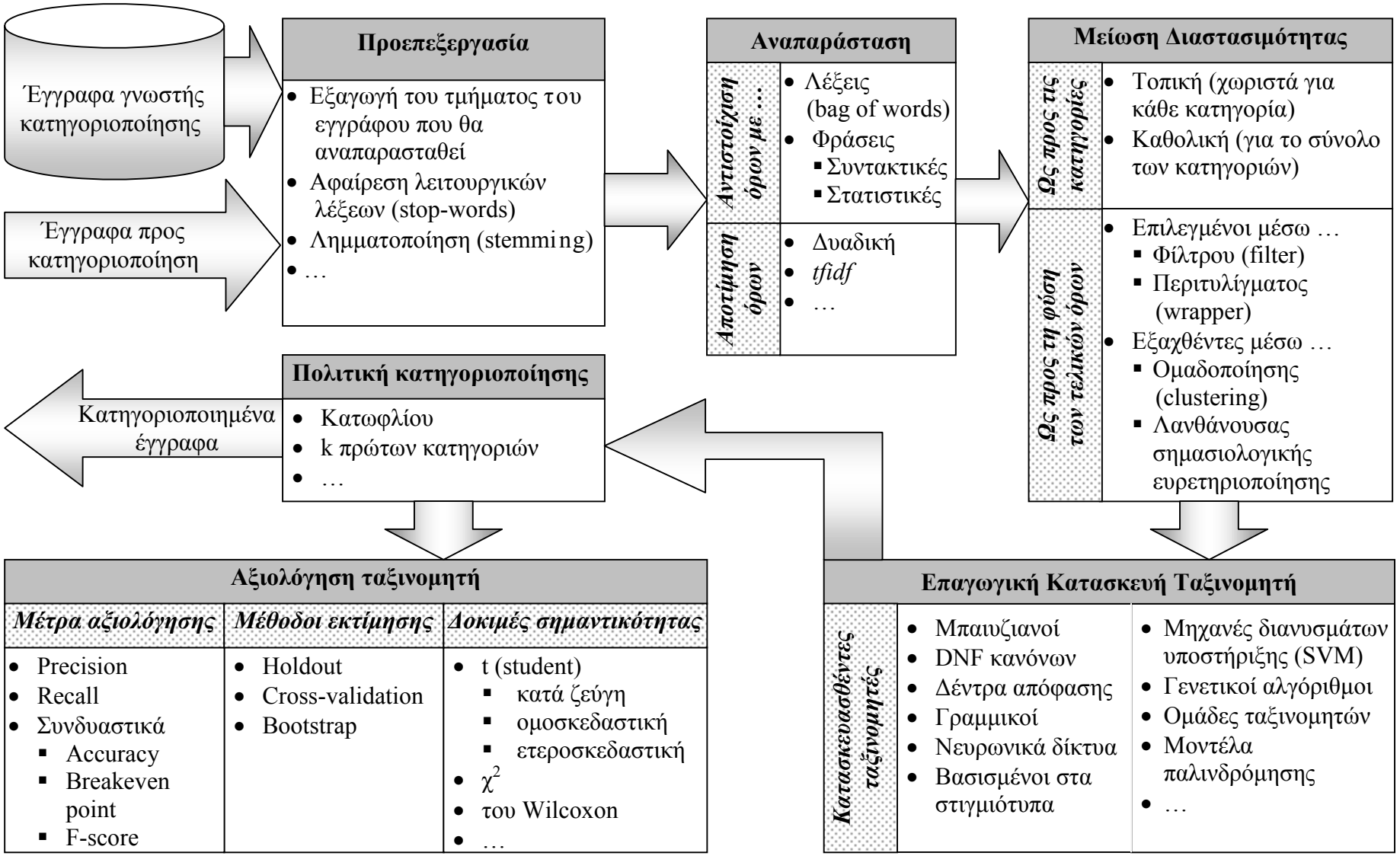
Μια τρίτη μέθοδος είναι η bootstrap ([Efron & Tibshirani, 1993]). Σε αυτήν, αν η συλλογή αποτελείται από m στιγμιότυπα, κατασκευάζονται b *bootstrap δείγματα (bootstrap samples),* επιλέγοντας ομοιόμορφα τυχαία για το καθένα *n* στιγμιότυπα *με επανάληψη*. Εφόσον τα δεδομένα δειγματολειπτούνται με επανάληψη, η πιθανότητα να μην περιέχεται ένα στιγμιότυπο στο δείγμα είναι $(1-1/n)^m \approx e^{-1} \approx 0.368$ για μεγάλο *n*. Επομένως, κάθε bootstrap δείγμα περιέχει περίπου 63.2% διαφορετικά στιγμιότυπα. Κάθε τέτοιο δείγμα αποτελεί το σύνολο εκπαίδευσης και όσα στιγμιότυπα δεν περιέχονται σε αυτό αποτελούν το σύνολο ελέγχου. Το εκτιμώμενο μέτρο είναι ο μέσος όρος του μέτρου για τις b επαναλήψεις. Το πλεονέκτημα που έχει αυτή η μέθοδος έναντι της CV είναι πως το b μπορεί να είναι οσοδήποτε μεγάλο, παρέχοντας καλύτερη εκτίμηση, ενώ στη CV ο μέγιστος αριθμός τμημάτων είναι *n*. Από την άλλη, η CV έχει το πλεονέκτημα πως τα σύνολα ελέγχου είναι ανεξάρτητα μεταξύ τους (αφού δεν περιέχουν κοινά στιγμιότυπα), με αποτέλεσμα οι k μετρήσεις να είναι πιο ανεξάρτητες σε σχέση με τη bootstrap, όπου τα σύνολα ελέγχου επικαλύπτονται μερικώς.

Τέλος, μια σημαντική πτυχή της αξιολόγησης είναι η σύγκριση διαφορετικών μεθόδων ως προς την αποτελεσματικότητά τους. Με όρους στατιστικής, το πρόβλημα αποτελεί μία περίπτωση *ελέγχου υποθέσεων* και διατυπώνεται ως το κατά πόσο μια παρατηρούμενη διαφορά στην αποτελεσματικότητα μεταξύ δύο μεθόδων αντικατοπτρίζει την πραγματική υπεροχή της μιας από την άλλη ή είναι τυχαία, ή όπως συνήθως λέγεται, το κατά πόσο η διαφορά είναι *στατιστικά σημαντική (statistically significant)*. Συνήθως αυτό αποφασίζεται από την πιθανότητα να παρατηρηθεί τόση διαφορά υποθέτοντας πως οι μέθοδοι δε διαφέρουν πραγματικά *(μηδενική υπόθεση –null hypothesis)*. Η διαφορά θεωρείται σημαντική (οπότε *απορρίπτουμε τη μηδενική υπόθεση*) αν και μόνο αν αυτή η πιθανότητα είναι κάτω από ένα όριο p, η επιλογή του οποίου γίνεται εμπειρικά, με συνήθη τιμή το 0.05. Η εκτίμηση της πιθανότητας γίνεται μέσω κάποιου στατιστικού ελέγχου, όπως δοκιμή t *(student test)* – κατά ζεύγη *(matched-pairs)*, ομοσκεδαστική ή ετεροσκεδαστική – δοκιμή χ^2, του Wilcoxon, κ.α. Αν και παρακάτω θα γίνει μια σύντομη αναφορά στο συγκεκριμένο test που υιοθετήθηκε για

τον έλεγχο της σημαντικότητας μεταξύ διαφόρων μεθόδων, η αναλυτική εξήγηση των προαναφερθέντων δοκιμών μπορεί να βρεθεί σε οποιοδήποτε εγχειρίδιο στατιστικής (π.χ. [Box et al. 1978]) και ελέγχου υποθέσεων και είναι πέραν του σκοπού αυτής της παρουσίασης.

2.C.IV) Σύνοψη της σχεδίασης

Στο ακόλουθο διάγραμμα συνοψίζονται τα τυπικά βήματα της διαδικασίας σχεδίασης ενός συστήματος αυτόματης κατηγοριοποίησης κειμένου, με τις κυριότερες επιλογές που υπάρχουν σε κάθε βήμα, όπως αυτές παρουσιάστηκαν παραπάνω. Σημειώνεται πως το βήμα της αξιολόγησης εφαρμόζεται μόνο κατά την πειραματική εξέλιξη του συστήματος και τη φάση ρύθμισης παραμέτρων. Επίσης, αν και εννοιολογικά είναι το τελευταίο βήμα, τυπικά ξεκινάει πριν από την επαγωγική δημιουργία του ταξινομητή, με τη επιλογή των (ενός ή παραπάνω) συνόλων εκπαίδευσης, ελέγχου και, ενδεχομένως, επικύρωσης.



Σχεδίαση Συστήματος Αυτόματης Κατηγοριοποίησης Κειμένων

3) ΠΕΡΙΒΑΛΛΟΝ ΔΙΕΞΑΓΩΓΗΣ ΤΩΝ ΠΕΙΡΑΜΑΤΩΝ

Σε αυτό το κεφάλαιο δίνεται το πλαίσιο στο οποίο τοποθετούνται τα πειράματα που διεξήχθησαν. Αναφέρεται η συγκεκριμένη συλλογή ηλεκτρονικών μηνυμάτων που χρησιμοποιήθηκε καθώς και ο τρόπος προεπεξεργασίας και αναπαράστασης των τελευταίων πριν τη χρήση τους από τους αλγορίθμους μάθησης. Θίγεται η έννοια του κόστους για μια λανθασμένη κατάταξη και των επακολούθων που αυτή έχει, αφ'ενός στην πολιτική κατηγοριοποίησης και αφ'ετέρου στον ορισμό κατάλληλων μέτρων αξιολόγησης της αποτελεσματικότητας ενός φίλτρου. Τέλος, παρουσιάζονται τα αποτελέσματα προηγούμενων πειραμάτων πάνω στην ίδια συλλογή με βάση τον απλοϊκό ταξινομητή Μπαίυζ.

3.Α) Συλλογή μηνυμάτων

Η έρευνα στην κατηγοριοποίηση κειμένου έχει ωφεληθεί από την ύπαρξη δημόσια διαθέσιμων χειρωνακτικά κατηγοριοποιημένων συλλογών κειμένων (όπως αυτή του ειδησεογραφικού πρακτορείου Reuters [Yang 1999]), οι οποίες χρησιμεύουν ως συλλογές-ορόσημα (benchmarks). Η δημιουργία τέτοιων συλλογών για την περιοχή του φιλτραρίσματος ηλεκτρονικών μηνυμάτων δεν είναι εξίσου εύκολη, αφού η δημοσίευση της αλληλογραφίας απλών χρηστών θα παραβίαζε το προσωπικό τους απόρρητο. Μία προσέγγιση σε αυτό το ζήτημα είναι η ανάμειξη spam μηνυμάτων με αρχειοθετημένα μηνύματα που έχουν εξαχθεί από δημόσιες λίστες αλληλογραφίας (mailing lists). Σε αυτή την προσέγγιση οφείλει τη δημιουργία της η συλλογή πάνω στην οποία πραγματοποιήθηκαν όλα τα πειράματα στα πλαίσια της εργασίας.

Η συγκεκριμένη συλλογή αποτελείται κατά ένα μέρος από spam μηνύματα και κατά το υπόλοιπο από μηνύματα που ελήφθησαν από τη Linguist list[*], μία ελεγχόμενη (κι επομένως χωρίς spam) λίστα πάνω στην επιστήμη και στο επάγγελμα του γλωσσολόγου. Το συνολικό σώμα των μηνυμάτων, που ονομάστηκε *Ling-Spam*, είναι δημόσια διαθέσιμο ως αναφορά για άλλους ερευνητές[†]. Η συλλογή αποτελείται από 2893 μηνύματα. Από αυτά:

- 2412 προέρχονται από τη λίστα Linguist.
- 481 είναι spam μηνύματα. Επισυνάψεις (attachments), ετικέτες HTML και διπλότυπα μηνύματα που ελήφθησαν την ίδια μέρα απομακρύνθηκαν.

Το ποσοστό των spam είναι έτσι το 16.6% περίπου του συνολικού σώματος, ένα νούμερο κοντά στους συνήθεις ρυθμούς λήψης spam σύμφωνα με τους [Sahami et al. 1998] και [Cranor & LaMacchia 1998].

Αν και τα μηνύματα της λίστας Linguist καλύπτουν μικρότερο εύρος θεμάτων σε σχέση με αυτά που λαμβάνει ένας τυπικός χρήστης, είναι λιγότερο τυποποιημένα απ' όσο θα ανέμενε κάποιος (π.χ. περιέχουν ανακοινώσεις για νέο λογισμικό, αγγελίες εργασίας, ακόμα και υψηλών τόνων απαντήσεις). Συνεπώς, είναι δυνατόν να προκύψουν προκαταρκτικά

[*] Το αρχείο της Linguist list υπάρχει στη διεύθυνση http://listserv.linguistlist.org/archives/linguist.html.
[†] Η συλλογή Ling-Spam είναι διαθέσιμη στη διεύθυνση
http://www.iit.demokritos.gr/~ionandr/lingspam_public.tar.gz

συμπεράσματα πάνω στο φιλτράρισμα spam μηνυμάτων με τη συλλογή Ling-Spam, τουλάχιστον μέχρι την εμφάνιση κάποιας καλύτερης συλλογής που θα είναι πιο κοντά στα πραγματικά μηνύματα των χρηστών, χωρίς όμως να παραβιάζει το προσωπικό τους απόρρητο[*]. Έστω κι έτσι, τα συμπεράσματα των πειραμάτων με βάση το Ling-Spam είναι άμεσα αξιοποιήσιμα για φίλτρα που προορίζονται για ανοιχτές, μη ελεγχόμενες, λίστες και ομάδες ενδιαφέροντος (newsgroups).

3.Β) Προεπεξεργασία και αναπαράσταση μηνυμάτων

Αρκετές από τις επιλογές που υπάρχουν ως προς την αναπαράσταση των κειμένων, την προεπεξεργασία τους και τη μείωση της διαστασιμότητας, οι οποίες αναφέρθηκαν στην ενότητα (2.C), βρήκαν εφαρμογή και στο πρόβλημα του φιλτραρίσματος των ηλεκτρονικών μηνυμάτων. Οι συγκεκριμένες επιλογές που έγιναν είναι:

- **Αναπαράσταση**
  - Τα μηνύματα αναπαρίστανται ως διανύσματα $\bar{x} = \langle x_1, x_2, x_3, \ldots, x_n \rangle$, όπου $x_1, \ldots, x_n$ είναι οι τιμές των χαρακτηριστικών $X_1, \ldots, X_n$, σύμφωνα με το μοντέλο του διανυσματικού χώρου.
  - Τα χαρακτηριστικά είναι δυαδικά, δηλαδή $X_i = 1$ αν ο όρος που αντιστοιχεί στο χαρακτηριστικό X_i είναι παρών στο μήνυμα, διαφορετικά $X_i = 0$.
  - Κάθε όρος X_i αντιστοιχεί σε μία λέξη, δηλαδή δείχνει αν μία συγκεκριμένη λέξη εμφανίζεται στο μήνυμα ή όχι.

- **Προεπεξεργασία**
  - Από κάθε μήνυμα, αναπαραστάθηκε μόνο το θέμα και το σώμα του, ενώ αφαιρέθηκαν τα υπόλοιπα πεδία, όπως η διεύθυνση του αποστολέα, η ημερομηνία και ώρα άφιξης του γράμματος, κ.τ.λ.. Ο στόχος ήταν να βασιστεί η εκπαίδευση των ταξινομητών στο περιεχόμενο και μόνο των μηνυμάτων και όχι σε "εξωγενή" πληροφορία.
  - Το Ling-Spam λημματοποιήθηκε[†], αντικαταστάθηκε δηλαδή κάθε λέξη με το λήμμα της (π.χ. το "earning" γίνεται "earn").

- **Μείωση διαστασιμότητας**
  - Εφαρμόστηκε κατωφλίωση συχνότητας των εγγράφων (*document frequency thresholding*) με κατώφλι 4. Όροι, δηλαδή, που αντιστοιχούν σε λέξεις εμφανιζόμενες σε λιγότερα από 4 μηνύματα αφαιρέθηκαν. Οι διαφορετικοί όροι στα σύνολα εκπαίδευσης μειώθηκαν έτσι, από περίπου 60.000 σε περίπου 15.000.

---

[*] Η συλλογή PU1 που έχει κατασκευαστεί [Androutsopoulos et al. 2000b], και η οποία περιέχει πραγματικά μηνύματα χρηστών κρυπτογραφημένα κατάλληλα, δεν ήταν διαθέσιμη από την αρχή της εργασίας.

[†] Χρησιμοποιήθηκε ο λημματοποιητής *morph*, ο οποίος περιέχεται στο σύστημα GATE. Περισσότερες λεπτομέρειες στη διεύθυνση http://www.dcs.shef.ac.uk/research/groups/nlp/gate.

- Για τους όρους που απέμειναν, εφαρμόστηκε μείωση της διαστασιμότητας του προβλήματος *καθολικά***,** δηλαδή κάθε μήνυμα είχε κοινή αναπαράσταση και για τις δύο κατηγορίες. Η μείωση έγινε μέσω επιλογής όρων με την προσέγγιση του φίλτρου, όπου ως συνάρτηση φίλτρου χρησιμοποιήθηκε το *πληροφοριακό κέρδος (information gain),* ή αλλιώς *αναμενόμενη αμοιβαία πιθανότητα (expected mutual information)* ([Lewis 1992]). Αυτή ορίζεται ως:

$$IG(X,C) = \sum_{x \in \{0,1\},\, c \in \{spam,\, legit\}} P(X=x, C=c) \cdot \log \frac{P(X=x, C=c)}{P(X=x) \cdot P(C=c)},$$

όπου X είναι ένα υποψήφιο feature και C η μεταβλητή που δηλώνει την κατηγορία. Οι πιθανότητες που εμφανίζονται στον τύπο εκτιμώνται από το σώμα των μηνυμάτων εκπαίδευσης μέσω *m-εκτίμησης* με p=0.5 και m=1 (βλ. ενότητα (2.B.I)). Τα features με τις N υψηλότερες τιμές *IG* επιλέγονται για την αναπαράσταση, με το N να παίρνει τιμές στα πειράματα από 50 έως 700 με βήμα 50.

3.C) Αξιολόγηση με βάση το κόστος

Μια ιδιαιτερότητα που παρουσιάζει η εφαρμογή του φιλτραρίσματος ηλεκτρονικών μηνυμάτων σε σχέση με πολλές εφαρμογές κατηγοριοποίησης κειμένου είναι η αναγκαιότητα της έννοιας του κόστους για μια λανθασμένη πρόβλεψη. Στις συνήθεις εφαρμογές, οι προβλέψεις ενός ταξινομητή μπορούν να διαχωριστούν απλά σε δύο σύνολα: στις σωστές και στις λανθασμένες. Δεν υπάρχει διαφοροποίηση μεταξύ δύο σωστών προβλέψεων, ούτε μεταξύ δυο λανθασμένων. Αντίθετα, οι προβλέψεις ενός συστήματος φιλτραρίσματος e-mails είναι ανάγκη να διαχωριστούν περαιτέρω. Ο λόγος είναι πως η κατάταξη ενός θεμιτού μηνύματος ως spam θεωρείται, συνήθως, αρκετά πιο σοβαρό λάθος από την κατάταξη ενός spam ως θεμιτού. Αν συμβολιστεί η πρώτη περίπτωση ως $L \to S$ και η δεύτερη ως $S \to L$, υποθέτουμε πως η $L \to S$ κοστίζει όσο λ $S \to L$.

Η εισαγωγή της έννοιας του κόστους έχει σημαντική επίδραση, κατά πρώτο λόγο στην πολιτική κατηγοριοποίησης του συστήματος (βλ. ενότητα (2.C.II)). Έστω $W_S(d_j)$ ο βαθμός βεβαιότητας ενός ταξινομητή πως το μήνυμα d_j είναι spam και $W_L(d_j)$ ο αντίστοιχος βαθμός πως το μήνυμα είναι θεμιτό. Αν το κόστος κατηγοριοποίησης ήταν το ίδιο για τα $L \to S$ και $S \to L$ (δηλ. λ=1), τότε η προφανής πολιτική κατηγοριοποίησης (αν ο ταξινομητής είναι αμερόληπτος) θα ήταν να κατατάσσεται ένα μήνυμα ως spam αν και μόνο αν ισχύει $W_S(d_j) > W_L(d_j)$ (σε περίπτωση ισότητας των βαρών, το μήνυμα κατατάσσεται ως θεμιτό, "λόγω αμφιβολιών"). Αν όμως η αποτυχία $L \to S$ κοστίζει όσο λ>1 αποτυχίες $S \to L$, το σύστημα θα πρέπει να είναι πιο επιφυλακτικό στο να χαρακτηρίσει ένα μήνυμα ως spam, και μάλιστα τόσο πιο επιφυλακτικό όσο μεγαλύτερη είναι η τιμή του λ. Πιο φορμαλιστικά, η πολιτική κατηγοριοποίησης σε αυτή την περίπτωση είναι πως ένα μήνυμα κατατάσσεται ως spam αν και μόνο αν ισχύει:

$$\frac{W_S(d_j)}{W_L(d_j)} > f(\lambda), \quad \lambda > 0, \qquad (3.1)$$

όπου $f(\lambda)$ είναι μία γνησίως αύξουσα συνάρτηση του λ με:

$$\lim_{\lambda\to 0} f(\lambda) \to 0\,, \quad f(1)=1\,, \quad \lim_{\lambda\to+\infty} f(\lambda) \to +\infty\,.$$

Το ποια είναι η βέλτιστη συνάρτηση $f(\lambda)$, δηλαδή ποια μεγιστοποιεί ένα κατάλληλο μέτρο αξιολόγησης βασισμένο στο κόστος (όπως αυτό που θα προταθεί στη συνέχεια), είναι εν γένει δύσκολο πρόβλημα, αφού εξαρτάται τόσο από το μέτρο καθεαυτό, όσο και από τον επιλεγμένο αλγόριθμο μάθησης και το σύνολο εκπαίδευσης, από τα οποία προκύπτει ο ταξινομητής ο οποίος παράγει τα βάρη $W_S(d_j)$ και $W_L(d_j)$. Στην εργασία επιλέχθηκε $f(\lambda)=\lambda$, ανεξαρτήτως του αλγορίθμου μάθησης που χρησιμοποιείται. Όπως αποδεικνύεται στο [Duda & Hart 1973], αν τα βάρη $W_S(d_j)$ και $W_L(d_j)$ είναι οι υπό συνθήκη πιθανότητες το d_j να είναι πραγματικά spam ($P(C=spam\,|\,d_j)$) και θεμιτό ($P(C=legitimate\,|\,d_j)$), αντίστοιχα, τότε η επιλογή της $f(\lambda)$ να είναι ίση με λ αποφέρει το ελάχιστο προσδοκώμενο κόστος (expected cost).

Δεδομένου πως το πρόβλημα του φιλτραρίσματος των spam mails είναι πρόβλημα μη επικαλυπτόμενης κατηγοριοποίησης (ένα mail πρέπει να καταταχθεί σε ακριβώς μία κατηγορία), αν τα βάρη $W_S(d_j)$ και $W_L(d_j)$ κανονικοποιηθούν έτσι ώστε να βρίσκονται στο διάστημα [0,1], θα ισχύει $W_S(d_j)+W_L(d_j)=1$. Έτσι, και με δεδομένη την προσέγγιση $f(\lambda)=\lambda$, η σχέση (3.1) απλοποιείται στην:

$$\frac{W_S(d_j)}{1-W_S(d_j)}>\lambda \;\Rightarrow\; W_S(d_j)>\frac{\lambda}{1+\lambda} \;\Rightarrow\; W_S(d_j)>t\,, \qquad (3.2)$$

όπου $t=\dfrac{\lambda}{1+\lambda}$ και $\lambda=\dfrac{t}{1-t}$.

Όπως φαίνεται και από τη σχέση (3.2), το t αποτελεί το κατώφλι κατάταξης του συστήματος. Αν το κανονικοποιημένο στο [0,1] βάρος $W_S(d_j)$ ξεπερνάει αυτό το κατώφλι, το μήνυμα κατατάσσεται ως spam, διαφορετικά κατατάσσεται ως μη spam.

Ο προσδιορισμός της τιμής του λ εξαρτάται, εν γένει, τόσο από το χρήστη (δηλαδή από το πόσο εκτιμά ο ίδιος το κόστος μιας κατάταξης $L\to S$ σε σχέση με μια $S\to L$), όσο και από την πολιτική διαχείρισης των κατηγοριοποιηθέντων ως spam μηνυμάτων. Για παράδειγμα, αν τα τελευταία διαγράφονται χωρίς άλλη επεξεργασία ή επέμβαση του χρήστη, το λ θα πρέπει να έχει "αρκετά μεγάλη" τιμή, καθώς οι περισσότεροι χρήστες θα θεωρούσαν απαράδεκτο να χάσουν ένα μη spam mail τους.

Μια εναλλακτική λύση που θα μπορούσε να προταθεί για ένα πραγματικό σύστημα είναι να υπάρχουν προκαθορισμένα (ή και οριζόμενα από το χρήστη) *σενάρια χρήσης*, το καθένα από τα οποία θα αντιστοιχούσε σε συγκεκριμένη πολιτική διαχείρισης των spam και συγκεκριμένη τιμή του λ. Αυτό διευκολύνει την αλληλεπίδραση με το χρήστη, ο οποίος δεν είναι υποχρεωμένος να επιλέξει ρητά το λ (αλλά μπορεί να το πράξει αν το επιθυμεί). Τα πειράματα που έγιναν στα πλαίσια της εργασίας θεωρούν τρία σενάρια χρήσης:

- Σύμφωνα με το πρώτο σενάριο, τα μηνύματα που κατατάσσονται ως spam διαγράφονται χωρίς άλλη επεξεργασία και έγκριση του χρήστη. Η τιμή που επιλέχθηκε για το λ είναι 999, που αντιστοιχεί σε κατώφλι $t = 0.999$. Το φίλτρο δηλαδή, πρέπει να έχει πάνω από 99.9% πίστη πως το μήνυμα είναι spam για να το κατάταξει έτσι, μια κατάταξη η οποία θα προκαλέσει τη διαγραφή του μηνύματος.
- Σύμφωνα με το δεύτερο σενάριο, το μήνυμα δε διαγράφεται άμεσα, αλλά στέλνεται πίσω στον αποστολέα, αναφέροντας πως το μήνυμά του αναγνωρίστηκε ως spam από το φίλτρο και ζητώντας από αυτόν να το επαναπροωθήσει σε μια ιδιωτική μη φιλτραριζόμενη διεύθυνση του χρήστη, αν υπάρχει (δείτε και το [Hall 1998]). Αυτή η ιδιωτική διεύθυνση δε θα διαφημίζεται πουθενά (π.χ. σε ιστοσελίδες ή σε λίστες αλληλογραφίας και ομάδες συζήτησης), οπότε σπάνια θα δέχεται spam. Επίσης, θα μπορούσε να ζητηθεί από τον αποστολέα να απαντήσει στο νέο γράμμα του σε ένα τυχαία επιλεγόμενο γρίφο (π.χ. «Συμπεριέλαβε στο θέμα του γράμματος την ερώτηση "ποια είναι η πρωτεύουσα της Γαλλίας" και την απάντησή της») για να διασφαλιστεί πως η απάντηση δε θα σταλθεί από ρομπότ που στέλνει spam. Μία λογική τιμή του λ για αυτή την περίπτωση είναι η 9 ($t = 0.9$): Το μπλοκάρισμα ενός θεμιτού μηνύματος τιμωρείται σχετικά ήπια, για να μοντελοποιήσει το γεγονός της καθυστέρησης και της ενόχλησης του αποστολέα, ο οποίος θα πρέπει να ξαναστείλει το μήνυμά του.
- Στο τρίτο και τελευταίο σενάριο, δεν απασχολεί τον αποδέκτη η επιπλέον δουλειά που επιβάλλει στον αποστολέα, οπότε δίνεται ίση βαρύτητα για τις αποτυχίες $L \to S$ και $S \to L$ (λ=1, $t = 0.5$). Επίσης, η ίδια τιμή θα μπορούσε να χρησιμοποιηθεί αν η πολιτική διαχείρισης των μηνυμάτων είναι να παρουσιάζονται αυτά στο χρήστη με ειδικό σημάδι που υποδεικνύει την κατάταξή τους, έτσι ώστε να μπορεί ο χρήστης να διαβάζει πρώτα τα κατάταγμένα ως θεμιτά μηνύματα και στο τέλος να επιλέγει αν θα διαβάσει ή θα διαγράψει τα καταταγμένα ως spam. Μια ακόμα επιλογή είναι να παρουσιάζονται τα νέα μηνύματα στο χρήστη κατά φθίνουσα διάταξη ως προς την πίστη $W_L(d_j)$ να είναι θεμιτά ([Drucker et al. 1999]).

Εκτός από την πολιτική κατηγοριοποίησης, η έννοια του κόστους επηρεάζει ουσιαστικά τον ορισμό του μέτρου αξιολόγησης της αποτελεσματικότητας του φίλτρου. Οποιοδήποτε τέτοιο μέτρο πρέπει να λαμβάνει υπόψη του την ασυμμετρία των αποτυχιών $L \to S$ και $S \to L$ συναρτήσει του λ. Η προσέγγιση που ακολουθήθηκε είναι να αντιμετωπίζεται κάθε θεμιτό μήνυμα ως λ spam: έτσι όταν ένα θεμιτό μήνυμα κατατάσσεται λάθος, αυτό ισοδυναμεί με την λανθασμένη κατάταξη λ spam, και όταν κατατάσσεται σωστά, ισοδυναμεί με λ spam επιτυχίες.

Κάθε μέτρο μπορεί να οριστεί με βάση τον *πίνακα ενδεχομένων*, ο οποίος αναφέρθηκε στην ενότητα (2.C.III.b). Αυτός παίρνει στο συγκεκριμένο πρόβλημα την παρακάτω μορφή:

Πίνακας συμπτώσεων		*Πραγματική κατηγορία*	
		spam	**legitimate**
Προβλεφθείσα κατηγορία	**spam**	$N_{S\to S}$	$N_{L\to S}$
	legitimate	$N_{S\to L}$	$N_{L\to L}$

Εδώ, $N_{S\to S}$ και $N_{L\to L}$ είναι το πλήθος των σωστών προβλέψεων για τα spam και τα θεμιτά μηνύματα, αντίστοιχα, ενώ $N_{S\to L}$ και $N_{L\to S}$ είναι το πλήθος των

λανθασμένων προβλέψεων για τα spam και τα θεμιτά, αντίστοιχα.

Αν οι κλάσεις spam και non-spam είχαν ίση βαρύτητα, δύο κατάλληλα μέτρα αξιολόγησης θα ήταν η *ακρίβεια (accuracy - Acc)* και ο *ρυθμός λαθών (error rate - Err)* ($Err = 1 - Acc$), τα οποία εκφράζονται ως:

$$Acc = \frac{N_{L\to L} + N_{S\to S}}{N_L + N_S}, \qquad Err = \frac{N_{L\to S} + N_{S\to L}}{N_L + N_S}$$

όπου $N_L = N_{L\to L} + N_{L\to S}$ το πλήθος των θεμιτών μηνυμάτων προς κατάταξη και $N_S = N_{S\to S} + N_{S\to L}$ το αντίστοιχο πλήθος για τα spam.

Η ακρίβεια και ο ρυθμός λαθών αποδίδουν ίση βαρύτητα στους δύο τύπους λαθών $L \to S$ και $S \to L$. Η γενίκευση αυτών των μέτρων έτσι ώστε να αντικατοπτρίζουν τη διαφορετική αποδιδόμενη βαρύτητα, σύμφωνα με την προσέγγιση να αντιμετωπίζεται ένα θεμιτό μήνυμα ως λ spam, είναι η *αποτιμημένη ακρίβεια (weighted accuracy –WAcc)* και ο *αποτιμημένος ρυθμός λαθών (weighted error rate –WErr)* ($WErr = 1 - WAcc$):

$$WAcc = \frac{\lambda \cdot N_{L\to L} + N_{S\to S}}{\lambda \cdot N_L + N_S} \qquad WErr = \frac{\lambda \cdot N_{L\to S} + N_{S\to L}}{\lambda \cdot N_L + N_S}$$

Οι τιμές της ακρίβειας και της αποτιμημένης εκδοχής της είναι συχνά απατηλά υψηλές. Αυτό συμβαίνει γιατί το πλήθος των θεμιτών μηνυμάτων είναι αρκετές φορές μεγαλύτερο από αυτό των spam, με αποτέλεσμα να υπερισχύουν οι όροι $N_{L\to L}$ και N_L έναντι των $N_{S\to S}$ και N_S. Για να σχηματιστεί μια πιο σωστή εικόνα για την αποτελεσματικότητα του φίλτρου, συνηθίζεται να συγκρίνεται η ακρίβεια ή ο ρυθμός λαθών με αυτά ενός απλοϊκού ταξινομητή, ο οποίος αποτελεί τη βάση αναφοράς ("baseline"). Εδώ θεωρούμε ως βάση αναφοράς την περίπτωση που δεν υπάρχει φίλτρο (ή ισοδύναμα το φίλτρο είναι *ο τετριμμένος απορρίπτης* για την κατηγορία spam): όλα τα θεμιτά μηνύματα γίνονται (σωστά) αποδεκτά και όλα τα spam περνάνε (λανθασμένα). Η αποτιμημένη ακρίβεια και ο αποτιμημένος ρυθμός λαθών για αυτή την περίπτωση είναι:

$$WAcc^b = \frac{\lambda \cdot N_L}{\lambda \cdot N_L + N_S} \qquad WErr^b = \frac{N_S}{\lambda \cdot N_L + N_S}$$

Για τη σύγκριση της επίδοσης του φίλτρου ως προς τη βάση, εισάγεται ο *ολικός λόγος κόστους (total cost ratio – TCR)*:

$$TCR = \frac{WErr^b}{WErr} = \frac{N_S}{\lambda \cdot N_{L\to S} + N_{S\to L}}$$

Μεγαλύτερο *TCR* υποδεικνύει καλύτερη επίδοση. Αν $TCR < 1$, η μη χρήση του φίλτρου είναι προτιμότερη. Αν το κόστος είναι ανάλογο του διατιθέμενου χρόνου, το *TCR* μετράει πόσος χρόνος χάνεται για την χειρωνακτική επεξεργασία όλων των spam μηνυμάτων (διαγραφή ή/και ανάγνωση) όταν δεν υπάρχει φίλτρο (N_S) ως προς το χρόνο που χάνεται για την επεξεργασία των spam που περνούν το φίλτρο $(N_{S\to L})$ συν το χρόνο που απαιτείται για την ανάκτηση των θεμιτών μηνυμάτων που (λανθασμένα) μπλοκαρίστηκαν $(\lambda \cdot N_{L\to S})$.

Το *TCR*, αν και απλό στη σύλληψή του, παρουσιάζει δύο μειονεκτήματα: Το ένα είναι πως δεν προσφέρει εύκολα διαισθητική αντίληψη, πέραν από αυτή που αναφέρθηκε στην

προηγούμενη παράγραφο. Αυτό γίνεται πιο εμφανές για το άνω του φράγμα, το οποίο είναι άπειρο, αν δε σημειωθεί κανένα λάθος κατάταξης (τέλειο φίλτρο). Δεν υπάρχει δηλαδή κάποιο σταθερό (πεπερασμένο) σημείο αναφοράς, με το οποίο να συγκρίνεται η απόδοση του, έτσι ώστε να υπάρχει η αίσθηση της απόστασης που το χωρίζει από το τέλειο φίλτρο. Επίσης προβληματικό, αν και σαφώς λιγότερα σημαντικό, είναι το γεγονός πως το κάτω φράγμα του δεν είναι σταθερό, αλλά εξαρτάται από τη σύνθεση του συνόλου ελέγχου. Συγκεκριμένα είναι: $TCR_{\min} = \frac{N_S}{\lambda \cdot N_L + N_S}$, αν όλα τα μηνύματα ταξινομηθούν λάθος. Πάλι δε μπορούμε να πούμε εύκολα το πόσο κοντά στον χείριστο ταξινομητή είμαστε αν δε λάβουμε υπόψη τη σύνθεση του δείγματος και το λ.

Αλληλένδετο με το παραπάνω πρόβλημα είναι η έλλειψη γραμμικότητας του *TCR* ως προς τις αποφάσεις του ταξινομητή, δηλαδή τις τυχαίες μεταβλητές $N_{S\to L}$ και $N_{L\to S}$. Αυτό δημιουργεί δυσκολία διαισθητικής αντίληψης των πραγματικών διαφορών στην απόδοση διαφορετικών αλγορίθμων (ή του ίδιου αλγορίθμου για διαφορετικά σύνολα εκπαίδευσης) πάνω στο ίδιο σύνολο ελέγχου. Μία παραπάνω λανθασμένη ή σωστή απόφαση επηρεάζει τον παρονομαστή και είναι δυνατόν να οδηγήσει σε μεγάλη μεταβολή του *TCR*, κυρίως σε σενάρια υψηλού λ (όπως θα φανεί στη συνέχεια) και περιπτώσεις υψηλής ακρίβειας, όπου τα $N_{S\to L}$ και $N_{L\to S}$ είναι αρκετά μικρά.

Ως απόρροια της μη γραμμικότητας, ο υπολογισμός του μέσου όρου των *TCR* για διαφορετικούς ταξινομητές δίνει παραπλανητικά αποτελέσματα: τα αρνητικά αποτελέσματα (κάτω της βάσης) κινούνται στο διάστημα $\left[\frac{N_S}{\lambda \cdot N_L + N_S}, 1\right]$, ενώ τα θετικά στο $(1, +\infty)$. Έτσι η παρουσία των αρνητικών δειγμάτων υποβιβάζεται στο μέσο όρο προς όφελος των θετικών. Γι' αυτό το λόγο, όταν θέλουμε το μέσο *TCR* μιας σειράς από ταξινομητές, πρώτα υπολογίζουμε το μέσο όρο των *WErr και WErr*^{b} (τα οποία *είναι* γραμμικά ως προς τις αποφάσεις των ταξινομητών) και στη συνέχεια παίρνουμε το *TCR* ως το λόγο του δεύτερου προς το πρώτο.

Παρά τα παραπάνω προβλήματα, το *TCR* χρησιμοποιείται στην εργασία για να υπάρχει συνέχεια με τα προηγούμενα αποτελέσματα που έχουν δημοσιευτεί πάνω στην ίδια συλλογή μηνυμάτων.

3.D) Αποτελέσματα προηγούμενων πειραμάτων

Τα πειράματα που διενεργήθηκαν κατά τη διάρκεια της εργασίας βασίζονται σημαντικά και αποτελούν συνέχεια μιας σειρά πειραμάτων που πραγματοποιήθηκαν πάνω στην ίδια συλλογή (Ling-Spam) στο Ε.ΚΕ.Φ.Ε. "Δημόκριτος" κατά την περίοδο 1999-2000. Ο αλγόριθμος μάθησης που χρησιμοποιήθηκε ήταν ο απλοϊκός ταξινομητής Μπαίυζ (NB). Στο [Androutsopoulos et al. 2000a] δημοσιεύονται τα πρώτα αποτελέσματα αυτών των πειραμάτων, όπου φαίνεται η επίδραση στην επίδοση του NB της λημματοποίησης, της απομάκρυνσης λειτουργικών/συχνών λέξεων που δίνονται σε διάφορες λίστες (*stop-lists*) και της διαστασιμότητας των διανυσμάτων αναπαράστασης των μηνυμάτων. Από αυτά

προκύπτει πως η λημματοποίηση αυξάνει ελαφρώς το *TCR* για μικρές διαστασιμότητες (50-150 features), όπου γενικά σημειώνονται οι καλύτερες επιδόσεις. Καθώς αυξάνεται η διαστασιμότητα, η πορεία του *TCR* είναι έντονα πτωτική, ένδειξη πως ο NB επηρεάζεται σημαντικά από το υπερταίριασμα με τα δεδομένα εκπαίδευσης, γεγονός που οφείλεται στην χαμηλή διαχωριστική ικανότητα των features που συμμετέχουν στην αναπαράσταση. Η χρήση stop-list, από την άλλη, δεν έχει πρακτική επίδραση στην απόδοση.

Με βάση αυτά τα στοιχεία, αποφασίστηκε πως στα πειράματα για την εργασία θα γίνει λημματοποίηση και δε θα χρησιμοποιηθεί stop-list. Αργότερα επακολούθησαν και πρόσθετα πειράματα, με σημαντικότερη διαφοροποίηση την εκτίμηση των πιθανοτήτων του τύπου του NB μέσω *m-εκτίμησης* (με τις συχνά επιλεγόμενες τιμές παραμέτρων p=0.5 και m=1), αντί της απλής εκτίμησης ως συχνότητα στα σύνολα εκπαίδευσης, που εφαρμόστηκε αρχικά. Τα νέα αποτελέσματα ήταν αισθητά καλύτερα από τα πρώτα. Η λημματοποίηση είχε πάλι θετική επίδραση, αλλά σε αντίθεση με πριν, η χρήση stop-list αποδείχθηκε επωφελής. Ωστόσο στα πειράματα της εργασίας, αν και χρησιμοποιούνται m-εκτιμήσεις για τις πιθανότητες, δε χρησιμοποιήθηκε stop-list, για το λόγο πως είχε ήδη ξεκινήσει η εργασία όταν έγιναν γνωστά τα νέα αποτελέσματα και δεν υπήρχε χρόνος να ξαναγίνουν όλα από την αρχή με stop-list. Πάντως, τόσο η επίδραση της stop-list, όσο και της λημματοποίησης δεν αποτελούσαν παραμέτρους προς διερεύνηση στα πλαίσια της εργασίας, καθώς οι υπόλοιπες παράμετροι που υπάρχουν είναι ήδη αρκετές, όπως θα παρουσιαστεί στο επόμενο κεφάλαιο.

Στο διάγραμμα (3-1) παριστάνονται οι καμπύλες *TCR* για τον NB ως συνάρτηση της διαστασιμότητας[*]. Κάθε καμπύλη αντιστοιχεί σε ένα από τα τρία θεωρούμενα σενάρια χρήσης. Για την ενίσχυση της αξιοπιστίας των αποτελεσμάτων, εφαρμόστηκε *διασταυρωμένη επικύρωση* 10 σημείων (βλ. ενότητα (2.C.III.b)).

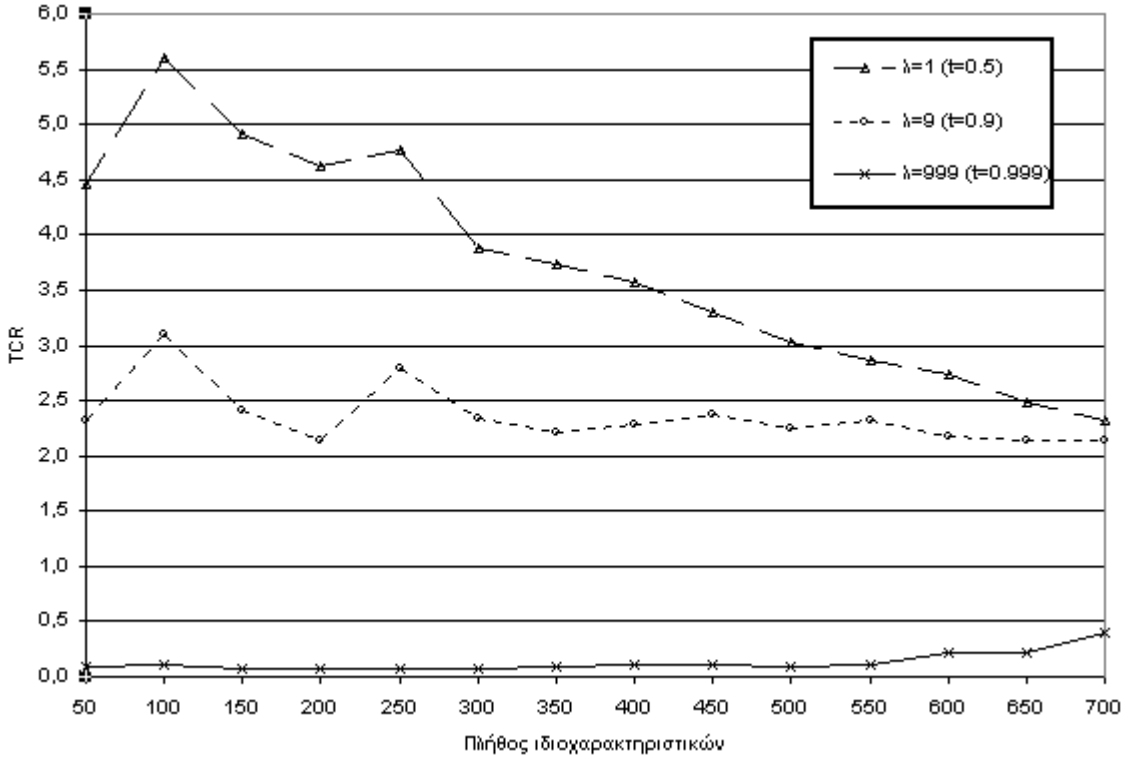
Η πρώτη παρατήρηση που μπορεί να γίνει είναι πως η απόδοση του ταξινομητή υποβιβάζεται όσο αυστηρότερος γίνεται αυτός ως προς την τιμωρία των $L \to S$ αστοχιών. Αυτό είναι αναμενόμενο, αφού η αποτιμημένη ακρίβεια $WAcc^b$ της βάσης αναφοράς αυξάνεται με την αύξηση του λ, και επομένως μειώνονται τα περιθώρια του φίλτρου να σημειώσει μεγαλύτερη *WAcc* απ'ό,τι αν δεν χρησιμοποιηθεί φίλτρο. Στο σενάριο διαγραφής των προβλεπόμενων ως spam μηνυμάτων ($\lambda = 999$), το *TCR* είναι συνεχώς κάτω της βάσης, γιατί ένα και μόνο λάθος $L \to S$ ισοδυναμεί με 999 $S \to L$. Στην πράξη αυτό σημαίνει πως στο συγκεκριμένο σενάριο δεν πρέπει να καταταχθεί κανένα θεμιτό μήνυμα λάθος για να υπάρξει άνω της βάσης απόδοση, κάτι που δεν κατορθώνει για καμία διαστασιμότητα ο NB. Για τα άλλα δύο σενάρια, η εικόνα είναι σαφώς καλύτερη της βάσης.

Μια δεύτερη παρατήρηση πάνω στο διάγραμμα (3-1) είναι η πτωτική πορεία της καμπύλης για $\lambda = 1$ και η σταθερή σχετικά πορεία της $\lambda = 9$ ως προς τη διαστασιμότητα. Μια πρώτη εξήγηση μπορεί να δοθεί από τα δύο επόμενα διαγράμματα. Στο 3-2 παριστάνεται η *ανάκληση* για τα spam (spam recall – *SR*) για κάθε σενάριο, η οποία ορίζεται ως το ποσοστό των spam που μπλοκαρίστηκαν από το φίλτρο, ενώ στο 3-3 δίνεται η *ορθότητα* για τα spam (spam precision – *SP*), η οποία ορίζεται ως το ποσοστό των σωστά καταταγμένων ως spam (βλ. ενότητα (2.C.III.a)):

[*] Αφορούν την περίπτωση που γίνεται λημματοποίηση και δε χρησιμοποιείται stop-list.

$$SR = \frac{N_{S\to S}}{N_{S\to S} + N_{S\to L}} \qquad SP = \frac{N_{S\to S}}{N_{S\to S} + N_{L\to S}}$$

Ισοδύναμα, η *SR* μπορεί να γραφεί ως $\frac{N_S - N_{S\to L}}{N_S}$, και δεδομένου πως το N_S είναι σταθερό για ένα δοθέν σύνολο ελέγχου, η *SR* είναι μέτρο των λαθών $S \to L$. Η *SP*, από την άλλη μεριά, γράφεται ως $\frac{N_S - N_{S\to L}}{N_S + (N_{L\to S} - N_{S\to L})}$ και είναι σαφές πως επηρεάζεται και από τα δύο είδη λάθους. Ωστόσο, αποδεικνύεται εύκολα πως αν $N_{L\to S} < N_{S\to S} - 1$, κάτι που στην πράξη συμβαίνει συνήθως, τότε ένα λάθος $L \to S$ υποβιβάζει την *SP* περισσότερο από ένα $S \to L$. Προσεγγιστικά λοιπόν μπορούμε να θεωρούμε πως η *SP* είναι μέτρο κυρίως των $L \to S$ λαθών.

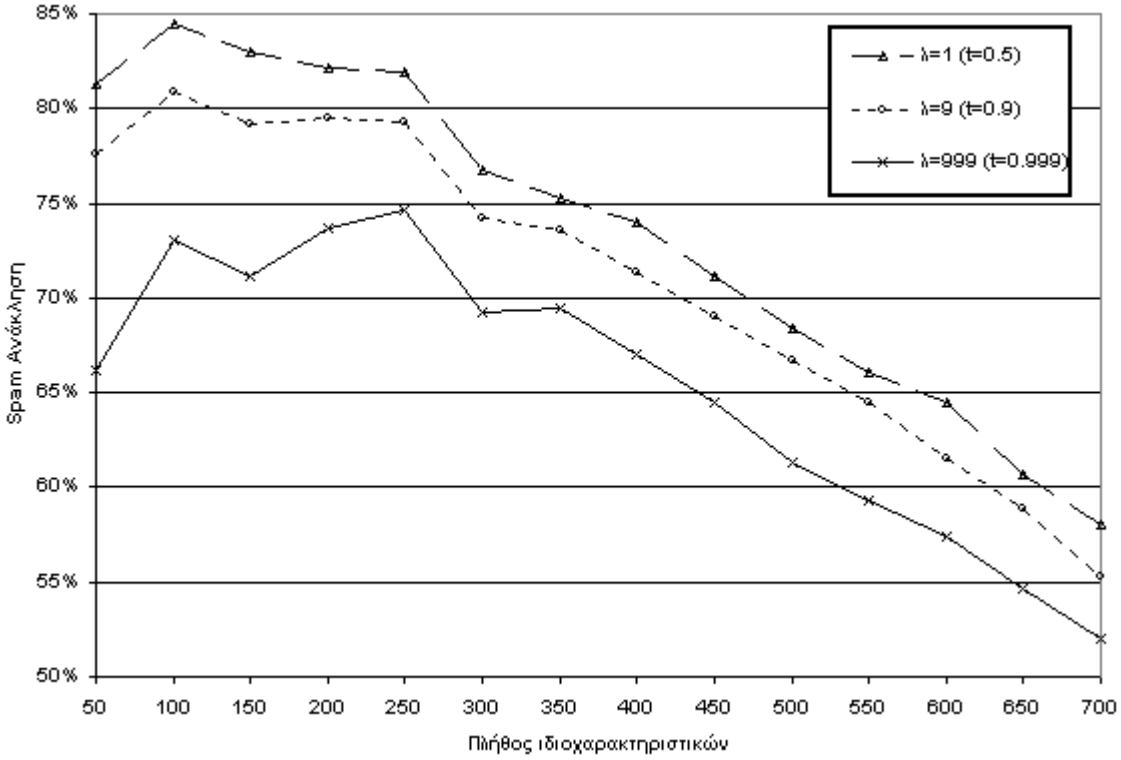


Διάγραμμα 3-1: *TCR* του NB (με λημματοποίηση / χωρίς stoplist).

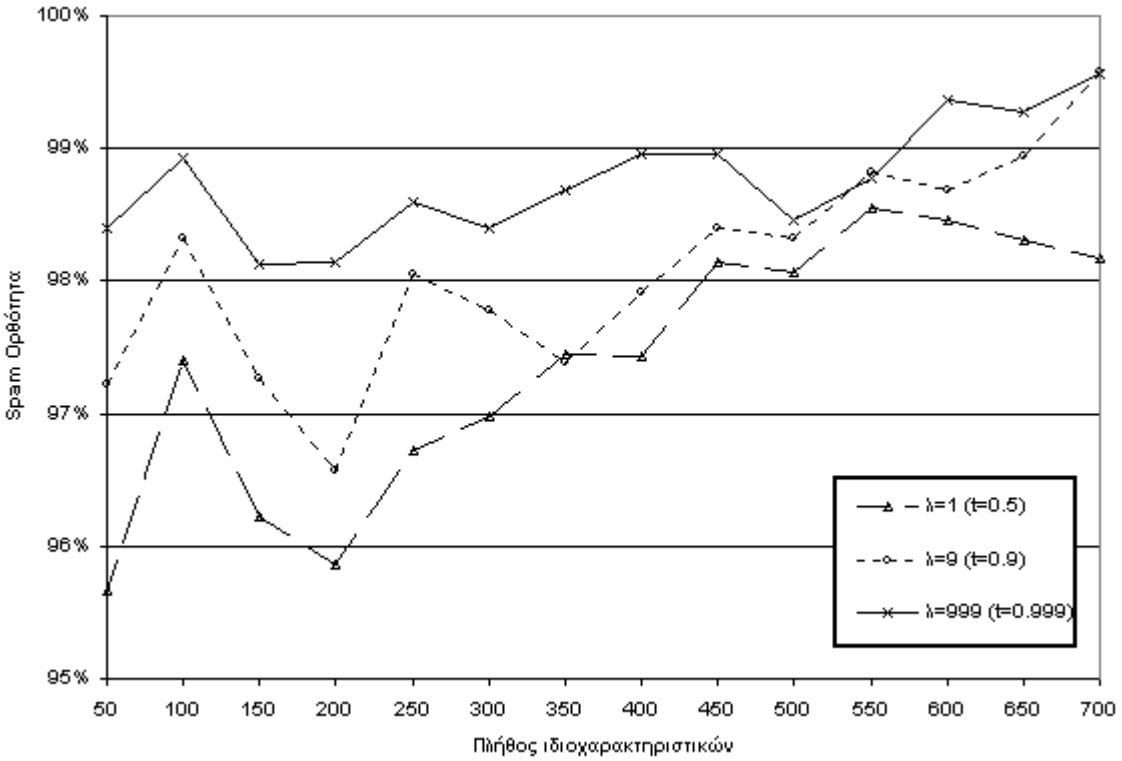
Η *SR* είναι για όλα τα σενάρια έντονα πτωτική, ενώ η, ούτως ή άλλως υψηλή, *SP* παρουσιάζει ελαφρά άνοδο. Για λ=1, η μικρή άνοδος της *SP* ($\cong 3\%$) δεν αντισταθμίζει την σημαντική πτώση της *SR* ($\cong 23\%$). Για λ=9, αν και παρατηρείται περίπου η ίδια μεταβολή των μεγεθών όπως και για λ=1, η *SP* εδώ είναι πιο σημαντική στη διαμόρφωση του *TCR*, αφού τα λάθη $L \to S$ τιμωρούνται 9 φορές περισσότερο από τα $S \to L$. Γι' αυτό και μια μικρή άνοδος της *SP* αντισταθμίζει περίπου μια μεγάλη πτώση της *SR*.

Στα διαγράμματα 3-2 και 3-3 δίνεται επίσης μια πρώτη αίσθηση των δυνατοτήτων της αυτόματης κατηγοριοποίησης μέσω τεχνικών μηχανικής μάθησης. Κι ένας απλός αλγόριθμος μάθησης, όπως είναι ο NB, μπορεί και συγκρατεί ακόμα και σε ένα πολύ αυστηρό σενάριο χρήσης πάνω από το 50% των spam μηνυμάτων, ενώ ξεπερνάει το 80% για λιγότερο αυστηρά

σενάρια. Η ορθότητά του δεν πέφτει κάτω του 95%, ενώ στο αυστηρό σενάριο ξεπερνάει το 99.5% ! Στα επόμενα κεφάλαια, οι επιδόσεις αυτές θα βελτιωθούν ακόμα περισσότερο με τη βοήθεια πιο εξελιγμένων τεχνικών.



Διάγραμμα 3-2: Spam ανάκληση του NB (με λημματοποίηση / χωρίς stoplist).



Διάγραμμα 3-3: Spam ορθότητα του NB (με λημματοποίηση / χωρίς stoplist).

4) ΠΕΙΡΑΜΑΤΑ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ ΤΩΝ *k*-ΚΟΝΤΙΝΟΤΕΡΩΝ ΓΕΙΤΟΝΩΝ

Σε αυτό το κεφάλαιο αρχίζει η παρουσίαση του κύριου μέρους της εργασίας, το οποίο συνίσταται στον πειραματισμό με αλγορίθμους και τεχνικές με στόχο τη βελτίωση της απόδοσης του αυτόματου ταξινομητή, έχοντας ως μέτρο το *TCR*. Στο βαθμό που ήταν δυνατό, επιχειρήθηκαν να ερμηνευθούν τα παρατηρούμενα αποτελέσματα, παρέχοντας ενδεχομένως χρήσιμες γενικεύσεις και για άλλες εφαρμογές αυτόματης κατηγοριοποιήσης. Το παρόν κεφάλαιο περιλαμβάνει την παρουσίαση των πειραμάτων που έγιναν με βάση τον αλγόριθμο μάθησης των *k*-κοντινότερων γειτόνων (*k*-NN) και το επόμενο τα αντίστοιχα πειράματα για μία συνδυαστική αρχιτεκτονική ταξινομητών βασισμένη στον *k*-NN και στον NB.

Ο *k*-NN έχει χρησιμοποιηθεί στο παρελθόν στην περιοχή της κατηγοριοποίησης κειμένου (π.χ. [Lam et al. 1999], [Li & Jain 1998], [Yang & Liu 1999]), επιδεικνύοντας ανάλογα καλά αποτελέσματα που έχει σημειώσει και σε άλλους τομείς. Το γεγονός αυτό, σε συνδυασμό με την εύκολη διαισθητική ερμηνεία της λειτουργίας του αλγορίθμου και την ύπαρξη δημόσια διαθέσιμου λογισμικού που τον υλοποιεί, οδήγησαν στην απόφαση να δοκιμαστεί κατά την εκπόνηση της εργασίας.

Ειδικότερα όσον αφορά την υλοποίηση, χρησιμοποιήθηκε κατά κύριο λόγο το πρόγραμμα TiMBL (Tilburg Memory Based Learner) [Daelemans et al. 2000]. Το TiMBL παρέχει το βασικό αλγόριθμο κατηγοριοποίησης βασισμένο στη μνήμη, πλαισιωμένο από επεκτάσεις όπως η αποτίμηση χαρακτηριστικών (feature weighting) και υποστηριζόμενο εσωτερικά από δομές ευρετηριοποίησης οι οποίες καθιστούν την αναζήτηση των γειτονικών στιγμιοτύπων κατά τη φάση κατάταξης πολύ πιο γρήγορη απ'ότι αν βασιζόταν στη συνήθη επίπεδη οργάνωση των δεδομένων που παρέχει μια άμεση υλοποίηση του αλγορίθμου. Προσφέρει ακόμα διεπαφή προς τον προγραμματιστή εφαρμογών σε γλώσσα C++, γεγονός που έκανε πιο ελκυστική την υιοθέτησή του, καθώς ο μηχανισμός αποτίμησης γειτόνων με βάση την απόσταση (distance weighting) δεν υποστηρίζεται άμεσα (τουλάχιστον μέχρι την έκδοση 3.0 που ήταν διαθέσιμη) και έπρεπε να υλοποιηθεί ξεχωριστά. Αν και το TiMBL σχεδιάστηκε με στόχο το χώρο της επεξεργασίας φυσικής γλώσσας, μπορεί να χρησιμοποιηθεί εξίσου και σε άλλα πεδία (π.χ. υποστηρίζει τόσο συμβολικά, όσο και αριθμητικά features).

Μία διαφορά του TiMBL από τον *k*-NN που πρέπει να τονιστεί είναι πως το *k* δεν δηλώνει το πλήθος των κοντινότερων γειτονικών στιγμιοτύπων που ορίζουν τη γειτονία, αλλά το πλήθος των κοντινότερων *αποστάσεων* από το άγνωστο στιγμιότυπο που ορίζουν τη γειτονία. Αν υπάρχουν περισσότεροι από ένας γείτονες σε κάποια από τις *k* αποστάσεις, ο αλγόριθμος θεωρεί παραπάνω από *k* γείτονες[*].

[*] Και ο "κλασικός" k-NN είναι δυνατόν να θεωρήσει περισσότερους από k γείτονες, μόνο όμως αν υπάρξουν ισοβαθμίες (π.χ., ο 1-NN θα θεωρήσει δύο τουλάχιστον γείτονες αν υπάρχουν δύο στιγμιότυπα που ισαπέχουν από το άγνωστο κατά ελάχιστη απόσταση).

4.Α) Παράμετροι προς διερεύνηση

Η τελική απόδοση ενός συστήματος είναι συνάρτηση πολλών παραμέτρων. Στην ιδανική περίπτωση αυτές οι παράμετροι είναι ανεξάρτητες μεταξύ τους, γεγονός που απλοποιεί ουσιαστικά την πειραματική βελτιστοποίηση του τελικού συστήματος. Ο λόγος είναι πως μπορεί να μελετηθεί κάθε παράμετρος χωριστά από τις υπόλοιπες, να εκτιμηθεί η βέλτιστη τιμή της και στη συνέχεια να μελετηθούν οι υπόλοιπες θεωρώντας αυτήν ως σταθερά με τιμή τη βέλτιστη. Έτσι, το πλήθος των βημάτων που απαιτούνται για την πειραματική βελτιστοποίηση του συστήματος είναι όσο και το *άθροισμα* του πλήθους των διαφορετικών τιμών που παίρνουν οι παράμετροί του (αν κάποιες είναι συνεχείς, είναι αναγκαία η διακριτοποίηση τους).

Στην πράξη, οι παράμετροι είναι συχνά αλληλοεξαρτώμενες. Στη χειρότερη περίπτωση που όλες οι παράμετροι αλληλοεξαρτώνται, οι δυνατοί συνδυασμοί που πρέπει να μελετηθούν είναι το καρτεσιανό γινόμενο των συνόλων τιμών των παραμέτρων, γεγονός που ανεβάζει το πλήθος των βημάτων για τη βελτιστοποίηση στο *γινόμενο* του πλήθους των τιμών των παραμέτρων. Κάτι τέτοιο είναι συνήθως απαγορευτικό από άποψη χρόνου. Μία πρακτική αντιμετώπιση του προβλήματος είναι να γίνουν κάποιες υποθέσεις ανεξαρτησίας μεταξύ συγκεκριμένων παραμέτρων, οι οποίες στην καλύτερη περίπτωση επιβεβαιώνονται μέσω δειγματοληπτικής επιλογής κάποιων τιμών των θεωρούμενων παραμέτρων και διενέργειας πειραμάτων με τις τιμές αυτές. Αν οι υποθέσεις επιβεβαιωθούν, προσεγγιστικά έστω, υιοθετούνται για τη συνέχεια της βελτιστοποίησης. Αφού σταθεροποιηθεί μία (υποτιθέμενη) ανεξάρτητη παράμετρος στη βέλτιστη, κατά τα πειράματα, τιμή της, η ίδια διαδικασία πρότασης και υιοθέτησης υποθέσεων μπορεί να συνεχιστεί για τις υπόλοιπες παράμετρους, στο βαθμό που επιβεβαιώνεται από τα πειραματικά δεδομένα. Επιχειρείται έτσι μια ευριστική διερεύνηση του χώρου των παραμέτρων αντί μιας εξαντλητικής αναζήτησης του βέλτιστου συνδυασμού. Αυτή η προσέγγιση ακολουθήθηκε και στην εργασία όσον αφορά τις παραμέτρους του συστήματος αυτόματης εξαγωγής ταξινομητή.

Μια διάκριση που μπορεί να γίνει ως προς το ρόλο των παραμέτρων είναι αν αυτές προέρχονται από τις ίδιες τις απαιτήσεις του προβλήματος ("εξωτερικές") ή είναι "εσωτερικές" της μεθόδου επίλυσής του. Οι πρώτες μπορούν να θεωρηθούν ανεξάρτητες, επειδή καθορίζονται από το χρήστη κατά βούληση, ενώ οι δεύτερες εξαρτώνται από την επιλογή των πρώτων και αποφασίζονται από το σχεδιαστή του συστήματος. Το πρόβλημα της βελτιστοποίησης συνίσταται στην επιλογή ενός συνδυασμού εξαρτημένων μεταβλητών για κάθε συνδυασμό από ανεξάρτητες, έτσι ώστε να μεγιστοποιείται μία συνάρτηση-στόχος.

Στην εφαρμογή της αυτόματης εξαγωγής ταξινομητή για φιλτράρισμα ηλεκτρονικών μηνυμάτων, μοναδική ανεξάρτητη μεταβλητή, σύμφωνα με τη μοντελοποίηση του προβλήματος, είναι το λ, το πόσο κοστίζει μια αποτυχία $L \to S$ σε σχέση με μια $S \to L$. Επομένως για κάθε λ πρέπει να εκτιμηθούν οι τιμές των υπόλοιπων "εσωτερικών" παραμέτρων. Ήδη στο κεφάλαιο 3 αναφέρθηκαν μερικές από τις τελευταίες: η χρήση ή μη λημματοποίησης και stop-list, οι παράμετροι της m-εκτίμησης και η μέθοδος επιλογής των features (feature selection). Αυτές αποτέλεσαν αντικείμενο προηγούμενων πειραμάτων ([Androutsopoulos et al. 2000a]) και στα πλαίσια της εργασίας θεωρούνται προκαθορισμένες

(με τη σιωπηλή υπόθεση πως δεν υπάρχει σημαντική αλληλεξάρτηση με τον αλγόριθμο μάθησης, δηλ. τον *k*-NN αντί του NB).

Μία παράμετρος που παρέμεινε αδέσμευτη είναι η διαστασιμότητα, δηλαδή το πλήθος των επιλεγόμενων χαρακτηριστικών αναπαράστασης. Το σύνολο τιμών της είναι το ίδιο με αυτό των προηγούμενων πειραμάτων: Τα N καλύτερα features (σύμφωνα με τη συνάρτηση πληροφοριακού κέρδους και τα δεδομένα εκπαίδευσης), με το N να κυμαίνεται από 50 έως 700, με βήμα 50. Οι υπόλοιπες παράμετροι είναι εσωτερικές του επιλεχθέντος αλγορίθμου μάθησης, του *k*-NN. Αυτές είναι:

1) Το *k*.
2) Η μέθοδος αποτίμησης των features (feature weighting).
3) Η συνάρτηση αποτίμησης των γειτόνων με βάση την απόσταση (distance weighting).

Στα διαγράμματα που θα ακολουθήσουν στο παρόν και στο επόμενο κεφάλαιο, θα απεικονισθούν τα πλέον χαρακτηριστικά αποτελέσματα για επιλεγμένες τιμές των υπό θεώρηση παραμέτρων. Πρέπει να σημειωθεί ωστόσο πως αυτά αποτελούν ένα μικρό μόνο μέρος του συνόλου των πειραμάτων που διενεργήθησαν και των αποτελεσμάτων που συγκεντρώθηκαν, τα οποία για λόγους έκτασης δεν παρατίθενται. Επίσης, συχνά κάποια αποτελέσματα είναι συναφή με άλλα εμφανιζόμενα, οπότε είναι περιττή η εμφάνισή τους. Η αναφορά σε αποτελέσματα που δεν παρουσιάζονται είναι γενικά επιγραμματική.

4.Β) Αποτίμηση χαρακτηριστικών

Η πρώτη παράμετρος για την οποία εξάχθηκαν ασφαλή γενικά συμπεράσματα ήταν η μέθοδος αποτίμησης των χαρακτηριστικών. Η μία επιλογή ήταν η απλή περίπτωση ισοβαρούς αποτίμησης, κατά την οποία γίνεται ίση αποτίμηση όλων των features. Οι δύο άλλες μέθοδοι που δοκιμάστηκαν ακολουθούν την προσέγγιση του φίλτρου και ήταν πιο επιτυχείς από την ισοβαρή. Μεταξύ των δύο, η αποτίμηση με βάση το *πληροφοριακό κέρδος* έδωσε καλύτερα αποτελέσματα. Το σημαντικότερο ήταν πως η ίδια εικόνα παρουσιάζεται γενικά ανεξαρτήτως των υπολοίπων παραμέτρων *λ*, *k* και διαστασιμότητας (η αλληλεξάρτηση με την παράμετρο αποτίμησης των γειτόνων δεν εξετάστηκε, για λόγο που θα αναφερθεί στην ενότητα (4.C)). Το γεγονός αυτό επιτρέπει τη δέσμευση της παραμέτρου και τη μείωση κατά ένα του πλήθους τους για τα επόμενα πειράματα.

4.Β.Ι) Μέτρα αποτίμησης

Ο "κλασικός" *k*-NN θεωρεί πως όλα τα features είναι ισάξια. Αυτή η προσέγγιση δεν είναι συνήθως η καλύτερη δυνατή. Στην ενότητα (2.Β.Ι) περιγράφηκαν οι τεχνικές του περιτυλίγματος και του φίλτρου, που επιχειρούν να αποδώσουν κατάλληλο βάρος σε κάθε feature, ανάλογα με την εκτιμώμενη αξία του στο διαχωρισμό των στιγμιοτύπων ανά κατηγορία. Στα πλαίσια της εργασίας ακολουθήθηκε η τεχνική του φίλτρου, δοκιμάζοντας δύο συχνά χρησιμοποιούμενες συναρτήσεις: το *πληροφοριακό κέρδος* (Information Gain – *IG*) και το *λόγο κέρδους* (Gain-Ratio – *GR*).

Για τη συνάρτηση πληροφοριακού κέρδους έγινε ήδη μια αναφορά στο κεφάλαιο 3, αφού η ίδια συνάρτηση χρησιμοποιήθηκε και για την επιλογή των features της αναπαράστασης. Πέρα από τον τύπο που δόθηκε εκεί, μία πιο κατανοητή ισοδύναμη μορφή του *IG* προκύπτει χρησιμοποιώντας μια θεμελιώδη έννοια της θεωρίας πληροφορίας, την εντροπία. Στη γενική της μορφή, αυτή η έκφραση του *IG* είναι:

$$IG(X,C) = H(C) - \sum_{x \in V_x} P(X = x) \cdot H(C \mid X = x), \qquad (4.1)$$

όπου X είναι πάλι ένα υποψήφιο χαρακτηριστικό, C η τυχαία μεταβλητή που παριστάνει την κατηγορία ενός στιγμιοτύπου, V_x το σύνολο τιμών του X και $H(C)$ η εντροπία της C. Η τελευταία ορίζεται ως:

$$H(C) \equiv -\sum_{c \in C} P(C = c) \cdot \log_2 P(C = c), \qquad (4.2)$$

όπου το άθροισμα γίνεται πάνω στο σύνολο των τιμών της τυχαίας μεταβλητής C. Η εντροπία στη θεωρία πληροφορίας είναι ένα μέτρο της αβεβαιότητας πρόβλεψης των τιμών μιας τυχαίας μεταβλητής. Μια φυσική της ερμηνεία είναι πως καθορίζει τον ελάχιστο αριθμό από δυαδικά ψηφία (bits) πληροφορίας που απαιτούνται για την κωδικοποίηση μιας αυθαίρετης τιμής της τυχαίας μεταβλητής.

Απουσία κάθε άλλης πληροφορίας, θα απαιτούνταν $H(C)$ bits πληροφορίας για την κωδικοποίηση μιας αυθαίρετης κατηγορίας. Ο τύπος (4.1) αφαιρεί από αυτή την ποσότητα την *αναμενόμενη* τιμή της εντροπίας αν είναι γνωστή η τιμή του χαρακτηριστικού X. Η αναμενόμενη εντροπία που δίνεται από το δεύτερο όρο είναι το άθροισμα των εντροπιών $H(C|X=x)$ για κάθε τιμή του X, χρησιμοποιώντας ως βάρος τις πιθανότητες εμφάνισης των τιμών του X. Το $IG(X,C)$ επομένως εκφράζει την αναμενόμενη μείωση της εντροπίας (άρα και της αβεβαιότητας) που θα προέλθει από τη γνώση της τιμής του X. Ισοδύναμα, είναι ο αριθμός των bits που εξοικονομούνται για την κωδικοποίηση της κατηγορίας δοθείσης της τιμής του X.

Ένα πλεονέκτημα του *IG* είναι πως δε λαμβάνει υπόψη μόνο την διαχωριστική ικανότητα της τιμής ενός feature, η οποία εκφράζεται από τον όρο $H(C|X=x)$, αλλά αξιολογεί και τη συχνότητα εμφάνισης της ($P(X=x)$). Έτσι, τιμές οι οποίες μειώνουν μεν πολύ την αβεβαιότητα πρόβλεψης της κατηγορίας (με την $H(C|X=x)$ αρκετά μικρή), αλλά εμφανίζονται σπάνια, δεν συνεισφέρουν σημαντικά στην αύξηση του *IG*. Επιπλέον, αυτό κάνει το *IG* πιο ανθεκτικό όσον αφορά την εκτίμηση των πιθανοτήτων. Αν και οι τελευταίες θα ήταν προτιμότερο να εκτιμώνται μέσω m-εκτίμησης για λόγους ομοιομορφίας με την διαδικασία επιλογής των όρων, το TiMBL τις εκτιμά με βάση τη συχνότητα τους στο σύνολο εκπαίδευσης[*].

Ένα μειονέκτημα που έχει το *IG* είναι πως υπερτονίζει την αξία χαρακτηριστικών με μεγάλο πλήθος τιμών. Τέτοια χαρακτηριστικά τείνουν να διαμερίζουν τα δεδομένα εκπαίδευσης σε πολύ μικρά υποσύνολα και κατά συνέπεια να καθορίζουν σε μεγάλο βαθμό από μόνα τους την κατηγορία ενός στιγμιοτύπου. Στην ακραία περίπτωση, ένα feature *κλειδί*, που καθορίζει μονοσήμαντα την κατηγορία, θα είχε μέγιστο *IG*. Στην πράξη όμως θα ήταν

[*] Μεταγενέστερα πειράματα έδειξαν πως δεν υπάρχει σημαντική διαφοροποίηση στη σειρά κατάταξης των χαρακτηριστικών είτε γίνεται m-εκτίμηση (με m=1 και p=0.5) είτε όχι.

άχρηστο αν κάθε νέο (άγνωστο) στιγμιότυπο είχε και νέα μη παρατηρηθείσα τιμή. Για παράδειγμα, ο αριθμός αστυνομικής ταυτότητας καθορίζει απολύτως έναν ασθενή που νοσηλεύεται σε ένα νοσοκομείο μια συγκεκριμένη περίοδο, άρα καθορίζει απολύτως και την ασθένειά του. Ωστόσο, κάθε νέος ασθενής έχει και διαφορετικό αριθμό ταυτότητας με αποτέλεσμα το χαρακτηριστικό αυτό, αν και άριστο στον προσδιορισμό της ασθένειας στο σύνολο εκπαίδευσης, να είναι άχρηστο για τον προσδιορισμό της ασθένειας ενός άγνωστου ασθενή.

Ένα εναλλακτικό του *IG* μέτρο που προτάθηκε για την αντιμετώπιση αυτού του προβλήματος είναι ο *λόγος κέρδους (Gain-Ratio – GR)* [Quinlan 1986]. Το *GR* διορθώνει την τιμή του *IG*, διαιρώντας την με έναν όρο που ονομάζεται *πληροφορία διαμερισμού (split information – SI)* του feature και η οποία ορίζεται ως:

$$SI(X) \equiv -\sum_{x \in V_x} P(X = x) \cdot \log_2 P(X = x) \qquad (4.3)$$

Το *SI(X)* δεν είναι τίποτα άλλο από την εντροπία του feature *X*. Το *GR* τότε ορίζεται ως:

$$GR(X, C) = \frac{IG(X, C)}{SI(X)} = \frac{H(C) - \sum_{x \in V_x} P(X = x) \cdot H(C \mid X = x)}{SI(X)} \qquad (4.4)$$

Η θέση του *SI(X)* στον παρονομαστή αποθαρρύνει την επιλογή χαρακτηριστικών με πολλές ισοκατανεμημένες σχετικά τιμές. Για ένα feature με N ισοκατανεμημένες τιμές, το *SI(X)* ισούται με $\log_2 N$, το οποίο αυξάνει με την αύξηση του N, ενώ μειώνεται καθώς μεγαλώνει η ανισοκατανομή εμφάνισης των τιμών. Μειονέκτημα του *GR* είναι ότι ο παρονομαστής τείνει στο μηδέν για features με πολύ ανισοκατανεμημένες τιμές.

4.B.II) Πειραματική σύγκριση μέτρων

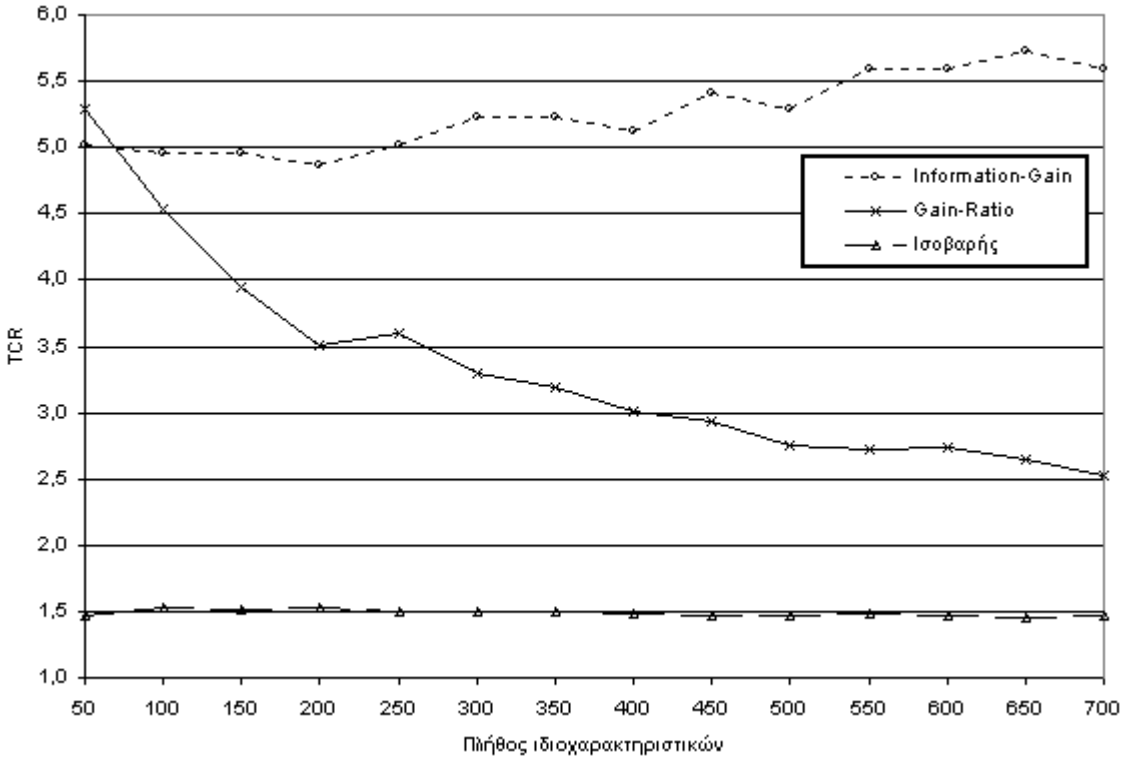
Στα περισσότερα διαγράμματα που θα παρουσιαστούν στη συνέχεια, κάθε διάγραμμα αφορά μία μόνο τιμή του λ (ένα σενάριο χρήσης) και οι καμπύλες σε αυτό δίνουν το *TCR* ως συνάρτηση της διαστασιμότητας. Κάθε σημείο μιας καμπύλης έχει προέλθει από *στρωματοποιημένη διασταυρωμένη επικύρωση* 10 σημείων (stratified 10-fold cross-validation), για την ενίσχυση της αξιοπιστίας των αποτελεσμάτων (βλ. ενότητα (2.C.III.b)). Η παράμετρος στην οποία διαφέρουν οι καμπύλες ενός διαγράμματος ποικίλει. Τα αποτελέσματα αυτής της ενότητας αντιπροσωπεύουν πειράματα χωρίς αποτίμηση γειτόνων με βάση την απόσταση (ή πιο σωστά με συνάρτηση αποτίμησης γειτόνων σταθερά). Στην επόμενη ενότητα θα διερευνηθεί η επίδραση και αυτής της παράμετρου.

Στα διαγράμματα (4-1) και (4-2) παρουσιάζονται οι καμπύλες για λ=1 και λ=999, αντίστοιχα, του 10-NN (k=10) για τα τρία μέτρα αποτίμησης των features που δοκιμάστηκαν*. Οι καμπύλες για λ=9 ακολουθούν την ίδια μορφή με τις αντίστοιχες για λ=1 (με μία κατακόρυφη μετατόπιση προς τα κάτω) και παραλείπονται χάριν συντομίας.

* Ένα ακόμα μέτρο για το οποίο έγιναν κάποια δειγματοληπτικά πειράματα ήταν το χ^2, μέτρο που βασίζεται στην ομώνυμη στατιστική κατανομή. Ωστόσο τα αποτελέσματά ήταν παρόμοια με του *IG* και γι'αυτό δεν κρίθηκε απαραίτητη η μελέτη του.

Για λ=1 (και λ=9), το συμπέρασμα είναι πως το *IG* είναι ανώτερο του *GR* και το τελευταίο με τη σειρά του ξεπερνάει την ισοβαρή αποτίμηση (equal weights – συντ. *EW*), ανεξαρτήτως της διαστασιμότητας. Αν και στα διαγράμματα εικονίζεται μόνο ο 10-NN, το ίδιο συμπέρασμα βγαίνει σε γενικές γραμμές και για τους 1-NN και 2-NN που επίσης δοκιμάστηκαν. Συγκρίνοντας το διάγραμμα (4-1) με το (3-1), διαπιστώνουμε ότι ο 10-NN *IG* είναι αποτελεσματικότερος και από τον NB, με εξαίρεση στα 100 features, ενώ ο NB γενικά ξεπερνάει τον 10-NN *GR* και τον 10-NN *EW*.

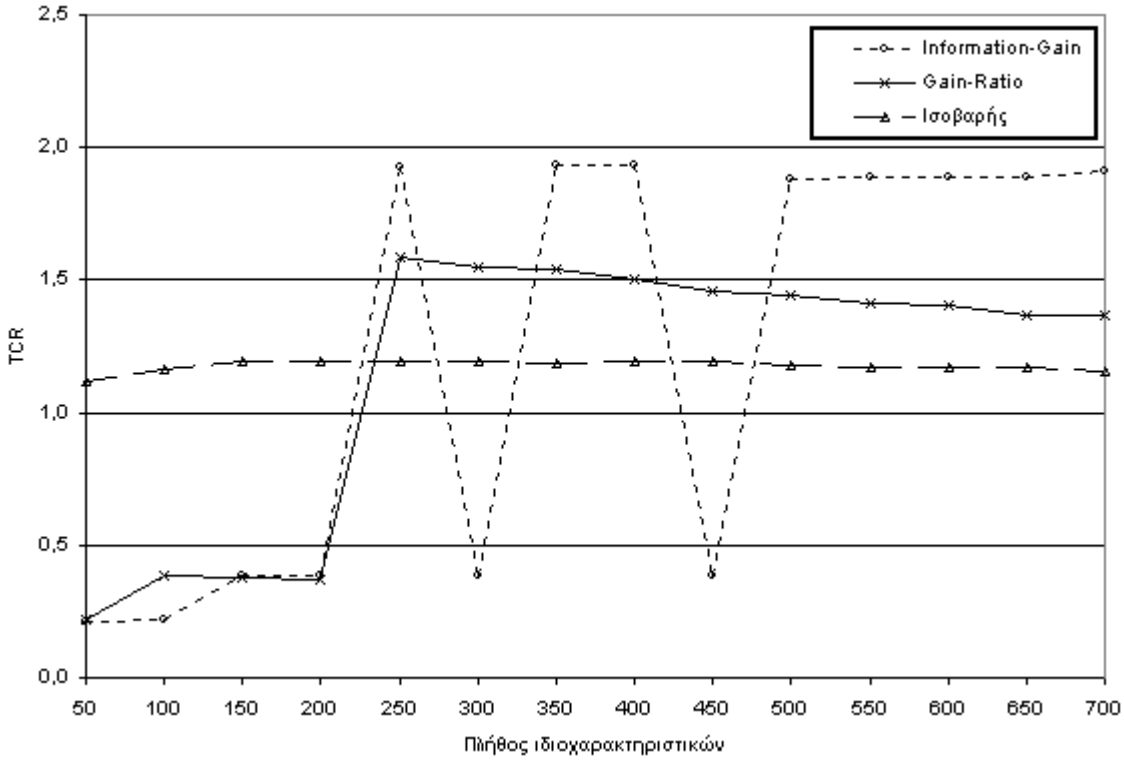
Χαρακτηριστική είναι επίσης η εξάρτηση κάθε μέτρου από τη διαστασιμότητα. Το *IG* βελτιώνεται ελαφρώς, το *GR* έχει έντονα πτωτική πορεία και το EW παραμένει σχεδόν ανεπηρέαστο, καθώς προστίθενται παραπάνω features στην αναπαράσταση.



Διάγραμμα 4-1: *TCR* για λ=1 (t=0.5) του 10-NN
για τρεις μέθοδους αποτίμησης χαρακτηριστικών (feature weighting).

Για λ=999, η εικόνα αλλάζει ριζικά. Κύριο χαρακτηριστικό σε αυτό το σενάριο, το οποίο παρατηρείται για τους περισσότερους συνδυασμούς παραμέτρων, είναι η απότομη μεταβολή του *TCR*, από αποδόσεις άνω της βάσης (=1) σε κάτω της βάσης και αντίστροφα, για διαφορετικές διαστασιμότητες. Η αστάθεια αυτή είναι αναμενόμενη, αν ληφθεί υπόψη η κυρίαρχη επίδραση των λαθών $L \to S$ στο *TCR* για τόσο μεγάλο λ. Όπως σημειώθηκε και στο κεφάλαιο 3 για να ξεπεράσει τη βάση ένα φίλτρο όταν λ=999 πρέπει να μην κάνει κανένα λάθος $L \to S$. Η μετάβαση από ένα, συνήθως, τέτοιο λάθος σε κανένα και αντίστροφα για διαφορετικές διαστασιμότητες είναι υπεύθυνη για την αστάθεια των καμπυλών. Το ζητούμενο λοιπόν εδώ δεν είναι να εντοπιστεί ο συνδυασμός παραμέτρων που μεγιστοποιεί το *TCR* για τα συγκεκριμένα σύνολα εκπαίδευσης και ελέγχου, αλλά να βρεθεί ένας *αξιόπιστος* συνδυασμός με όσο γίνεται περισσότερο ικανοποιητική απόδοση, υπό την έννοια

πως θα εγγυάται σε μεγάλο βαθμό τη διατήρηση της απόδοσης πάνω από τη βάση και για διαφορετικά σύνολα δεδομένων.



Διάγραμμα 4-2: *TCR* για λ=999 (t=0.999) του 10-NN
για τρεις μέθοδους αποτίμησης χαρακτηριστικών (feature weighting).

Με βάση τον παραπάνω στόχο, το πιο αξιόπιστο από τα τρία μέτρα είναι το *EW*, αφού παραμένει διαρκώς πάνω από τη βάση. Ωστόσο η spam ανάκλησή του (*SR*) δεν ξεπερνάει το 17%, δηλαδή λιγότερα από το ένα πέμπτο των spam μπλοκάρονται από το φίλτρο, απόδοση που δεν είναι αρκετή για πρακτική εφαρμογή. Δεύτερο πιο αξιόπιστο είναι το *GR* για περισσότερα από 250 features, όπου επίσης μένει πάνω από τη βάση. Η ανάκληση του *GR* φτάνει κοντά στο 37% (με 100% ορθότητα), υπερδιπλάσια του *EW*, αλλά επίσης μη ικανοποιητική. Τέλος, το *IG* φαίνεται το λιγότερο αξιόπιστο από τα τρία, αφού μόνο για 500 features και πάνω δείχνει να σταθεροποιείται άνω της βάσης, και αυτό με αρκετή επιφύλαξη δεδομένης της αστάθειας του μεταξύ 200 και 500 features. Πάντως η ανάκληση του βρίσκεται γύρω στο 47%, με ένα στα δύο περίπου spam να μπλοκάρονται, επίδοση σαφώς ικανοποιητικότερη από του *EW*.

Αντίθετα από τον 10-NN, οι επιδόσεις των 1-NN και 2-NN είναι γενικά κάτω της βάσης για όλα τα μέτρα. Μία πρώτη γενική εντύπωση από τα στοιχεία αυτά είναι πως προϋπόθεση για αξιοπρεπή επίδοση σε αυτό το σενάριο είναι η μεγάλη διαστασιμότητα και η μεγάλη θεωρούμενη γειτονία (π.χ. k=10). Αυτή η εντύπωση ενισχύθηκε στα πειράματα που ακολούθησαν και θα παρουσιαστούν στη συνέχεια.

4.Β.ΙΙΙ) Θεωρητική διερεύνηση

Σε αυτή την ενότητα θα επιχειρηθούν κάποιες ερμηνείες πάνω στη συμπεριφορά των τριών μεθόδων αποτίμησης των features και της αλληλεπίδρασής τους με το λ, το k και τη διαστασιμότητα. Αν και έγινε προσπάθεια να στηριχθεί η επιχειρηματολογία σε θεωρητικές βάσεις, αρκετές προτάσεις αποτελούν υποθέσεις που ισχύουν διαισθητικά και κατά προσέγγιση, επιβεβαιώνονται όμως σε μεγάλο βαθμό από πειραματικά δεδομένα.

4.Β.ΙΙΙ.α) Ισοβαρής αποτίμηση (*EW*)

Τα πρώτα πειράματα που έγιναν χρονικά ήταν με ισοβαρή αποτίμηση (*EW*), δημιουργώντας αίσθηση με τα χαμηλότερα του προσδοκώμενου αποτελέσματα. Ο k-NN είναι από τους πλέον επιτυχημένους αλγόριθμους και υπήρχε εκ των προτέρων η πεποίθηση πως θα αναδεικνύονταν καλύτερος από τον NB, και σίγουρα όχι τόσο κακός για μεγάλο k.

Μία παρατήρηση που έγινε ως προς την επιλογή των γειτόνων μέσω *EW* κατέδειξε την αιτία της χαμηλής απόδοσης. Αυτή είναι ο πάρα πολύ μεγάλος αριθμός ισοπαλιών μεταξύ των γειτόνων που συμβαίνουν, με συνέπεια το k να μην έχει σχέση με το πλήθος των γειτόνων που πραγματικά λαμβάνονται υπόψη. Π.χ., για k=10 και στα 700 features, τυπικά μεγέθη γειτονιών είναι 100-200, ενώ για το ίδιο k στα 50 features ξεπερνούν και τους 2000 (!!), με όλα τα στιγμιότυπα εκπαίδευσης να είναι 2604. Με δεδομένη αυτή την παρατήρηση, δεν είναι περίεργη η χαμηλή απόδοση, η οποία οφείλεται ουσιαστικά στην πολύ χαμηλή ανάκληση. Απ' τη στιγμή που η πλειοψηφούσα κλάση είναι αυτή των θεμιτών μηνυμάτων, είναι εύλογο πως θα υπερισχύει σε γειτονίες των εκατοντάδων και των χιλιάδων γειτόνων, αφού όλα κι όλα τα spam στο σύνολο εκπαίδευσης είναι γύρω στα 433. Έτσι τα περισσότερα άγνωστα μηνύματα, ακόμα κι αν είναι spam, θα έχουν σε τόσο ευρείες γειτονίες περισσότερους μη spam γείτονες και επομένως θα ταξινομούνται λανθασμένα.

Το επόμενο ερώτημα είναι βέβαια τι ευθύνεται για τις πολλές ισοπαλίες. Η απάντηση είναι πως ευθύνεται η ισοβαρής αποτίμηση. Αυτό ίσως προκαλεί έκπληξη σε όσους είναι εξοικειωμένοι με τον k-NN, δεδομένου πως ο αλγόριθμος εφαρμόζεται συνήθως με ισοβαρή αποτίμηση χωρίς πρόβλημα. Παραδοσιακά όμως ο k-NN εφαρμόζεται κυρίως σε προβλήματα με συνεχή αριθμητικά features, όπου είναι απίθανο να ισαπέχουν δύο αυθαίρετα στιγμιότυπα από ένα άγνωστο. Εδώ όμως τα features είναι συμβολικά και μάλιστα δυαδικά, οπότε η απόσταση δύο στιγμιοτύπων, όπως υπολογίζεται από τον τύπο (2.11), εκφράζει το πλήθος των features στα οποία διαφέρουν τα στιγμιότυπα. Δεν είναι καθόλου σπάνιο να υπάρχουν πολλά στιγμιότυπα που να διαφέρουν κατά το ίδιο πλήθος από features (όχι απαραίτητα στα ίδια features) από ένα άγνωστο, με αποτέλεσμα όλα αυτά να θεωρούνται ισαπέχοντες γείτονες του τελευταίου.

Στα *IG*, *GR* και γενικά σε οποιαδήποτε συνεχή συνάρτηση αποτίμησης των features αυτό το πρόβλημα δεν υφίσταται, απ'τη στιγμή που κάθε feature αποτιμάται διαφορετικά. Έτσι, για να ισαπέχουν δύο στιγμιότυπα από ένα άγνωστο δεν αρκεί η ασθενής συνθήκη να διαφέρουν από αυτό κατά το ίδιο πλήθος από features. Στην πράξη, ισοπαλίες συμβαίνουν μόνο μεταξύ διανυσμάτων που ταυτίζονται. Στα πειράματα με τα *IG* και *GR*, το μέγεθος της γειτονιάς για ένα στιγμιότυπο σπάνια ξεπερνούσε το k κατά διψήφιο αριθμό στιγμιοτύπων, με

πιο συνήθη την υπέβαση κατά 1-4, ανάλογα και με τη διαστασιμότητα. Επακόλουθο των ελάχιστων υπερβάσεων είναι και το γεγονός πως οι προβλέψεις των *IG* και *GR* είναι σχεδόν ανεξάρτητες του λ για k=1[*]. Πράγματι, οι μόνες διαφοροποιήσεις οφείλονται στις λίγες ισοπαλίες. Για παράδειγμα, αν ισοβαθμήσουν τρία στιγμιότυπα ως κοντινότερα σε ένα άγνωστο, με το ένα να είναι θεμιτό και τα άλλα δύο spam, το άγνωστο θα καταταχθεί ως spam για λ=1 (αφού $2/1 > 1$), αλλά θεμιτό για λ=9 (αφού $2/1 < 9$). Αν όμως δεν υπάρχει ισοβαθμία, το στιγμιότυπο θα καταταχθεί στην ίδια κατηγορία με το μοναδικό γείτονα που θα θεωρήσει, ανεξαρτήτως του λ. Με την ίδια συλλογιστική, για k=3 οι προβλέψεις είναι σχεδόν ίδιες για λ=9 και 999, αλλά διαφέρουν για λ=1.

4.B.III.b) Σύγκριση των μέτρων *IG – GR – EW*

Το κύριο ερώτημα εδώ είναι γιατί διαφέρει τόσο η απόδοση του *IG* από το *GR*, ανεξαρτήτως σχεδόν του λ, του k και της διαστασιμότητας. Το *GR* είναι μια κανονικοποιημένη μορφή του *IG*. Δεν θα περίμενε κανείς να έχουν τέτοια διαφορά.

Μια πρώτη παρατήρηση που μπορεί να γίνει είναι πως από το *GR* δε θα αναμενόταν αισθητή βελτίωση σε σχέση με το *IG* για τη συγκεκριμένη αναπαράσταση των μηνυμάτων που επιλέχθηκε. Όπως αναφέρθηκε στην ενότητα (4.B.I), το *GR* προτάθηκε για να αντιμετωπίσει το πρόβλημα της υπερεκτίμησης από το *IG* χαρακτηριστικών με πολλές και σχετικά ισοκατανεμημένες τιμές. Αυτό το πρόβλημα δεν υφίσταται εδώ, λόγω της δυαδικής αναπαράστασης που χρησιμοποιήθηκε. Πέραν του ότι υπάρχουν δύο μόνο δυνατές τιμές για κάθε feature (που υποδηλώνουν την παρουσία ή απουσία μιας λέξης σε ένα μήνυμα), αυτές είναι και έντονα ανισοκατανεμημένες: Η απουσία μιας λέξης είναι κατά κανόνα πολύ πιο συχνή από την παρουσία της. Έτσι, το *GR* δεν φαίνεται καταρχήν να υπόσχεται εδώ κάτι καλύτερο από το *IG*.

Τα αποτελέσματα δείχνουν πως το *GR* όχι μόνο δεν βελτιώνει την απόδοση, αλλά την υποβαθμίζει σημαντικά. Παραπάνω μετρήσεις που έγιναν έδειξαν πολλές και μεγάλες αποκλίσεις στη σειρά κατάταξης των features για τις δύο μεθόδους (π.χ. λέξεις που αποτιμώνται στις 10 καλύτερες από το *IG* συχνά βρίσκονται στη δεύτερη εκατοντάδα για το *GR*, και αντίστροφα). Αυτό δε θα μπορούσε παρά να έχει σημαντικότατες διαφορές στην απόδοση του ταξινομητή, δεν αρκεί όμως για να δικαιολογήσει το γιατί τελικά είναι το *IG* αυτή που κάνει πιο σωστή αποτίμηση από το *GR* και δε συμβαίνει το αντίστροφο.

Μία πρώτη προσπάθεια ερμηνείας είναι η εξής: Το *GR* ευνοεί features με μικρή πληροφορία διαμερισμού *SI(X),* (ή απλά εντροπία), δηλαδή features που υπάρχει μικρή αβεβαιότητα ως προς την παρουσία τους. Οι λέξεις που αντιστοιχούν σε τέτοια features είναι είτε πολύ συχνές ή πολύ σπάνιες[†]. Η αξία όμως αυτών των λέξεων είναι πολύ αμφίβολη, επειδή ακριβώς είναι γενικά πολύ σπάνιες ή συνηθισμένες λέξεις, με αποτέλεσμα το *GR* να κάνει κακή αποτίμηση των λέξεων. Ο συλλογισμός αυτός δεν είναι πλήρης, διότι για πολύ

[*] Προσοχή: Η ανεξαρτησία των προβλέψεων από το λ δε συνεπάγεται και την ανεξαρτησία του *TCR* από αυτό, αφού η αξιολόγηση των σωστών και λανθασμένων προβλέψεων βασίζεται στο λ.

[†] Στην πράξη, μόνο πολύ σπάνιες, αφού δεν υπάρχουν "πολύ συχνές" λέξεις - λέξεις με συχνότητα της τάξης του 90-95%. Αντίθετα, υπάρχουν λέξεις με συχνότητα της τάξης του 0.1%

συχνά ή σπάνια features είναι δυνατόν να μειωθεί και ο αριθμητής του *GR*, το *IG*. Πράγματι, το τελευταίο συνυπολογίζει την πιθανότητα εμφάνισης κάθε τιμής του feature, τιμωρώντας αυστηρά την ύπαρξη συχνών τιμών με μεγάλη εντροπία (=αβεβαιότητα) της κατηγορίας δοθείσης της τιμής *H(C|X=x)*, όπως είναι η απουσία μιας λέξης από ένα μήνυμα. Έτσι, δεν είναι προφανές ότι οι πολύ σπάνιες λέξεις θα υπερεκτιμηθούν, αφού μειώνονται ταυτόγχρονα ο αριθμητής και ο παρονομαστής.

Μπορούμε να δείξουμε ποιοτικά για το συγκεκριμένο πρόβλημα πως συνήθως ο παρονομαστής του *GR* μειώνεται πιο γρήγορα από τον αριθμητή του καθώς η ανισοκατανομή των τιμών του feature αυξάνεται. Μεγαλύτερη ανισοκατανομή σημαίνει πως η συχνή τιμή γίνεται συχνότερη και η σπάνια σπανιότερη. Εφόσον στην κατηγοριοποίηση κειμένου η σπάνια τιμή για ένα δυαδικό feature που αντιστοιχεί σε κάποιον όρο είναι αυτή που δηλώνει την παρουσία του όρου, θα δείξουμε πως το *GR* ευνοεί τους σπάνιους όρους (για τους οποίους δηλαδή παρουσιάζεται μεγαλύτερη ανισοκατανομή τιμών).

Έστω πως η τιμή 0 για ένα feature δηλώνει την απουσία της αντίστοιχης λέξης και η τιμή 1 την παρουσία της. Ο τύπος (4.1) τότε απλοποιείται σε:

$$IG(X,C) = H(C) - \big(P(X=0)\cdot H(C \mid X=0) + P(X=1)\cdot H(C \mid X=1)\big) \qquad (4.5)$$

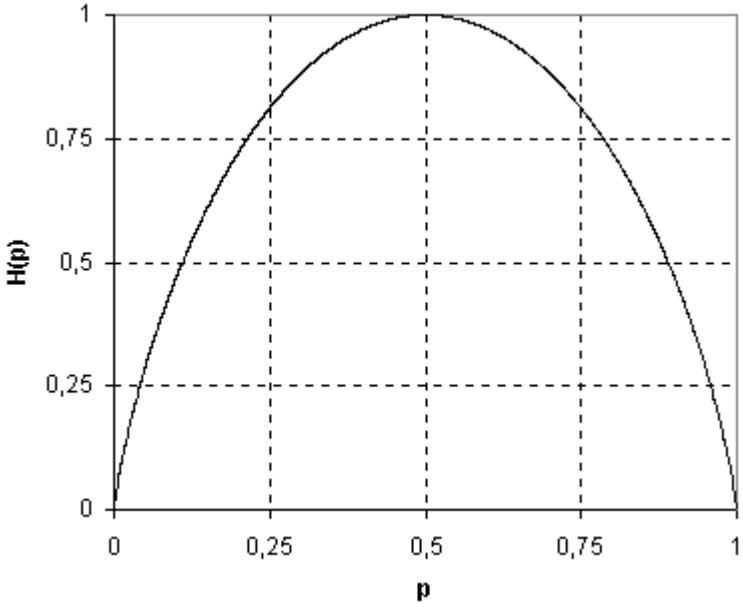
Έστωσαν επίσης δύο features ανάλογης αξίας (διαχωριστικής ικανότητας), το ένα με (σχετικά) σπάνια παρουσία (*rare*) και το άλλο με συχνότερη (*freq*). Ισχύει δηλαδή:

$$P(freq=1) > P(rare=1) \quad \text{και άρα} \quad P(rare=0) > P(freq=0) \qquad (4.6)$$

Εφόσον τα *rare* και *freq* είναι ανάλογης αξίας, ο όρος *H(C|X=0)* είναι περίπου ίσος για τα δύο features και σχετικά μεγάλος, κοντά στο *H(C)*, λόγω του ότι η *απουσία* μιας λέξης από ένα μήνυμα ελάχιστα συνηγορεί υπέρ κάποιας κλάσης. Αντίστοιχα, ο όρος *H(C|X=1)* είναι επίσης περίπου ίσος για τα δύο features και σχετικά μικρός, τόσο μικρότερος όσο πιο καλό διαχωριστικό είναι το feature, αφού η *παρουσία* του σε ένα mail συνηγορεί σημαντικά υπέρ της μίας από τις δύο κλάσεις. Κατά συνέπεια, όσο μεγαλύτερη είναι είναι η συχνότητα απουσίας *P(X=0)* από τη συχνότητα παρουσίας *P(X=1)*, τόσο μεγαλύτερο θα είναι και το άθροισμα στην (4.5) και επομένως θα μειώνεται το *IG*. Γι' αυτό και ο σπάνιος όρος *rare* θα έχει μικρότερο *IG* από το συχνότερο *freq*. Όσο για το *SI(X),* το οποίο εξ ορισμού είναι η εντροπία του X, μειώνεται με την αύξηση της ανισοκατανομής, άρα επίσης είναι μικρότερο για τον όρο *rare.*

Ωστόσο, από την (4.5) φαίνεται πως το *IG* μεταβάλλεται γραμμικά ως προς την *P(X=0)*, και συνεπώς μικρή διαφορά σε αυτήν για διαφορετικά features οδηγεί σε μικρή διαφορά στο αποδιδόμενο από το *IG* βάρος. Από την άλλη μεριά, το *SI(X)* δε μεταβάλλεται γραμμικά ως προς την *P(X=0)*, και μάλιστα, λόγω της υψηλής τιμής του τελευταίου (άνω του 90-95% συνήθως), μικρή μεταβολή σε αυτήν οδηγεί σε μεγάλη μεταβολή του *SI(X)*, όπως μπορεί να φανεί και στο διάγραμμα (4-3). Έτσι στο *GR* η επίδραση του παρονομαστή είναι εντονότερη από αυτήν του *IG*, με τελικό αποτέλεσμα να υπερτονίζονται τα features που αντιστοιχούν σε σπάνιες λέξεις (τουλάχιστον αν εφαρμόζεται δυαδική αναπαράσταση). Η προδιάθεση αυτή μειώνει την ικανότητά του *GR* για καλή αποτίμηση. Αντίθετα το *IG*, αν και το ίδιο δεν είναι αμερόληπτο με την κλίση που έχει για συχνές λέξεις, η κλίση του αυτή είναι πιο ήπια από του

GR, πετυχαίνοντας σωστότερη αποτίμηση. Παρ' όλα αυτά, και οι δύο μέθοδοι δίνουν καλύτερα αποτελέσματα από την ισοβαρή αποτίμηση (*EW*), αφού δεν αντιμετωπίζουν το πρόβλημα των πολλών ισοπαλιών που υποβιβάζει την απόδοση της τελευταίας.



Διάγραμμα 4-3:
Η καμπύλη εντροπίας για ένα δυαδικό feature ως συνάρτηση της πιθανότητας εμφάνισης μίας εκ των δύο τιμών της.

4.B.III.c) Επίδραση της διαστασιμότητας

Η επιδείνωση του *GR* κατά την αύξηση της διαστασιμότητας δεν είναι δύσκολο να αιτιολογηθεί μετά από την παραπάνω ανάλυση. Αυξάνοντας τη διαστασιμότητα, αυξάνονται και τα περιθώρια λάθους του *GR*, καθώς έχει διαθέσιμη μεγαλύτερη "δεξαμενή" από σπάνια features για να υπερεκτιμήσει. Λαμβάνοντας υπόψη πως η επιλογή των N καλύτερων όρων (feature selection) έγινε με το *IG* κι επομένως η κατάταξη τους είναι σχετικά καλή, ο περιορισμός του *GR* σε μικρά πλήθη από, εν γένει καλά, features το αναγκάζει να περιορίσει την κακή γενικά αποτίμηση που κάνει μόνο σε αυτά, και όχι άλλα, πιθανότατα χειρότερα features που δεν επιλέχτηκαν.

Σχετικά τώρα με την ελαφρά ανοδική πορεία των καμπυλών με *IG*, αυτή οφείλεται στο ότι επιτρέπει στον *k*-NN να σταθμίζει καλά την αξία των καλύτερων features σε σχέση με την αξία των χειρότερων που προστίθενται όσο μεγαλώνει η διαστασιμότητα. Τα "χειρότερα" features δεν είναι όλα άχρηστα, απλά πρέπει να τους αποδοθεί μικρότερη σημασία σε σχέση με τα καλύτερα. Το *IG* φαίνεται να αποδίδει γενικά σωστή σημασία και καταφέρνει έτσι να επωφεληθεί και από τα χειρότερα features. Ένα πλεονέκτημα που δημιουργεί η εύρεση ενός καλού μέτρου αποτίμησης είναι πως εξαλείφει σε ένα βαθμό την ανάγκη για επιλογή των χαρακτηριστικών, αφού οσοδήποτε κακό και να είναι ένα feature μπορεί να μετέχει στην αναπαράσταση μεν, με ελάχιστο βάρος δε. Στην πράξη όμως η επιλογή των features πρέπει να γίνει λόγω του υψηλού κόστους σε χρόνο και μνήμη που συνεπάγεται η μεγάλη διαστασιμότητα.

Τέλος, η στασιμότητα των καμπυλών για ισοβαρή αποτίμηση, κυρίως για μεγάλα *k*, όπως και η χαμηλή απόδοση, οφείλεται στους πολλούς υπεράριθμους γείτονες. Ένα στιγμιότυπο μπορεί να απέχει από ένα άλλο από 0 έως τη διαστασιμότητα, και επομένως οι ισοπαλίες

είναι λιγότερο πιθανές για μεγάλες διαστασιμότητες. Έτσι, αν και το πλήθος των γειτόνων μειώνεται μεν με την αύξηση της διαστασιμότητας, δε μειώνεται αρκετά – από το επίπεδο των χιλιάδων πέφτει στις εκατοντάδες. Η γειτονιά παραμένει έτσι πολύ μεγάλη, με την καθολικά συχνότερη κλάση (εν προκειμένω των θεμιτών μηνυμάτων) να υπερισχύει τις περισσότερες φορές.

4.B.IV) Επίδραση της παραμέτρου *k*

Όπως αναφέρθηκε και στην εισαγωγή της ενότητας (4.B), η πειραματική σύγκριση των μεθόδων αποτίμησης των features στηρίζει αρκετά την υπόθεση πως η συγκεκριμένη παράμετρος είναι σε μεγάλο βαθμό ανεξάρτητη από τις υπόλοιπες. Έτσι τα υπόλοιπα πειράματα πραγματοποιήθηκαν μόνο για την καλύτερη μέθοδο (εν προκειμένω την *IG*), μειώνοντας τη διάσταση παραμέτρων του προβλήματος βελτιστοποίησης κατά ένα.

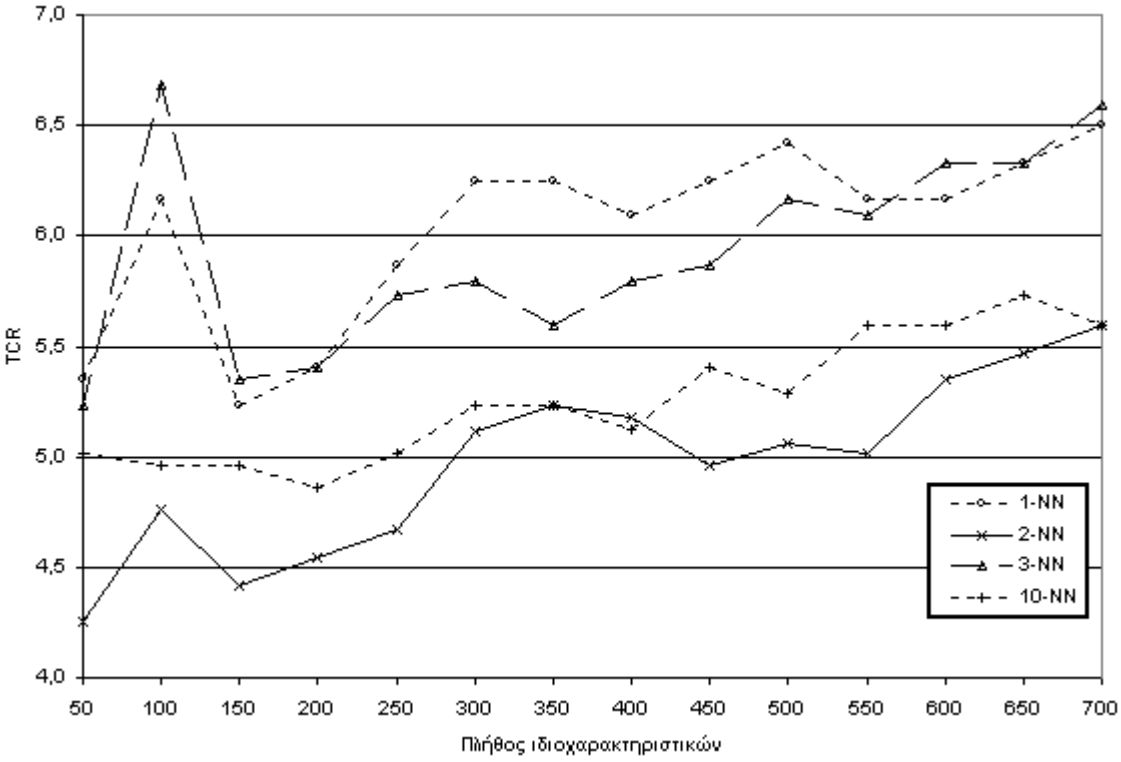
Η επόμενη παράμετρος που έπρεπε να διερευνηθεί ως προς την επίδρασή της στην απόδοση είναι το *k*. Δυστυχώς σε αυτή την περίπτωση τα αποτελέσματα δεν ήταν τόσο ξεκάθαρα όσο με τη διερεύνηση της αποτίμησης των features. Το *k*, πέραν από τη μεγάλη σημασία που έχει η επιλογή του για διαφορετικά σενάρια και η οποία εξηγείται επαρκώς διαισθητικά, εξαρτάται και από τη διαστασιμότητα κατά τρόπο μη ερμηνεύσιμο. Αυτό που παρατηρήθηκε ήταν πως για δεδομένο λ δεν υπάρχει μια απόλυτη σειρά κατάταξης των *k*, ανεξάρτητη της διαστασιμότητας, δηλαδή για αναπαραστάσεις διαφορετικής διαστασιμότητας αλλάζει και η συγκριτική απόδοση των διαφόρων *k*. Το χειρότερο για την ερμηνεία αυτής της εικόνας είναι πως η εξάρτηση από τη διαστασιμότητα γίνεται συχνά κατά ασυνεχή τρόπο, εννοώντας πως για ένα διάστημα τιμών διαστασιμότητας (π.χ. 50-250 features) μία τιμή του *k* υπερταιρεί μιας άλλης, για ένα επόμενο διάστημα (π.χ. 300-400) προηγείται η δεύτερη, για ένα τρίτο διάστημα (π.χ. 450-600) προηγείται πάλι η πρώτη, κ.ο.κ. Δεν μπορεί επομένως να επιχειρηθεί μία, διαισθητική έστω, εξήγηση για το φαινόμενο, η οποία ενδεχομένως να ήταν δυνατό να δοθεί αν η μία τιμή προηγούνταν μόνο για "μικρές" διαστασιμότητες (από 50 έως N) και η άλλη για "μεγάλες" (από N+50 έως 700). Η εντύπωση που δίνεται πάντως είναι πως αυτές οι ανακατατάξεις δεν έχουν κάποια θεμελιώδη συσχέτιση με το *k*.

Στα διαγράμματα (4-4), (4-5) και (4-6) δίνονται κάποιες χαρακτηριστικές καμπύλες για διάφορες τιμές του *k*, για λ=1, 9 και 999, αντίστοιχα. Οι καμπύλες που επιλέχθηκαν για παρουσίαση είναι οι πλέον ενδεικτικές και, στο βαθμό που ήταν δυνατόν, οι λιγότερο επικαλυπτόμενες μεταξύ τους. Πειράματα ωστόσο έγιναν για όλα τα *k* από 1 έως και 10, για κάθε λ και διαστασιμότητα, και η γενική τους εικόνα θα περιγραφεί στο κείμενο. Ομοίως, οι παρατηρήσεις και τα πορίσματα που ακολουθούν αφορούν το σύνολο των πειραμάτων και όχι μόνο τα παρισταθέντα στα διαγράμματα.

Στο διάγραμμα (4-4) για λ =1 παριστάνεται το *TCR* για τέσσερις τιμές του *k*. Το πρόβλημα της εξάρτησης από τη διαστασιμότητα κάνει αισθητή την παρουσία του, αν και σε πολύ μικρότερο βαθμό απ'ότι αν παρουσιάζονταν και οι δέκα καμπύλες για καθένα από τα θεωρούμενα *k*. Ο 1-NN είναι για τις περισσότερες διαστασιμότητες καλύτερος, με την

καθολικά βέλτιστη τιμή ωστόσο να είναι του 3-NN για 100 features. Ο 3-NN μπορεί να θεωρηθεί έτσι δεύτερος, θέση που διατηρεί προσεγγιστικά και για το σύνολο των δέκα καμπυλών. Ακολουθούν ο 5-NN (με έντονη επικάλυψη με τον 3-NN), ο 7-NN και ο 9-NN, οι οποίοι δεν παριστάνονται στο διάγραμμα. Στη συνέχεια ο 10-NN, ο 8-NN και ο 6-NN δείχνουν να προηγούνται, με συνεχείς ανακατατάξεις για διαφορετικές διαστασιμότητες. Τελευταίοι έρχονται ο 4-NN (ο οποίος όμως είναι καθολικά πρώτος για 50 features !) και ο 2-NN.

Με μια πρόχειρη ματιά στο διάγραμμα, η επίδοση των διαφόρων καμπυλών φαίνεται τυχαία. Η (προσεγγιστική) σειρά κατάταξης των καμπυλών όμως δίνει ένα πρώτο ερέθισμα: Οι περιττού *k* *k*-NN προηγούνται, εν γένει, από αυτούς άρτιου *k*. Επίσης, οι μικροί περιττοί *k* είναι καλύτεροι των μεγαλύτερων, ενώ το αντίστροφο περίπου ισχύει για τους άρτιους, όπου ο 10-NN είναι από τους κορυφαίους και ο 2-NN ο χειρότερος.



Διάγραμμα 4-4: *TCR* του *k*-NN για λ=1 (t=0.5) για διάφορα *k*
(με Information-Gain αποτίμηση χαρακτηριστικών).

Αν και αυτές οι παρατηρήσεις μοιάζουν να περιπλέκουν τα πράγματα αντί να τα ξεδιαλύνουν, η ερμηνεία τους είναι άμεση. Όπως έχει ήδη αναφερθεί, ο *k*-NN με *IG* θεωρεί ελάχιστους υπεράριθμους γείτονες, αφού στην πράξη δύο στιγμιότυπα πρέπει να ταυτίζονται για να ισαπέχουν από ένα άλλο άγνωστο. Κατά συνέπεια, τις περισσότερες φορές η γειτονιά αποτελείται από ακριβώς *k* γείτονες. Αν το *k* είναι άρτιο, είναι δυνατόν να υπάρξουν ισοπαλίες[*], με τους *k*/2 γείτονες να είναι spam και τους υπόλοιπους *k*/2 θεμιτά μηνύματα.

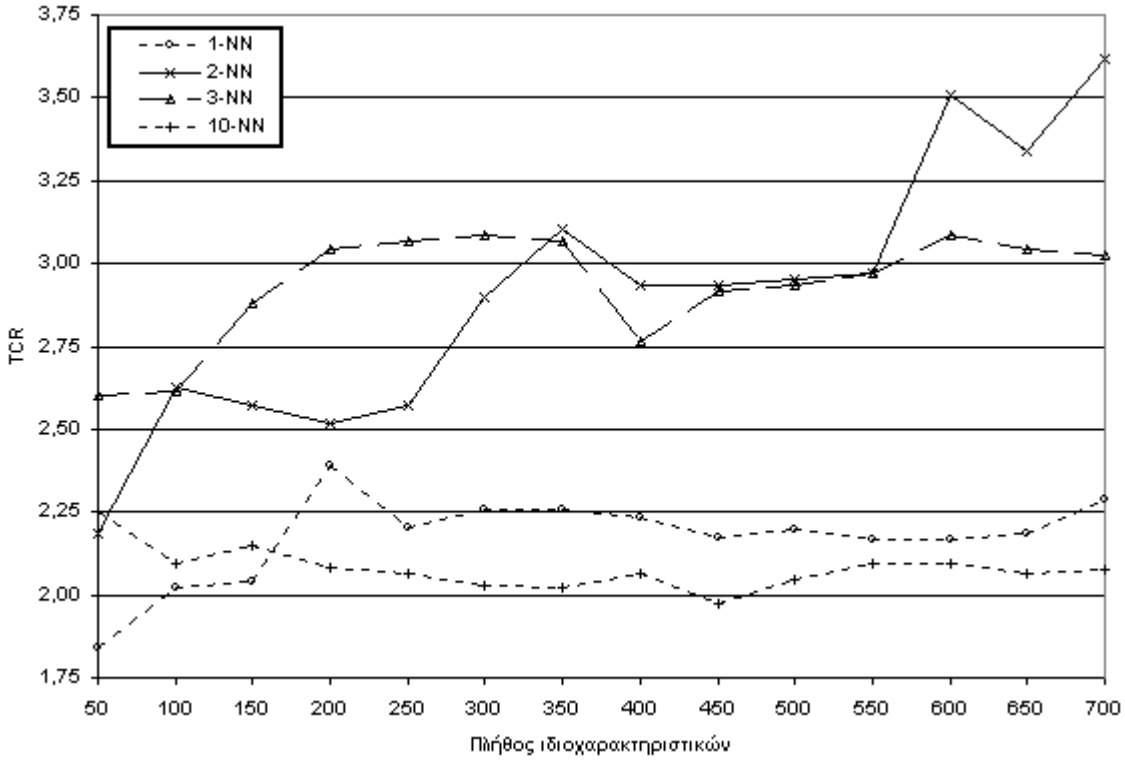
[*] Δεν πρέπει να υπάρξει σύγχυση ανάμεσα στις ισοπαλίες μεταξύ των κλάσεων μέσα σε μια κατανομή γειτόνων, όπως είναι το νόημα εδώ, και των ισοπαλιών μεταξύ στιγμιοτύπων ως προς την απόσταση τους από ένα τρίτο στιγμιότυπο που αναφέρθηκε στην ενότητα (4.B.III.a)

Γνωρίζοντας πως σε περίπτωση ισοπαλίας το μήνυμα κατατάσσεται πάντα ως θεμιτό, είναι σαφές πως κάποια spam θα περάσουν το φίλτρο με αυτό τον τρόπο. Τα πειραματικά δεδομένα στηρίζουν αυτή την εξήγηση, καθώς η spam ανάκληση (*SR*) για άρτια *k* είναι γενικά χαμηλότερη από αυτή για περιττά, κάτι που δε συμβαίνει για τη spam ορθότητα (*SP*).

Η βελτίωση της απόδοσης για μικρά περιττά *k* σε σχέση με μεγαλύτερα εξηγείται πιθανότατα από τη μεγάλη ανισοκατανομή των δύο κατηγοριών. Διαισθητικά, όσο μεγαλώνει το *k*, μεγαλώνει και η πιθανότητα η πλειοψηφία της γειτονιάς να είναι θεμιτά μηνύματα και άρα μεγαλώνει η τάση του ταξινομητή να κατατάσσει τα πάντα ως θεμιτά. Αυτή η τάση οδηγεί τον ταξινομητή σε περισσότερες αποτυχίες $S \to L$ και σε λιγότερες $L \to S$, με τις πρώτες όμως να υπερισχύουν των δεύτερων. Ο λόγος είναι πως όσο μεγαλώνει το *k*, μια πλειοψηφία θεμιτών γειτόνων ενός spam είναι πιο πιθανή από την πλειοψηφία spam γειτόνων ενός θεμιτού μηνύματος, εξαιτίας ακριβώς του πολλαπλάσιου συνολικού πληθυσμού των θεμιτών έναντι των spam. Και αυτό το επιχείρημα υποστηρίζεται από τα πειράματα, καθώς η *SR* ανεβαίνει και η *SP* πέφτει με τη μείωση του *k*, με την άνοδο της *SR* να υπερβαίνει την πτώση της *SP*.

Το επιχείρημα της τάσης κατάταξης των μηνυμάτων από τον *k*-NN ως θεμιτών με την αύξηση του *k* ισχύει φυσικά και για άρτιες τιμές του *k*. Εδώ όμως υπάρχει και μια αντίρροπη τάση λόγω των περισσότερων ισοπαλιών που κατά κανόνα παρατηρούνται σε σχέση με περιττές τιμές. Η αύξηση του *k* μειώνει γενικά τη συχνότητα των ισοπαλιών, και, όπως έχει αναφερθεί ήδη, στις ισοπαλίες τα μηνύματα κατατάσσονται ως spam. Το φαινόμενο είναι ιδιαίτερα έντονο για *k*=2, λόγω του ότι τότε η ισοπαλία αντιπροσωπεύει το ένα από τα τρία δυνατά ενδεχόμενα (αν εξαιρεθούν οι σχετικά σπάνιοι υπεράριθμοι γείτονες): Ή και οι δύο γείτονες να είναι spam, ή και οι δύο να είναι θεμιτοί, ή να είναι από ένας σε κάθε κλάση. Αντίστοιχα λ.χ. για *k*=10, οι περιπτώσεις ισοπαλίας είναι μεν περισσότερες (όσες και οι συνδυασμοί των 10 ανά 5), έχουν όμως συνήθως αθροιστικά μικρότερη πιθανότητα εμφάνισης. Η τελευταία πρόταση δεν έχει τεκμηριωθεί άμεσα με μετρήσεις και δεν ισχύει πάντοτε, αλλά είναι σύμφωνη με την υπόθεση πως απομακρυνόμενοι από ένα spam στιγμιότυπο αυξάνονται και οι μη spam γείτονές του. Μεταξύ των δύο αντίρροπων τάσεων κατάταξης δεν υπάρχει καθολικά νικητής, και εκεί οφείλεται η μεγάλη συνάφεια των καμπυλών με άρτιο *k* για διαφορετικές διαστασιμότητες .

Το διάγραμμα (4-5) παρουσιάζει τις καμπύλες για τις ίδιες τιμές του *k* για *λ*=9. Η εικόνα εδώ διαφέρει κατά το ότι οι κορυφαίες επιδόσεις σημειώνονται για τους 2-NN και 3-NN, χωρίς να ξεχωρίζει ένας από τους δύο ανεξαρτήτως της διαστασιμότητας. Η φαινομενικά "τυχαία" ανακατάταξη των καμπυλών για διαφορετικές διαστασιμότητες που εμφανίζεται για λ=1 κάνει έτσι κι εδώ την παρουσία της. Μεγαλύτερες τιμές του *k* υποβαθμίζουν σταδιακά την απόδοση. Τέλος ο 1-NN είναι ο χειρότερος μέχρι τα 150 features, αλλά ανακάμπτει ελαφρώς από τα 200 και πάνω, φτάνοντας τα επίπεδα των 7-NN και 8-NN.



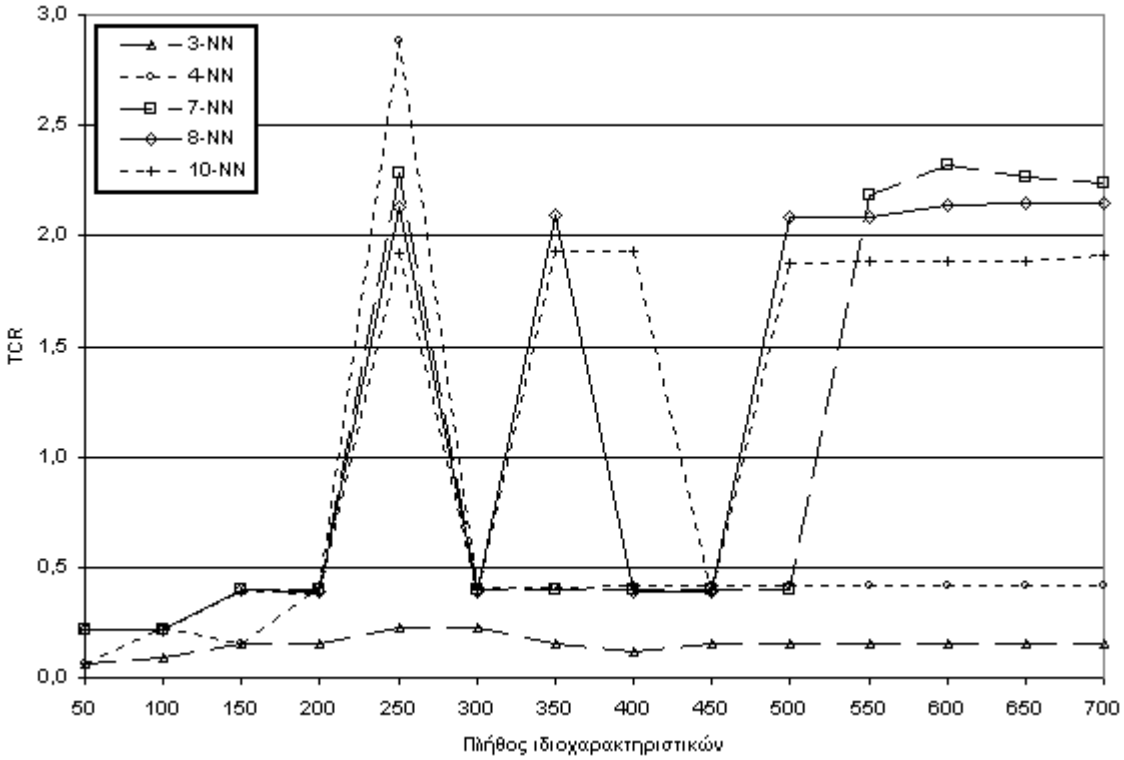
Διάγραμμα 4-5: *TCR* του *k*-NN για λ=9 (t=0.9) για διάφορα *k* (με Information-Gain αποτίμηση χαρακτηριστικών).

Αρχικά ίσως προβληματίζει το γεγονός πως ο 2-NN, από τελευταίος για λ=1 είναι πρώτος (μαζί με τον 3-NN) για λ=9 και ο 1-NN, από πρώτος για λ=1 καταλήγει από τους χειρότερους για λ=9. Ωστόσο και τα δύο αποτελέσματα ερμηνεύονται εύκολα αν ληφθεί υπόψη η πολιτική κατηγοριοποίησης που εφαρμόζεται για κάθε λ, μαζί με τα ίδια επιχειρήματα που χρησιμοποίηθηκαν και για λ=1. Όπως τονίστηκε στην ενότητα (4.B.III.a), οι προβλέψεις του 1-NN είναι σχεδόν ανεξάρτητες του λ, λόγω των ελάχιστων υπεράριθμων γειτόνων που θεωρούνται όταν γίνεται αποτίμηση μέσω του *IG*. Το αποτέλεσμα είναι να μην δείχνει μεγαλύτερη επιφυλακτικότητα στο να κατατάσσει ένα μήνυμα ως spam για λ=9, αντίθετα με τους *k*-NN για *k*>1. Γι'αυτό και η *SP* του, από μέτρια που ήταν (συγκριτικά με τα άλλα *k*) για λ=1 καταλήγει τελευταία και με διαφορά από τους άλλους *k*-NN για λ=9. Αν και από την άλλη πλευρά έχει την υψηλότερη *SR*, άρα και τις λιγότερες $S \to L$ αστοχίες, το *TCR* για λ=9 δίνει μεγαλύτερο βάρος στις αστοχίες $L \to S$, οι οποίες είναι περισσότερες για τον 1-NN.

Προηγουμένως υποστηρίχθηκε πως με την αύξηση του *k* ελαττώνονται οι $L \to S$ και αυξάνονται οι $S \to L$ αποτυχίες. Δεδομένου πως το *TCR* για λ=9 τιμωρεί αυστηρότερα τις πρώτες, θα αναμενόταν πως οι υψηλότερες αποδόσεις παρουσιάζονται για μεγάλο *k*. Το γεγονός πως το βέλτιστο μέγεθος γειτονίας είναι μεταξύ του 2 και 3 οφείλεται στο ότι, ακόμα και για μικρά *k*, τα λάθη $L \to S$ είναι αρκετά σπάνια (η *SP* είναι της τάξης του +97% για *k*>=2) και κατά συνέπεια τα περιθώρια περαιτέρω βελτίωσης είναι αρκετά μικρά. Αυτό που συμβαίνει λοιπόν είναι πως για *k*>3, η ελάχιστη ή ανύπαρκτη βελτίωση στην *SP*, ακόμα και ενισχυμένη από το λ, δεν αρκεί για να απορροφήσει την πτώση της *SR*, με αποτέλεσμα την σταδιακή υποχώρηση του *TCR* καθώς μεγαλώνει η γειτονιά.

Αξίζει επίσης να παρατηρηθεί πως για λ=9 (όπως και για 999) δεν έχουν αντίκτυπο οι ισοπαλίες στην απόδοση των k-NN με άρτιο k, διότι στην ουσία το πλήθος των θεμιτών γειτόνων πολλαπλασιάζεται με λ. Έτσι λ.χ., αν ο 2-NN προσδιορίσει για ένα μήνυμα ένα spam γείτονα και έναν θεμιτό, αυτό δεν είναι πλέον ισοπαλία αλλά σαφής απόφαση κατάταξης του μηνύματος ως θεμιτό. Η μόνη πιθανή περίπτωση ισοπαλίας για λ=9 και για τα k που δοκιμάστηκαν είναι να εντοπίσει ο 10-NN γειτονιά αποτελούμενη από 1 θεμιτό και 9 spam, περίπτωση που ενδεχομένως να μη συνέβη ποτέ στην πράξη και δεν επηρεάζει τα αποτελέσματα.

Τέλος, στο διάγραμμα (4-6) για λ=999 παρατηρείται πάλι η απότομη μεταβολή του *TCR* από αποδόσεις άνω της βάσης σε κάτω της βάσης και αντίστροφα, όπως και στο διάγραμμα (4-2). Οι k-NN για k<4 δεν ξεπερνούν ποτέ τη βάση, δηλαδή για καμιά δοκιμασθείσα διαστασιμότητα δε φτάνουν στο 100% *SP*. Για k=4 και 5, το κατορθώνουν μόνο για τα 250 features, γεγονός που δεν εμπνέει εμπιστοσύνη, αφού στην πράξη δεν μπορεί να προσδιορισθεί η ακριβής διαστασιμότητα που ξεπερνάει τη βάση. Πιο αξιόπιστοι είναι οι k-NN για k>6 και μόνο για μεγάλη διαστασιμότητα (τουλάχιστον 500-550 features).



Διάγραμμα 4-6: *TCR* του k-NN για λ=999 (t=0.999) για διάφορα k (με Information-Gain αποτίμηση χαρακτηριστικών).

Από τους ταξινομητές που πετυχαίνουν *SP*=100%, καλύτερος για την ίδια διαστασιμότητα είναι αυτός με το μικρότερο k, λόγω της καλύτερης *SR* που συνήθως επιφέρει, σύμφωνα με όσα ειπώθηκαν και για τα άλλα λ. Η *SR* που επιτυγχάνεται για το βέλτιστο συνδυασμό παραμέτρων (k=4, 250 features) είναι περίπου 65,3%, ενώ για το βέλτιστο "αξιόπιστο"

συνδυασμό παραμέτρων (k=7, 600 features) αγγίζει το 57%, ικανοποιητικό νούμερο για το δεδομένο επίπεδο αυστηρότητας του φίλτρου.

4.C) Αποτίμηση γειτόνων με βάση την απόσταση

Τα πλεονεκτήματα της αποτίμησης των γειτονικών στιγμιοτύπων με βάση την απόσταση τους από το στιγμιότυπο προς κατάταξη (distance weighting) στον k-NN είναι από καιρό γνωστά (π.χ. [Bailey & Jain 1978], [Dudani 1976], [McLeod et al.1987]). Θεωρήθηκε επομένως σκόπιμο να διερευνηθεί και αυτή η παράμετρος έχοντας βάσιμες ελπίδες για βελτίωση της απόδοσης. Πράγματι τα πειράματα επιβεβαίωσαν την κατά κοινή ομολογία θετική επίδραση της αποτίμησης των γειτόνων.

Στα πειράματα που έγιναν για τη διερεύνηση της αποτίμηση των features (ενότητα (4.B)), η παράμετρος distance weighting δε λήφθηκε υπόψη. Η επιλογή αυτή δικαιολογείται από τη λο γική που κρύβεται πίσω από την ιδέα των αποτιμήσεων. Η αποτίμηση γειτόνων ακολουθεί την επαγωγική προδιάθεση του k-NN πως οι εγγύτεροι γείτονες είναι πιο σχετικοί με το άγνωστο στιγμιότυπο από τους πιο απομακρυσμένους όσον αφορά την τιμή της συνάρτησης-στόχου, και κατά συνέπεια πρέπει να αποτιμηθούν περισσότερο. Προυπόθεση για την αποτελεσματικότητα στην πράξη αυτής της εύλογης πεποίθησης είναι πως ο υπολογισμός της απόστασης δίνεται από ένα μέτρο που εκφράζει όσο γίνεται πιο πιστά τη συσχέτιση μεταξύ των στιγμιοτύπων. Ένα τέτοιο μέτρο απαιτεί μια κατάλληλη συνάρτηση αποτίμησης των features. Έτσι, δεν έχει νόημα να δοκιμαστεί αποτίμηση γειτόνων αν το μέτρο που εκτιμά την απόσταση είναι κακό – το πιθανότερο είναι να δώσει χειρότερα αποτελέσματα, αφού θα θεωρούνται πιο σημαντικά τα στιγμιότυπα που είναι λιγότερο σχετικά με το άγνωστο παρά τα πραγματικά σχετικά. Γι' αυτό και αναζητήθηκε πρώτα ένα καλό μέτρο αποτίμησης των features και όταν αυτό προσδιορίστηκε (IG), στη συνέχεια μελετήθηκε η καταλληλότητα διαφόρων συναρτήσεων αποτίμησης γειτόνων.

4.C.I) Συναρτήσεις αποτίμησης γειτόνων

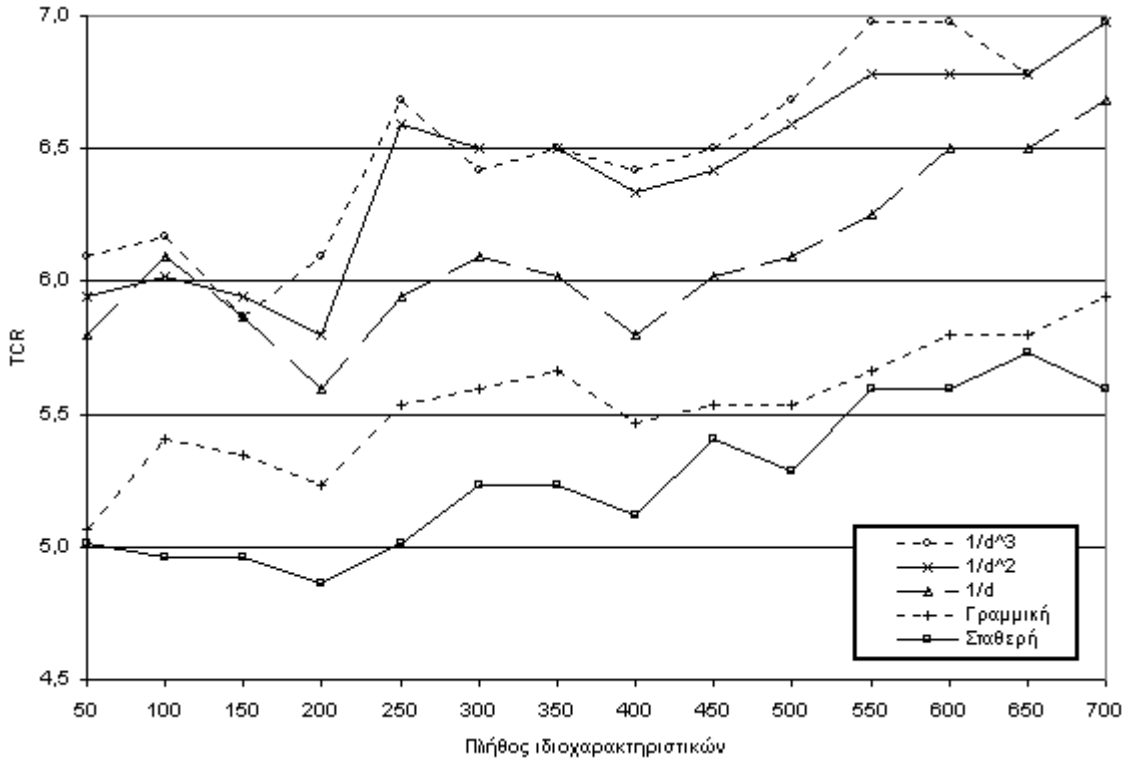
Πειράματα έγιναν για τέσσερις συναρτήσεις αποτίμησης γειτόνων. Η μία ήταν γραμμική: $f_0(d) = d_{\max} - d$, όπου d η απόσταση του γειτονικού στιγμιοτύπου από το άγνωστο και $d_{\max}$ η μέγιστη δυνατή απόσταση. Αυτή ισούται με το άθροισμα των βαρών όλων των features (εφόσον η μέγιστη απόσταση σημειώνεται αν δύο διανύσματα διαφέρουν σε όλα τους τα features) , όπως εκτιμώνται από το IG.

Οι άλλες τρεις ήταν υπερβολές:

$f_n(d) = \frac{1}{d^n}$, με το n να παίρνει τις τιμές 1,2 και 3. Όπως αναφέρθηκε και στην περιγραφή του k-NN στο κεφάλαιο 2, όταν ένας τουλάχιστον γείτονας ταυτίζεται με το άγνωστο στιγμιότυπο (δηλ. d=0), το τελευταίο κατατάσσεται στην κλάση που πλειοψηφεί για μηδενική απόσταση, ενώ όσοι γείτονες βρίσκονται σε μεγαλύτερες του μηδέν αποστάσεις αγνοούνται.

Στο διάγραμμα (4-7) παριστάνεται το TCR του 10-NN για λ=1 για τις τέσσερις συναρτήσεις αποτίμησης όπως και για την περίπτωση σταθεράς, (που αντιστοιχεί σε ισοβαρή

αποτίμηση, όπως ίσχυε για τα πειράματα που περιγράφηκαν στην ενότητα 4.Β) ως βάση αναφοράς. Το *k* επιλέχθηκε να είναι μεγάλο, αφ'ενός μεν επειδή οι συναρτήσεις αποτίμησης των γειτόνων διαφοροποιούνται περισσότερο μεταξύ τους για μεγάλες γειτονίες και αφ'ετέρου διότι συνήθως στην πράξη η χρήση τους συνδυάζεται με μεγάλο *k*. Άλλωστε, ένας από τους λόγους που χρειάζεται μια καλή συνάρτηση αποτίμησης γειτόνων είναι πως απαλλάσσει σε μεγάλο βαθμό το σχεδιαστή από την ανάγκη προσεκτικής επιλογής του *k*, γιατί από μια τιμή και μετά, οι παραπάνω γείτονες που προστίθενται δεν επηρεάζουν σχεδόν ποτέ την απόφαση του ταξινομητή. Το διάγραμμα για *λ*=9 δεν προσφέρει κάτι παραπάνω ως προς τη σύγκριση των καμπυλών και δεν παρουσιάζεται. Επίσης για *λ*=999, το διάγραμμα παραλείπεται διότι η βελτίωση εκεί δεν είναι ουσιαστική, καθώς όσα σημεία είχαν κάτω της βάσης απόδοση, παρέμειναν εκεί παρά την αποτίμηση, ενώ τα υπόλοιπα αυξήθηκαν ανεπαίσθητα.



Διάγραμμα 4-7: *TCR* για *λ*=1 (*t*=0.5) του 10-NN για τρεις συναρτήσεις αποτίμησης γειτόνων με βάση την απόσταση (distance weighting).

Το διάγραμμα δείχνει ξεκάθαρα τη βελτίωση που συνοδεύει την αποτίμηση γειτόνων, η οποία μάλιστα είναι τόσο μεγαλύτερη όσο δραστικότερα μειώνεται η σημασία των μακρινών γειτόνων. Είναι χαρακτηριστικό πως η κατάταξη των τεσσάρων χαμηλότερων καμπυλών παραμένει ουσιαστικά η ίδια για όλες τις διαστασιμότητες. Αντίθετα, η υπεροχή της $f_3(d)$ από την $f_2(d)$, όποτε συμβαίνει, είναι ελάχιστη, με τη δεύτερη μάλιστα να προηγείται για αρκετές διαστασιμότητες. Αυτό υποδεικνύει πως συναρτήσεις για μεγαλύτερα n δεν υπόσχονται παραπέρα βελτίωση και γι'αυτό δε δοκιμάστηκαν.

Η ερμηνεία που δόθηκε για αυτά τα αποτελέσματα ήταν η εξής: Οι k κοντινότερες αποστάσεις από τα άγνωστα στιγμιότυπα υπολογίζονται ως άθροισμα των IG βαρών των features στα οποία διαφέρει ένα υποψήφιο γειτονικό στιγμιότυπο από το άγνωστο (βλ. εξίσωση (2.12)). Τα βάρη αυτά στα πειράματα είναι συνήθως αρκετά μικροί αριθμοί, κάτω της μονάδας. Ο λόγος είναι πως για την περίπτωση δύο κλάσεων το *IG* είναι το πολύ μονάδα, όση και η μέγιστη εντροπία των κλάσεων *H(C)*, ενώ για τη συγκεκριμένη κατανομή των κατηγοριών στη συλλογή Ling-Spam είναι το πολύ 0.66. Έτσι, μικρές είναι εν γένει και οι διαφορές μεταξύ των αποστάσεων διαφορετικών γειτόνων, αφού ο ένας ταιριάζει σε λίγα περισσότερα features με το άγνωστο στιγμιότυπο απ' ότι άλλος. Η τελευταία παρατήρηση δικαιολογείται από το ότι κάθε ζευγάρι από στιγμιότυπα έχει πάρα πολλά μηδενικά features, λόγω της απουσίας των περισσότερων λέξεων αναπαράστασης από τα αντίστοιχα μηνύματα

Αν χρησιμοποιήσουμε μία γραμμική συνάρτηση, όπως η $f_1(d)$, στα κοντινότερα στιγμιότυπα δίνονται βάρη που είναι λίγο μεγαλύτερα από αυτά που δίνονται σε μακρινότερα στιγμιότυπα. Έτσι συνολικά σε όλη τη θεωρούμενη γειτονιά, ο παράγοντας που εξακολουθεί να παίζει κυρίαρχο λόγο στην απόφαση του ταξινομητή είναι το πλήθος των γειτόνων από κάθε κλάση και όχι η απόσταση τους από το στόχο. Για παράδειγμα, αν μια γειτονιά αποτελείται από 4 spam και 6 θεμιτά στιγμιότυπα, είναι συνήθως δύσκολο να κατατάξει ο ταξινομητής το άγνωστο μήνυμα ως spam αν χρησιμοποιείται γραμμική συνάρτηση αποτίμησης, ακόμα και αν τα τέσσερα spam είναι πιο κοντά από τα έξι θεμιτά. Τα δύο παραπάνω θεμιτά μηνύματα θα απέχουν συνήθως συγκρίσιμη απόσταση με τα κοντινά και το λίγο παραπάνω βάρος που θα δοθεί στα τελευταία δεν είναι αρκετό για να υπερκαλύψει το βάρος δύο παραπάνω στιγμιοτύπων της άλλης κλάσης. Στην πράξη, η κύρια αιτία κέρδους από μια "μετριοπαθή" συνάρτηση αποτίμησης είναι η σωστή επίλυση ισοπαλιών (π.χ. 5-5), όπου εκεί πραγματικά παίζει ρόλο η απόσταση των γειτόνων κάθε κλάσης από το στόχο. Αντίθετα, μια συνάρτηση που δίνει σημαντική ώθηση στους κοντινούς γείτονες μπορεί να αναστρέψει την απόφαση του ταξινομητή υπέρ της μειοψηφίας, αν αυτή αθροιστικά βρίσκεται πλησιέστερα στο άγνωστο στιγμιότυπο

Ένας ακόμα λόγος αυξημένης απόδοσης των υπερβολικών συναρτήσεων που διαπιστώθηκε εκ των υστέρων μετά από προσεκτικότερη εξέταση των κατατάξεων είναι η σωστή πρόβλεψη των περιπτώσεων ταύτισης ενός ή παραπάνω γειτόνων με το στόχο. Είναι σχεδόν σίγουρο πως σε τέτοιες περιπτώσεις, το άγνωστο στιγμιότυπο ανήκει στην ίδια κλάση με τα ταυτιζόμενα γειτονικά, έστω κι αν στην ευρύτερη γειτονιά πλειοψηφεί η άλλη κλάση. Οι συναρτήσεις $f_n(d)$ κατατάσσουν σωστά όλα αυτά τα στιγμιότυπα. Φυσικά, αυτό δεν έχει να κάνει με το ότι οι συναρτήσεις είναι υπερβολές: οποιαδήποτε συνάρτηση μπορεί – και ίσως πρέπει – να οριστεί έτσι ώστε να χειρίζεται ιδιαίτερα τις περιπτώσεις ακριβούς ταύτισης.

Από το διάγραμμα φαίνεται επίσης πως οι καμπύλες δεν παρουσιάζουν σημαντική διαφορά ως προς την εξάρτησή τους από τη διαστασιμότητα. Και οι πέντε μπορούν να θεωρηθούν αύξουσες στο εύρος των 700 features, αλλά και στις περιπτώσεις που φθίνουν τοπικά (π.χ. 200 features), πέφτουν και οι πέντε. Δίνεται έτσι πειραματικά έρεισμα στην υπόθεση ανεξαρτησίας μεταξύ αποτίμησης γειτόνων και διαστασιμότητας.

4.C.II) Επίδραση της παραμέτρου *k*

Στο σημείο αυτό θα μπορούσαμε να υιοθετήσουμε ως συνάρτηση αποτίμησης των γειτόνων την $f_3(d) = 1/d^3$ και να αρκεστούμε στα πειράματα που έγιναν με τον 10-NN, υποθέτοντας πως μικρότερα *k* δεν αναμένεται να βελτιώσουν την απόδοση. Ο λόγος, που αναφέρθηκε ήδη, είναι πως η επιλογή του *k* είναι λιγότερο κρίσιμη όταν γίνεται αποτίμηση γειτόνων, τουλάχιστον από ένα ελάχιστο *k* και πάνω, ενώ για μικρότερες γειτονίες τα οφέλη της αποτίμησης γίνονται ολοένα και λιγότερα. Παρ'όλα αυτά, μιας και από τις τέσσερις αρχικές παραμέτρους έχουν μείνει αδέσμευτες οι δύο (*k* και διαστασιμότητα) και για τις άλλες δύο έχει προκύψει βέλτιστη επιλογή, ανεξάρτητη των άλλων παραμέτρων (*IG* αποτίμηση χαρακτηριστικών και $1/d^3$ αποτίμηση γειτόνων), είναι πλέον εφικτή μια "εξαντλητική"* αναζήτηση του βέλτιστου συνδυασμού των αδέσμευτων παραμέτρων. Η αναζήτηση αυτή έδειξε πως:

- Για λ=1, η απόδοση γενικά βελτιώνεται όσο η ακτίνα γειτονίας αυξάνεται, με την καλύτερη τιμή να είναι μεταξύ 7 και 8, ανάλογα και με τη διαστασιμότητα, ενώ για *k*=9 και 10 χειροτερεύει ελαφρώς. Όλες οι καμπύλες δείχνουν να βελτιώνονται με την αύξηση της διαστασιμότητας, αν και τοπικά για κάποιο διάστημα μπορεί να είναι φθίνουσες. Αυτά είναι και τα μόνα γενικά συμπεράσματα που μπορούν να προκύψουν, καθώς οι επικαλύψεις των καμπυλών είναι πολλές και συχνές και δεν προσφέρονται για πιο λεπτομερείς αναλύσεις.
- Για λ=9, η εικόνα δεν είναι και πολύ διαφορετική σε σύγκριση με την περίπτωση που δε γίνεται αποτίμηση γειτόνων, πέραν από μια κατακόρυφη μετατόπιση όλων των καμπυλών προς τα πάνω†. Πάλι οι 2-NN και 3-NN έρχονται πρώτοι για τις περισσότερες διαστασιμότητες, ενώ μεγαλύτερες γειτονίες ρίχνουν προοδευτικά την απόδοση. Ο λόγος που το βέλτιστο μέγεθος γειτονίας παραμένει μεταξύ 2 και 3 και όχι μεγαλύτερο οφείλεται στο ότι για αυτό το σενάριο, η *SR* για μεγάλο *k* είναι αρκετά χαμηλότερη απ'ό,τι για μικρό. Για παράδειγμα, χωρίς αποτίμηση γειτόνων, η διαφορά στην *SR* του 3-NN από του 8-NN κυμαίνεται από 13% μέχρι 19% υπέρ του πρώτου. Αν και η αποτίμηση αποδεικνύεται πιο επωφελής για μεγάλα *k* παρά για μικρά, το άνοιγμα μεταξύ τους απλά ελαττώνεται, παραμένοντας όμως μεγάλο. Έτσι π.χ. η διαφορά στην *SR* του 3-NN από του 8-NN, αφού γίνει αποτίμηση, κυμαίνεται μεταξύ 7.5% και 12.5% περίπου. Σε αυτό το προβάδισμα στην *SR* οφείλεται τελικά η υπεροχή των 2-NN και 3-NN ως προς το *TCR*, καθώς η ελάχιστα καλύτερη *SP* των υπολοίπων (κάτω του 1% για τον 3-NN), αν και υπερεκτιμώμενη λόγω του λ, δεν επαρκεί για να ανατρέψει τη διαφορά.
- Για λ=999, τέλος, δεν έχει να προστεθεί τίποτα σε αυτά που αναφέρθηκαν στα πειράματα χωρίς αποτίμηση γειτόνων πλην μιας μικρής θετικής κατακόρυφης μετατόπισης όλων των καμπυλών.

---

* Στην πραγματικότητα δεν είναι εξαντλητική, αφού το k στα πειράματα κυμαίνεται μεταξύ 1 και 10 και η διαστασιμότητα μεταξύ 50 και 700 με βήμα 50.

† Εκτός του 1-NN για τον οποίο δεν έχει νόημα η αποτίμηση απόστασης.

Στον πίνακα (4.1) παρατίθενται οι βέλτιστες επιδόσεις για κάθε σενάριο και οι τιμές του *k* και της διαστασιμότητας στις οποίες επιτεύχθηκαν. Για όλα τα σενάρια το μέτρο αποτίμησης των features είναι το *IG* και συνάρτηση αποτίμησης γειτόνων η $1/d^3$.

λ	*k*	Διαστασιμότητα	Spam Ανάκληση (*SR*)	Spam Ορθότητα (*SP*)	*TCR*
1	8	600	88,60%	97,39%	7,18
9	2	700	81,93%	98,79%	3,64
999	4	250	68,02%	100%	3,12
999 Stable*	7	600	59,91%	100%	2,49

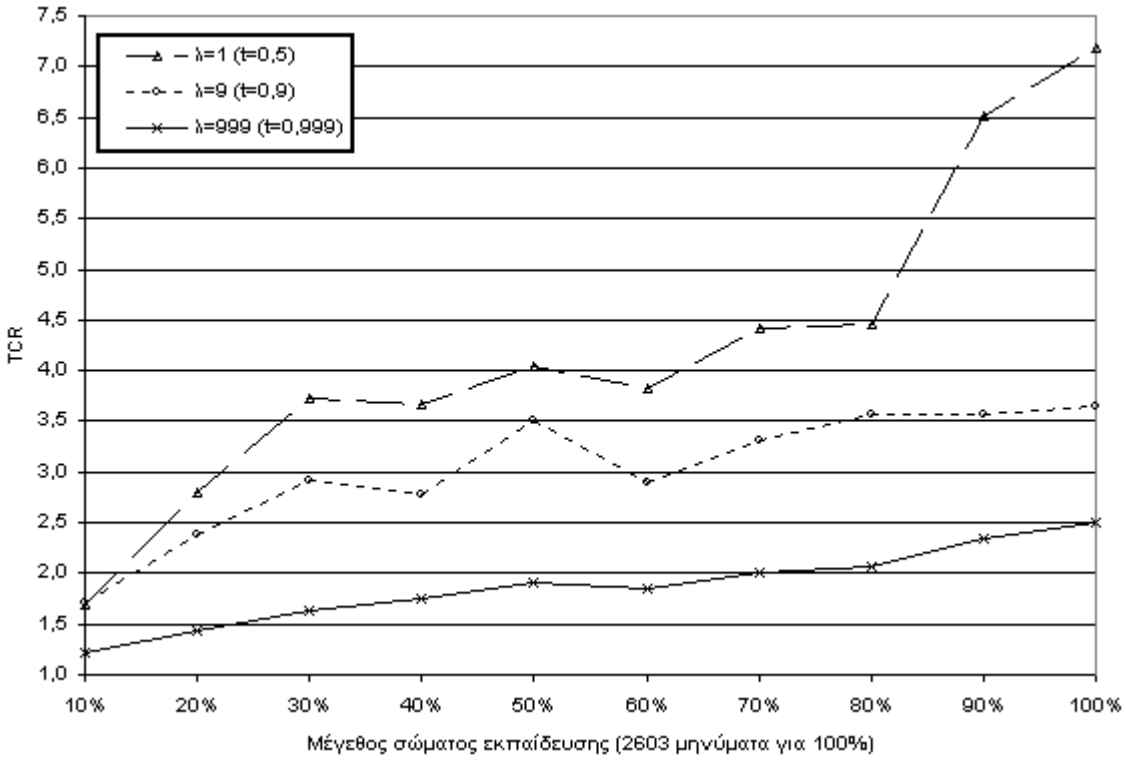
Πίνακας 4.1: Βέλτιστες επιδόσεις ανά σενάριο
και οι συνδυασμοί παραμέτρων για τις οποίες σημειώθηκαν.

4.D) Επίδραση του μεγέθους του σώματος εκπαίδευσης

Στο διάγραμμα (4-8) παριστάνεται η επίδοση του *k*-NN ως συνάρτηση του μεγέθους του σώματος εκπαίδευσης. Όπως και στα προηγούμενα πειράματα, έγινε στρωματοποιημένη διασταυρωμένη επικύρωση 10 σημείων, με τη διαίρεση του συνόλου των δεδομένων σε 10 ίσα (περίπου) τμήματα. Σε καθεμιά από τις 10 επαναλήψεις, ένα τμήμα αποτελούσε το σύνολο ελέγχου και από τα υπόλοιπα 9 το x% χρησιμοποιήθηκε για εκπαίδευση, με το x να παίρνει τιμές από 10 έως 100 με βήμα 10. Οι καμπύλες αφορούν τους βέλτιστους συνδυασμούς παραμέτρων (configurations) για κάθε σενάριο, όπως αυτές εκτιμήθηκαν για το 100% του σώματος εκπαίδευσης και δίνονται στον πίνακα (4.1). Γενικά, οι καμπύλες που εμφανίζουν την απόδοση ενός συστήματος μηχανικής μάθησης πάνω σε μια συγκεκριμένη συλλογή δεδομένων για μια εφαρμογή, εκφρασμένη σε κατάλληλο για την εφαρμογή μέτρο, συναρτήσει του μεγέθους του σώματος εκπαίδευσης λέγονται *καμπύλες μάθησης* (*learning curves*). Ο κανόνας είναι πως οι καμπύλες μάθησης είναι εν γένει αύξουσες καθώς αυξάνονται τα δεδομένα εκπαίδευσης, τουλάχιστον μέχρι κάποιο σημείο από το οποίο και μετά τείνουν ασυμπτωτικά στο βέλτιστο. Σε αυτό το σημείο θεωρείται πως το σύστημα έχει φτάσει στα όριά του και δεν πρόκειται να βελτιωθεί με επιπλέον εκπαίδευση.

Στο διάγραμμα επαληθεύεται το γεγονός της βελτίωσης του ταξινομητή με την προσθήκη δεδομένων εκπαίδευσης, και για τα τρία σενάρια. Η άνοδος των καμπυλών είναι ομαλή σε γενικές γραμμές, με την εξαίρεση του άλματος από 4.5 σε 6.5 *TCR* για λ=1. Η άνοδος του *TCR* μεταφράζεται σε άνοδο της *SR* και αρκετά μικρότερη μείωση της *SP* (π.χ. για λ=1, η μετάβαση απ'το 80% στο 100% διορθώνει την *SR* κατά 10% περίπου και επιβαρύνει την *SP* κατά 1.6%). Και οι τρεις καμπύλες πάντως δείχνουν να έχουν δυνατότητες παραπέρα βελτίωσης αν δοθεί περισσότερη εκπαίδευση, κυρίως για λ=1.

* Το "αξιόπιστο" βέλτιστο, υπό την έννοια πως ανήκει σε διάστημα στο οποίο, με βάση τα πειραματικά δεδομένα, η ορθότητα φαίνεται να διατηρείται συνεχώς στο 100%.



Διάγραμμα 4-8: *TCR* για μεταβλητό μέγεθος σώματος εκπαίδευσης για τον καλύτερο συνδυασμό παραμέτρων για κάθε σενάριο (λ).

5) ΠΕΙΡΑΜΑΤΑ ΜΕ ΟΜΑΔΕΣ ΤΑΞΙΝΟΜΗΤΩΝ

Το δεύτερο μέρος της εργασίας περιλαμβάνει πειράματα στα οποία συνδυάζονται ο αλγόριθμος των *k*-κοντινότερων γειτόνων (*k*-NN) με τον απλοϊκό ταξινομητή Bayes (NB) μέσω ενός κοινού σχήματος. Το συγκεκριμένο σχήμα αποτελεί μία από τις πολλές τεχνικές που έχουν προταθεί και οι οποίες χρησιμοποιούν ανεξάρτητα κατασκευασθέντες ταξινομητές για να σχηματίσουν έναν νέο, κατά τρόπο τέτοιο ώστε ο τελευταίος να έχει μικρότερο ρυθμό λαθών γενίκευσης (generalization error rate) από καθέναν απ' τους αρχικούς. Οι τεχνικές αυτές ονομάζονται *ομάδες ταξινομητών* (*classifier ensembles*) ή *επιτροπές ταξινομητών* (*classifier committees*) και είναι μια από τις πιο ενεργές ερευνητικές περιοχές στο χώρο της επιβλεπόμενης μηχανικής μάθησης (supervised learning). Επισημαίνεται εδώ πως κοινό χαρακτηριστικό των ομάδων ταξινομητών είναι ότι χρησιμοποιούν τους αρχικούς ταξινομητές ως "μαύρα κουτιά", δηλ. ως συστήματα που δέχονται στιγμιότυπα για είσοδο και δίνουν στην έξοδο την προβλεπόμενη τιμή της συνάρτησης-στόχου για τα στιγμιότυπα αυτά. Διακρίνονται έτσι από τους υβριδικούς αλγορίθμους, οι οποίοι προσπαθούν να συνδυάσουν ενδογενώς παραπάνω από έναν αλγορίθμους μάθησης.

Αρχικά θα δοθούν κάποια γενικά στοιχεία για τις ομάδες ταξινομητών, για τα πιο γνωστά σχήματα που έχουν μελετηθεί και χρησιμοποιηθεί στο παρελθόν και για τη συγκεκριμένη τεχνική που εφαρμόστηκε στην εργασία. Ακολούθως θα παρατεθούν κάποια προκαταρκτικά πειραματικά αποτελέσματα που οδήγησαν στην απόφαση για πειραματισμό με ομάδες ταξινομητών. Αναφέρονται κατόπιν οι δύο παραλλαγές της μεθοδολογίας που δοκιμάστηκαν και τα πλεονεκτήματα και μειονεκτήματα της καθεμιάς. Τέλος παρουσιάζονται τα αποτελέσματα των πειραμάτων, κάποια στατιστικά στοιχεία σχετικά με τις προβλέψεις του σύνθετου ταξινομητή σε σύγκριση με αυτές των απλών και το κεφάλαιο κλείνει με τη σύνοψη των καλύτερων παρατηρηθέντων επιδόσεων.

5.Α) Ομάδες ταξινομητών

Η βασική ιδέα πίσω από τις ομάδες ταξινομητών είναι πως, δοθείσας μιας εργασίας που απαιτεί γνώσεις ειδικού (expert) για να εκτελεστεί με επιτυχία, πολλοί ειδικοί ίσως είναι αποτελεσματικότεροι από έναν, αν οι ατομικές αποφάσεις τους συνδυαστούν κατάλληλα. Μία τέτοια "επιτροπή ειδικών" χαρακτηρίζεται από δύο επιλογές:

(i) Την επιλογή των ειδικών που μετέχουν σε αυτήν.

(ii) Την επιλογή του τρόπου συνδυασμού των αποφάσεών τους.

Σχετικά με το πρώτο θέμα, έχει παγιωθεί στη βιβλιογραφία της μηχανικής μάθησης πως για να είναι αποτελεσματική μια επιτροπή, τα μέλη της θα πρέπει να είναι όσο το δυνατόν πιο ανεξάρτητα μεταξύ τους, δηλ. θα πρέπει οι αποφάσεις τους να είναι όσο γίνεται πιο ασυσχέτιστες. Αυτό ακούγεται αρκετά λογικό, αφού αν τις περισσότερες φορές όλα τα μέλη συμφωνούν μεταξύ τους, οι αποφάσεις της επιτροπής θα ακολουθούν συνήθως τη γνώμη της πλειοψηφίας και κατά συνέπεια θα συμβαίνουν τα ίδια λάθη που θα συνέβαιναν και με ένα μόνο ταξινομητή.

Πολλές μέθοδοι κατασκευής ταξινομητών που προορίζονται για μέλη επιτροπής έχουν αναπτυχθεί, άλλες ειδικές για συγκεκριμένους αλγορίθμους μάθησης και άλλες γενικές. Οι τελευταίες χωρίζονται σε τέσσερις κατηγορίες:

1) Στην δειγματοληψία των δεδομένων εκπαίδευσης

Σε αυτή την κατηγορία οι διαφορετικοί ταξινομητές προκύπτουν από εκπαίδευση σε διαφορετικά υποσύνολα των διαθέσιμων δεδομένων. Η μέθοδος δουλεύει ιδιαίτερα καλά για *ασταθείς* αλγορίθμους μάθησης – αλγορίθμους οι οποίοι παράγουν ταξινομητές αρκετά διαφορετικούς με μικρές αλλαγές του συνόλου εκπαίδευσης. Τα δέντρα απόφασης και τα νευρωνικά δίκτυα είναι παραδείγματα ασταθών αλγορίθμων, ενώ αντίθετα ο *k*-NN, ο NB και οι μέθοδοι γραμμικής παλινδρόμησης (linear regression) είναι γενικά πολύ σταθεροί.

Η πιο άμεση μέθοδος δειγματοληψίας των δεδομένων λέγεται *bagging* (*Bootstrap AGGregation*). Ένας αλγόριθμος μάθησης εκπαιδεύεται σε διαφορετικά δείγματα του συνόλου των δεδομένων εκπαίδευσης, όπως ακριβώς αυτά που χρησιμοποιούνται για τη στατιστική πιστοποίηση των αποτελεσμάτων ενός αλγορίθμου (βλ. ενότητα (2.C.III.b)). Επίσης δανεισμένη από τις τεχνικές στατιστικής επικύρωσης είναι οι *επιτροπές διασταυρωμένης επικύρωσης* (*cross-validated committees*) [Parmanto et al. 1996], όπου τα υποσύνολα των δεδομένων στα οποία εκπαιδεύονται οι ταξινομητές καθορίζονται από τη γνωστή μέθοδο διασταυρωμένης επικύρωσης.

Μία πιο πολύπλοκη μέθοδος που έχει εμφανιστεί τα τελευταία χρόνια και έχει επικεντρώσει την προσοχή πολλών ερευνητών λόγω της θεαματικής επιτυχίας της είναι η μέθοδος της λεγόμενης *προώθησης* (*boosting*) [Schapire 1990], και κυρίως η βελτίωσή της, η προσαρμοστική προώθηση (Adaptive boosting – AdaBoost) [Freund & Schapire 1995]. Η βασική ιδέα της έγκειται στην, εννοιολογικά, σειριακή εκπαίδευση των ταξινομητών από έναν αλγόριθμο μάθησης, ο ένας μετά τον άλλο, αντί της παράλληλης που γίνεται συνήθως. Με αυτό τον τρόπο, ο ν-οστός ταξινομητής λαμβάνει υπόψη την επίδοση των προηγούμενων ν-1 ταξινομητών για καθένα στιγμιότυπο, επικεντρώνοντας έτσι την προσπάθειά του στη σωστή πρόβλεψη εκείνων που οι προηγούμενοι απέτυχαν περισσότερο. Η εστίαση σε αυτά τα "δύσκολα" στιγμιότυπα γίνεται με τη διατήρηση βαρών για τα στιγμιότυπα. Σε κάθε επανάληψη τα βάρη ενημερώνονται, μειούμενα για τα στιγμιότυπα που προβλέπονται σωστά και αυξανόμενα για αυτά που προβλέπονται λάθος. Σε κάθε επανάληψη, τα στιγμιότυπα επιλέγονται τυχαία με αντικατάσταση, ανάλογα με το σχετικό βάρος τους[*]. Διαδοχικά λοιπόν συμμετέχουν όλο και δυσκολότερα στιγμιότυπα για εκπαίδευση και έλεγχο. Τελικά, όλοι οι ταξινομητές που προκύπτουν με αυτό τον τρόπο συνδυάζονται μέσω ενός κανόνα, στον οποίο ο κάθε ταξινομητής συμμετέχει με βάρος εξαρτώμενο από την ακρίβειά του στο σύνολο που εκπαιδεύθηκε. Το bagging, το boosting και διάφορες παραλλαγές τους συγκρίνονται πειραματικά στο [Kohavi & Bauer 1998].

[*] Αν ο αλγόριθμος μάθησης που χρησιμοποιείται για boosting μπορεί να αξιοποιήσει άμεσα αυτά τα σχετικά βάρη, τότε δεν απαιτείται τυχαία επαναδειγματοληψία (*resampling*) των δεδομένων. Αυτή η παραλλαγή λέγεται boosting με επαναζύγιση (*reweighting*) και γενικά δίνει καλύτερα αποτελέσματα.

2) Στην αλλαγή της αναπαράστασης των δεδομένων εισόδου

Αυτή η τεχνική είναι κατάλληλη μόνο για εφαρμογές που περιέχουν χαρακτηριστικά με υψηλό πλεονασμό, όπως οι λέξεις στην κατηγοριοποίηση κειμένου.

3) Στην αλλαγή της συνάρτησης-στόχου

Εδώ ανήκει μια πρωτότυπη τεχνική που ονομάζεται *κωδικοποίηση διόρθωσης λαθών εξόδου* (*error-correcting output coding*) [Dietterich & Bakiri 1995]. Αυτή μετασχηματίζει το αρχικό πρόβλημα πολλών κλάσεων σε πολλά προβλήματα δύο κλάσεων, για καθένα από τα οποία εκπαιδεύεται ένας ταξινομητής. Καθεμιά από τις αρχικές κλάσεις αντιστοιχίζεται σε μία συμβολοσειρά δυαδικών ψηφίων (bit-string) και το ίδιο συμβαίνει για κάθε άγνωστο στιγμιότυπο που κατατάσσεται από τους εκπαιδευθέντες ταξινομητές. Τελικά, το στιγμιότυπο κατατάσσεται στην κλάση εκείνη της οποίας η κωδική λέξη είναι πιο κοντά (κατά απόσταση Hamming) στη δική του. Η κατάλληλη διάσπαση του προβλήματος σε υποπροβλήματα δύο κλάσεων μπορεί να γίνει με τη βοήθεια μεθόδων σχεδίασης καλών κωδίκων διόρθωσης λαθών που χρησιμοποιούνται στη ψηφιακή επεξεργασία σήματος.

4) Στην εισαγωγή τυχαιότητας στον αλγόριθμο μάθησης

Αυτή η τεχνική μπορεί να χρησιμοποιηθεί σε αλγορίθμους μάθησης που είτε χρησιμοποιούν κάποιες αρχικές παραμέτρους (π.χ. στα νευρωνικά δίκτυα, τα αρχικά βάρη των συνάψεων μπορούν να επιλεγούν τυχαία, δίνοντας αρκετά διαφορετικούς ταξινομητές) ή αποτελούν επεκτάσεις ντετερμινιστικών αλγορίθμων (π.χ. [Dietterich & Kong 1995] για το γνωστό αλγόριθμο δέντρων απόφασης C4.5 [Quinlan 1993]).

Όσον αφορά τον τρόπο συνδυασμού των αποφάσεων των ταξινομητών, ο πιο απλός κανόνας είναι η πλειοψηφική ψηφοφορία (*majority voting*), κατά την οποία επιλέγεται η κλάση που προβλέπει η πλειοψηφία των ταξινομητών. Μια βελτίωση αυτού του σχήματος είναι δυνατή αν κάθε ταξινομητής δίνει στην έξοδό του την πίστη του ως προς τις αποφάσεις του, οπότε εκλέγεται η κλάση που κέρδισε τη μέγιστη μέση εμπιστοσύνη. Άλλες τεχνικές εκτιμούν το βάρος κάθε ταξινομητή στην επιτροπή, ανάλογα με την αποτελεσματικότητά του σε ένα σύνολο δεδομένων επικύρωσης. Αυτά τα βάρη χρησιμοποιούνται στη συνέχεια, π.χ. μέσω γραμμικού συνδυασμού, για την πρόβλεψη της κλάσης από την επιτροπή. Μία ακόμα πολιτική είναι να επιλέγεται δυναμικά, ανά στιγμιότυπο προς κατάταξη, ο ταξινομητής εκείνος ο οποίος είχε την καλύτερη επίδοση στα Ν στιγμιότυπα επικύρωσης που συγγενεύουν περισσότερο με το άγνωστο και αυτός να αποφασίζει (*dynamic classifier selection*) [Li & Jain 1998]. Οι δύο προηγούμενες ιδέες μπορούν να συνδυαστούν, έτσι ώστε να λαμβάνονται υπόψη όλοι οι ταξινομητές, αλλά η απόφασή τους να μετράει με βάση την αποτελεσματικότητά τους στα Ν στιγμιότυπα επικύρωσης που μοιάζουν περισσότερο με το άγνωστο (*adaptive classifier combination*) [Li & Jain 1998]. Τέλος, μια ακόμα μέθοδος είναι η λεγόμενη *συσσωρευμένη γενίκευση* (*stacked generalization*), που βασίζεται σε μια διεπίπεδη ιεραρχία ταξινομητών. Λόγω του ότι αυτή η τεχνική εφαρμόστηκε στην εργασία για τον

πειραματισμό με τις ομάδες ταξινομητών, θα περιγραφεί σε μεγαλύτερο βάθος στην επόμενη ενότητα.

Οι ομάδες ταξινομητών έχουν δώσει μέχρι στιγμής πολύ ενθαρρυντικά αποτελέσματα και αποτελούν μία από τις πλέον ελπιδοφόρες κατευθύνσεις στη μηχανική μάθηση. Αλγόριθμοι όπως ο AdaBoost και οι επεκτάσεις του θεωρούνται σήμερα η καλύτερη επιλογή για πολλά πρακτικά προβλήματα. Ήδη αρχίζει να εμφανίζεται και θεωρητική αιτιολόγηση της επιτυχίας τους, πέρα από τα καλά αποτελέσματα ([Schapire et al. 1997]). Το κύριο μειονέκτημα, ωστόσο, αυτών των αλγορίθμων είναι το υψηλό κόστος τους σε υπολογιστικό χρόνο και μνήμη, καθώς περιλαμβάνουν την εκπαίδευση δεκάδων ή και εκατοντάδων ταξινομητών. Μία λύση σε αυτό το πρόβλημα είναι η χρήση παραλληλίας, τουλάχιστον στην περίπτωση που κάθε ταξινομητής μπορεί να εκπαιδευθεί ανεξάρτητα από τους άλλους. Επίσης αντικείμενο έρευνας είναι το κατά πόσο μπορεί να μετατραπεί μια επιτροπή ταξινομητών σε μια πιο συμπαγή και κατανοητή μορφή, αφαιρώντας ίσως τα μέλη με υψηλό βαθμό συσχέτισης ή με μετασχηματισμούς της αναπαράστασής τους.

Συσσωρευμένη γενίκευση

Η *συσσωρευμένη γενίκευση* (*stacked generalization*) [Wolpert 1992], η οποία αναφέρεται πιο απλά και ως stacking, είναι μια γενική μέθοδος που χρησιμοποιεί ένα υψηλού επιπέδου μοντέλο (ταξινομητή) για να συνδυάσει χαμηλότερου επιπέδου μοντέλα με σκοπό την επίτευξη μεγαλύτερης ακρίβειας πρόβλεψης. Στην πράξη έχει εφαρμοστεί τόσο για μοντέλα κατασκευασμένα για εργασίες κατηγοριοποίησης (διακριτής συνάρτησης-στόχου) [Wolpert 1992], όσο και για μοντέλα παλινδρόμησης (regression) [Breiman 1996], ακόμα και για μη επιβλεπόμενη μάθηση [Smyth & Wolpert 1997].

Η γενική ιδέα του stacking έχει ως εξής: Ένας ή περισσότεροι αλγόριθμοι μάθησης εκπαιδεύονται αρχικά σε ένα πλήθος από υποσύνολα των αρχικών δεδομένων, παράγοντας αντίστοιχο αριθμό από μοντέλα. Στη συνέχεια, κάθε στιγμιότυπο από τα αρχικά δεδομένα αντιστοιχίζεται σε ένα νέο, το οποίο αναπαριστά την πρόβλεψη κάθε μοντέλου για το αρχικό στιγμιότυπο, καθώς και την πραγματική τιμή της συνάρτησης-στόχου. Στο βήμα αυτό πρέπει να εξασφαλιστεί πως τα μοντέλα δημιουργούνται από σύνολα εκπαίδευσης που δεν περιλαμβάνουν το στιγμιότυπο που αναπαρίσταται (για το οποίο δηλαδή κάνουν πρόβλεψη), ακριβώς όπως ισχύει στη διασταυρωμένη επικύρωση. Τα νέα δεδομένα σχηματίζουν ένα νέο πρόβλημα μάθησης, και στο δεύτερο βήμα της μεθόδου ένας αλγόριθμος μάθησης καλείται να το λύσει, παράγοντας το μοντέλο του δεύτερου επιπέδου. Σύμφωνα με την ορολογία του Wolpert, τα αρχικά δεδομένα και τα μοντέλα που κατασκευάζονται για αυτά στο πρώτο βήμα αναφέρονται ως *δεδομένα επιπέδου-0* και *μοντέλα επιπέδου-0*, αντίστοιχα, ενώ τα τα νέα δεδομένα που προέρχονται από τις προβλέψεις των μοντέλων επιπέδου-0 και ο αλγόριθμος μάθησης που καλείται στο δεύτερο βήμα αναφέρονται αντίστοιχα ως *δεδομένα επιπέδου-1* και *γενικευτής επιπέδου-1*.

Σχηματικά, τα μοντέλα επιπέδου-0 μπορούν να θεωρηθούν ως τα "απλά μέλη" της επιτροπής ταξινομητών, ενώ το μοντέλο που κατασκευάζεται από τα δεδομένα επιπέδου-1 αντιστοιχεί στον "πρόεδρο" της επιτροπής, ο οποίος εκπαιδεύεται πάνω στις προβλέψεις των

μελών για κάθε στιγμιότυπο. Παρακάτω θα χρησιμοποιούνται οι όροι "μέλος" και "πρόεδρος" αντί των "μοντέλο επιπέδου-0" και "μοντέλο επιπέδου-1". Στη φάση της κατάταξης ενός νέου στιγμιοτύπου, κάθε μέλος κάνει ανεξάρτητα την πρόβλεψή του και στη συνέχεια ο πρόεδρος παίρνει την τελική απόφαση, λαμβάνοντας υπόψη τις γνώμες των μελών σύμφωνα με το μοντέλο που έχει κατασκευαστεί.

Μερικές από τις επιλογές που είναι, πιθανότατα, κρίσιμες για την αποτελεσματικότητα αυτού του σχήματος είναι:

i) Οι αλγόριθμοι μάθησης για την κατασκευή των μελών.
ii) Ο αλγόριθμος μάθησης για την κατασκευή του προέδρου.
iii) Το πλήθος των μελών.
iv) Η αναπαράσταση των δεδομένων επιπέδου-1.

Για τα ζητήματα αυτά, τα οποία κατά την πρώτη δημοσίευση της μεθόδου ο Wolpert χαρακτήρισε ως "μαύρη τέχνη", δεν έχουν δοθεί ακόμα γενικές απαντήσεις. Θεωρητικά, οποιοσδήποτε αλγόριθμος μάθησης μπορεί να χρησιμοποιηθεί στα (i) και (ii), και πράγματι έχουν δοκιμαστεί διάφοροι γνωστοί αλγόριθμοι χωρίς να προκύψουν ξεκάθαρα συμπεράσματα. Ως προς το (iv), πέραν από την περίπτωση που περιγράφηκε παραπάνω, κατά την οποία τα δεδομένα επιπέδου-1 είναι απλά οι προβλέψεις των μελών, μία άλλη επιλογή είναι δυνατή αν τα μέλη βγάζουν ως έξοδο το βαθμό εμπιστοσύνης τους για την κατάταξη του στιγμιοτύπου σε κάθε μία από τις κλάσεις. Σε αυτή την περίπτωση, τα δεδομένα του επιπέδου-1 μπορούν να αποτελούνται από το βαθμό εμπιστοσύνης που δίνει το κάθε μέλος σε κάθε μία κλάση.

5.Β) Κίνητρο συνδυασμού NB με *k*-NN

Η βασική ιδέα των πειραμάτων με επιτροπές ταξινομητών στην εργασία ήταν να συνδυαστούν οι δύο αλγορίθμοι μάθησης που είχαν ήδη χρησιμοποιηθεί στη συλλογή μηνυμάτων Ling-Spam, δηλαδή ο NB και ο *k*-NN. Πέραν από τα πολύ ενθαρρυντικά αποτελέσματα που έχουν σημειώσει οι ομάδες ταξινομητών σε διάφορα προβλήματα μηχανικής μάθησης όπου εφαρμόστηκαν, επιπλέον κίνητρο για τον πειραματισμό με αυτές στα πλαίσια του φιλτραρίσματος spam μηνυμάτων έδωσαν και κάποιες μετρήσεις που προηγήθηκαν. Αυτές έγιναν με σκοπό να εκτιμηθεί το κατά πόσο οι ταξινομητές που κατασκευάζονται με βάση τους NB και *k*-NN είναι ασυσχέτιστοι μεταξύ τους. Μόνο στην περίπτωση που η συσχέτισή τους κρινόταν μικρή, όπως και έγινε, θα πραγματοποιούνταν τα παρακάτω πειράματα.

Ένας απλός, αλλά ενδεικτικός, τρόπος να εκτιμηθεί στην πράξη η εν λόγω συσχέτιση ήταν να μετρηθούν τα λάθη ταξινόμησης που κάνουν από κοινού οι ταξινομητές και τα λάθη που κάνει μόνο ο ένας εκ των δύο. Αν τα τελευταία ήταν αρκετά περισσότερα από τα πρώτα, ο πρόεδρος θα είχε αρκετά περιθώρια να "μάθει" να επιλέγει το σωστό μέλος κάθε φορά. Αντίθετα, αν οι ταξινομητές έκαναν συνήθως τα ίδια λάθη, με δεδομένη τη γενικά υψηλή τους ακρίβεια στο σύνολο επικύρωσης, ο πρόεδρος θα μάθαινε να εμπιστεύεται σχεδόν πάντα τη γνώμη τους όταν αυτοί συμφωνούνε και κατά συνέπεια θα αναπαρήγαγε τα λάθη τους. Αν και είναι δυνατόν να διαφωνήσει ο πρόεδρος και με τα δύο μέλη, στην πράξη αυτή η

απόφαση είναι πιο παρακινδυνευμένη από το να συμφωνήσει μαζί τους, όπως θα φανεί και μέσα από τα αποτελέσματα των πειραμάτων στη συνέχεια.

Στον πίνακα (5.1) παρατίθεται το μέσο ποσοστό λαθών (ως προς τα τμήματα (folds) της διασταυρωμένης επικύρωσης) που κάνουν οι ταξινομητές των δύο αλγορίθμων, είτε μόνο ο ένας εκ των δύο ή και οι δύο μαζί. Τα νούμερα αναφέρονται στους ταξινομητές με παραμέτρους που έδωσαν τη βέλτιστη παρατηρηθείσα επίδοση για λ=1 και 9. Τα στοιχεία για λ=999 παραλείπονται, καθώς αποφασίστηκε να μη γίνουν πειράματα για το αυστηρό σενάριο χρήσης του φίλτρου. Αυτό έγινε λόγω του ότι για αυτό το σενάριο ο NB δε φαίνεται να καταφέρνει να περάσει τη βάση, με αποτέλεσμα η αξία του ως μέλους επιτροπής να καθίσταται αμφίβολη.

Από τον πίνακα φαίνεται πως τα κοινά λάθη των δύο ταξινομητών είναι αρκετά λιγότερα κατά μέσο όρο από το άθροισμα αυτών που κάνει μόνο ο ένας εκ των δύο. Τα περιθώρια υψηλής απόδοσης, λοιπόν, είναι μεγάλα για μια επιτροπή που θα μάθει να επιλέγει το σωστό μέλος ανά περίπτωση όταν αυτά διαφωνούν, έστω κι αν αποτυγχάνει σε όλα τα στιγμιότυπα που αποτυγχάνουν και τα δύο μέλη μαζί.

λ	**Μηνύματα**	***k*-NN μόνο**	**NB μόνο**	***k*-NN είτε NB**	***k*-NN και NB**
1	*Νόμιμα*	0,29%	0,37%	**0.66%**	**0,08%**
	Spam	5,41%	6,86%	**12.27%**	**8,52%**
	Όλα	1,14%	1,45%	**2.59%**	**1,49%**
9	*Νόμιμα*	0,08%	0,25%	**0.33%**	**0,08%**
	Spam	9,56%	9,56%	**19.12%**	**10,19%**
	Όλα	1,66%	1,80%	**3.46%**	**1,76%**

Πίνακας 5.1: Ποσοστά λαθών κατάταξης των βέλτιστων ταξινομητών για κάθε αλγόριθμο (*k*-NN και NB) ανά σενάριο (λ) και κατηγορία μηνυμάτων.

5.C) Σχεδιαστικές επιλογές

Λόγω του ότι οι υπό εξέταση αλγόριθμοι μάθησης ήταν μόνο δύο, αποκλείστηκε η λύση της πλειοψηφικής ψηφοφορίας και των αποτιμημένων παραλλαγών της. Μια εναλλακτική πρόταση είναι να κατασκευαστούν πολλοί ταξινομητές με βάση τους δύο αλγορίθμους με διαφοροποίηση κάποιων εκ των σχεδιαστικών παραμέτρων τους, π.χ. τις παραμέτρους της m-εκτίμησης για τον NB ή το *k* για τον *k*-NN. Οι ταξινομητές όμως που παράγονται με αυτό τον τρόπο από έναν αλγόριθμο έχουν πιθανότατα έντονη συσχέτιση μεταξύ τους, γεγονός που τους καθιστά ακατάλληλους για μέλη μιας επιτροπής. Η λύση του bagging επίσης απορρίφθηκε, αφού όπως αναφέρθηκε, αυτή ενδείκνυται για ασταθείς αλγορίθμους ενώ οι NB και *k*-NN είναι και οι δύο πολύ σταθεροί. Η μέθοδος της προώθησης πάλι, χρησιμοποιεί έναν αλγόριθμο μάθησης μόνο, ενώ στόχος ήταν να διερευνηθεί η δυνατότητα συνδυασμού των NB και *k*-NN μέσα σε μια επιτροπή.

Η λύση που υιοθετήθηκε τελικά ήταν αυτή της συσσώρευσης (stacking), καθώς με μια μικρή επέκταση της περιγραφής που δόθηκε στην προηγούμενη ενότητα, μπορεί να

εφαρμοστεί επιτυχώς ακόμα και για δύο μόλις ταξινομητές-μέλη. Καταρχήν, ακολουθήθηκε η εκδοχή κατά την οποία δε λαμβάνονται υπόψη οι ίδιες οι προβλέψεις των μελών (αν το μήνυμα κατατάσσεται ως spam ή ως θεμιτό), αλλά η εμπιστοσύνη τους (εκφρασμένη ως αριθμός στο [0,1]) ως προς το ενδεχόμενο να είναι το μήνυμα spam. Αυτή η εμπιστοσύνη είναι για τον NB απλά η εκ των υστέρων πιθανότητα να είναι το μήνυμα spam και η οποία υπολογίζεται ούτως ή άλλως από τον αλγόριθμο. Για τον *k*-NN, η αντίστοιχη εμπιστοσύνη δίνεται από το λόγο του βάρους των spam γειτόνων προς το άθροισμα των βαρών όλων των γειτόνων στη γειτονιά, όπου βάρος των γειτόνων για μια κλάση είναι το άθροισμα στον τύπο (2.15) για την κλάση αυτή (βλ. ενότητα (2.B.II)). Στο [Ting & Witten 1999] δημοσιεύονται πειραματικά αποτελέσματα που δείχνουν το προβάδισμα της χρήσης βαθμών εμπιστοσύνης έναντι των απλών προβλέψεων των μελών, τουλάχιστον για την περίπτωση που ο πρόεδρος της επιτροπής κατασκευάζεται από αλγόριθμο βασισμένο σε μνήμη (memory-based), όπως ο *k*-NN που χρησιμοποιήθηκε εδώ.

Η επέκταση έγινε στην αναπαράσταση των δεδομένων του επιπέδου-1 και είναι η εξής: Εκτός της πίστης των δύο απλών μελών της επιτροπής του να είναι το μήνυμα spam, ένα διάνυσμα του επιπέδου-1 θα περιέχει επιπλέον και έναν αριθμό από features σαν αυτά που αποτελούν την αναπαράσταση του επιπέδου-0, δηλαδή features που δηλώνουν την παρουσία ή απουσία μιας λέξης από ένα μήνυμα. Με άλλα λόγια, ο πρόεδρος της επιτροπής δε βασίζεται μόνο στις κρίσεις των δύο μελών, αλλά θεωρεί και (κάποια από) τα "πρωτογενή" δεδομένα. Αυτά στη γενική περίπτωση μπορεί να διαφέρουν από εκείνα που περιέχονται στην αναπαράσταση του επιπέδου-0, δηλ. να αποτελούν υποσύνολο ή υπερσύνολό τους, να επικαλύπτονται μερικώς ή και να είναι τελείως ξένα.

Λόγω της πληθώρας των επιλογών που είναι δυνατές και των χρονικών περιορισμών που υπήρχαν, τα πειράματα που διενεργήθηκαν με stacking δεν είχαν στόχο την, ευριστική έστω, βελτιστοποίηση του τελικού ταξινομητή, όπως τα πειράματα που έγιναν με τον απλό *k*-NN. Ο κύριος στόχος εδώ ήταν να καταδειχθούν οι δυνατότητες της συνδυαστικής χρήσης ταξινομητών μέσω μιας επιτροπής και το συγκριτικό πλεονέκτημα της τελευταίας ως προς κάθε μέλος χωριστά. Διαφορετικές επιλογές, πάντως, είναι πιθανό να οδηγούν σε ακόμα υψηλότερες επιδόσεις.

Μία επιλογή που έγινε προκειμένου να κρατηθεί το πλήθος των πειραμάτων μικρό ήταν να κατασκευαστούν τα μέλη της επιτροπής σύμφωνα με τις "βέλτιστες" τιμές παραμέτρων, όπως αυτές εκτιμήθηκαν από τα προηγούμενα πειράματα. Αν και αυτό ακούγεται αρκετά εύλογο, δεν είναι βέβαιο πως για να μεγιστοποιηθεί η ακρίβεια μιας επιτροπής πρέπει να γίνει το ίδιο και για τα μέλη της. Δεν αποκλείεται καταρχήν μια επιτροπή με υποβέλτιστα μέλη να είναι καλύτερη από μία με βέλτιστα, αν λ.χ. τα μέλη της είναι περισσότερο ανεξάρτητα, διευκολύνοντας έτσι το έργο του προέδρου.

Μια ακόμα επιλογή ήταν πως ο πρόεδρος της επιτροπής θα είναι επίσης ο *k*-NN. Αντίθετα όμως με τη χρήση του ως μέλους, όπου οι παράμετροί του παρέμειναν σταθερές, έγιναν διαφορετικά πειράματα με τον πρόεδρο να παίρνει κάθε συνδυασμό τιμών *k* (από 1 έως 10)

και διαστασιμότητας (από 50 έως 700 features με βήμα 50, συν βέβαια τα δύο features που δηλώνουν το βαθμό εμπιστοσύνης των δύο μελών[*]).

5.C.I) Συσσώρευση διασταυρωμένης επικύρωσης

Δύο ήταν οι μεθοδολογίες πειραματισμού που δοκιμάστηκαν. Η μία είναι αυτή που προτάθηκε από τον Wolpert και ακολουθήθηκε και από άλλους στη συνέχεια (π.χ. [Breiman 1996], [LeBlanc & Tibshirani 1993]). Η μεταφορά της στο συγκεκριμένο πρόβλημα υπό μελέτη και για τις συγκεκριμένες επιλογές που έγιναν περιγράφεται ως εξής:

1. Όπως και στα πειράματα με ένα ταξινομητή, γίνεται διασταυρωμένη επικύρωση. Τα αρχικά δεδομένα χωρίζονται σε 10 (γενικά m) περίπου ίσα τμήματα. Στο βήμα i της διασταυρωμένης επικύρωσης:
   a. Ένα τμήμα είναι το σύνολο ελέγχου TS_i, στο οποίο θα εξεταστεί ο πρόεδρος και τα υπόλοιπα 9 αποτελούν το σύνολο εκπαίδευσης TR_i. Στη συνέχεια γίνεται εσωτερικά στο TR_i διασταυρωμένη επικύρωση 3 (γενικά n) σημείων. Στο βήμα j της εσωτερικής διασταυρωμένης επικύρωσης:
      i. Το 1/3 του TR_i αποτελεί το σύνολο επικύρωσης V_{ij} και τα υπόλοιπα 2/3 το σύνολο εκπαίδευσης T_{ij} των μελών της επιτροπής.
      ii. Τα δύο μέλη της επιτροπής ("βέλτιστοι" *k*-NN και NB) εκπαιδεύονται ξεχωριστά πάνω στο T_{ij} και οι ταξινομητές C_{ij} που κατασκευάζονται εξετάζονται στο V_{ij}.
      iii. Για κάθε στιγμιότυπο του V_{ij} κατασκευάζεται ένα νέο, το οποίο περιλαμβάνει τους βαθμούς εμπιστοσύνης των C_{ij} ως προς το ενδεχόμενο να είναι το στιγμιότυπο spam, συν ένα πλήθος N (καθοριζόμενο έξω από την περιγραφόμενη διαδικασία) από features σαν αυτά του αρχικού στιγμιοτύπου ("πρωτογενή" features). Έτσι σχηματίζεται το σύνολο V'_{ij} (δεδομένα επιπέδου-1).
   b. Η ένωση $TR'_i = \bigcup_{j=1}^{3} V'_{ij}$ των V'_{ij} αποτελεί το τελικό σύνολο εκπαίδευσης του προέδρου για το βήμα i της διασταυρωμένης επικύρωσης. Ο πρόεδρος εκπαιδεύεται στο TR'_i
   c. Τα δύο μέλη της επιτροπής εκπαιδεύονται ξεχωριστά πάνω στο TR_i και οι ταξινομητές C_i που κατασκευάζονται αξιολογούνται στο TS_i. Η διαδικασία που γίνεται στο βήμα (1.a.iii) για την κατασκευή των V'_{ij} από τα V_{ij}, εφαρμόζεται επίσης και για την κατασκευή του TS'_i από το TS_i.
   d. Ο πρόεδρος αξιολογείται στο TS'_i.
2. Η τελική (αποτιμημένη) ακρίβεια υπολογίζεται από το μέσο όρο της ακρίβειας των 10 προέδρων.

[*] Το TiMBL αποτιμάει τα αριθμητικά features αφού τα διακριτοποιήσει σε ένα πλήθος από ισομήκη διαστήματα μεταξύ της ελάχιστης και μέγιστης τιμής τους. Η διακριτοποίηση αυτή όμως δε χρησιμοποιείται στον υπολογισμό της απόστασης. Η τελευταία υπολογίζεται για δύο αριθμητικές τιμές x και y ως: $d(x,y) = \frac{|x-y|}{\max - \min}$ όπου max και min η μέγιστη και η ελάχιστη δοθείσα τιμή του feature αντίστοιχα. Για λεπτομέρειες, δείτε το εγχειρίδιο του TiMBL ([Daelemans et al. 2000]).

Αν και η διαδικασία φαίνεται πολύπλοκη, στην πραγματικότητα δεν είναι. Αυτό που χρειάζεται εξήγηση είναι ο ρόλος της εσωτερικής διασταυρωμένης επικύρωσης. Αντίθετα με την εξωτερική που γίνεται για λόγους στατιστικής ορθότητας και δεν αποτελεί μέρος του stacking, η εσωτερική έχει σκοπό την κατασκευή των δεδομένων του επιπέδου-1. Εφόσον ο πρόεδρος της επιτροπής χρειάζεται να εκπαιδευθεί πάνω στις κρίσεις των μελών, το σύνολο εκπαίδευσης πρέπει να χωριστεί σε ένα τμήμα εκπαίδευσης των μελών και ένα τμήμα εξέτασης των μελών (επικύρωσης). Αν αυτός ο διαμερισμός γίνει μία μόνο φορά, ο πρόεδρος θα έχει στη διάθεση του μόνο τις προβλέψεις των μελών στο σύνολο επικύρωσης, δηλαδή ένα υποσύνολο των διαθέσιμων δεδομένων. Με τη διασταυρωμένη επικύρωση όμως, διαφορετικοί ταξινομητές επιπέδου-0 (μέλη) εκπαιδεύονται πάνω σε διαφορετικά υποσύνολα των αρχικών δεδομένων και εξετάζονται στα υπόλοιπα, έτσι ώστε τελικά ο πρόεδρος να αποκτήσει το ίδιο πλήθος δεδομένων με τα αρχικά.

Στα βήματα a και b πραγματοποιείται η εκπαίδευση του προέδρου, ενώ στα βήματα c και d γίνεται η αξιολόγησή του. Σημειώνεται πως στο βήμα c, τα μέλη εκπαιδεύονται με βάση όλα τα δεδομένα που έχουν στη διάθεσή τους (TR_i) για να δώσουν τις εκτιμήσεις τους για το TS_i. Ο πρόεδρος, έχοντας εκπαιδευθεί πάνω στο TR'_i στο βήμα b, λαμβάνει υπόψη τις εκτιμήσεις των μελών και κάποια από τα πρωτογενή features και κατατάσσει τα στιγμιότυπα του TS_i.

Ονομάζουμε τη μεθοδολογία αυτή *συσσώρευση διασταυρωμένης επικύρωσης* (*cross-validation stacking,* για συντομία *CV συσσώρευση*), γιατί τα σύνολα εκπαίδευσης TR'_i του προέδρου προκύπτουν μετά από διασταυρωμένη επικύρωση πάνω στα αρχικά σύνολα εκπαίδευσης TR_i.

Το πλεονέκτημα αυτής της CV συσσώρευσης είναι πως ο πρόεδρος εκπαιδεύεται με τόσα στιγμιότυπα όσα θα είχε στη διάθεσή του και αν δεν γινόταν stacking, και όχι λιγότερα. Ωστόσο ένα πρόβλημα που ενδεχομένως δημιουργείται είναι το εξής: Τα δεδομένα του επιπέδου-1 V'_{ij} προκύπτουν από μέλη που έχουν εκπαιδευθεί σε μικρότερα σύνολα εκπαίδευσης από εκείνα που έχουν στη διάθεσή τους στη φάση κατάταξης. Για την περίπτωση διασταυρωμένης επικύρωσης 3 σημείων που γίνεται στην εργασία, τα μέλη αρχικά εκπαιδεύονται στα 2/3 των δεδομένων εκπαίδευσης για να παράγουν τα στιγμιότυπα επιπέδου-1, ενώ στη φάση κατάταξης του συνόλου ελέγχου, εκπαιδεύονται πάνω σε όλα τα διαθέσιμα δεδομένα. Αν η απόδοση των μελών είναι σημαντικά χαμηλότερη στην πρώτη περίπτωση, υπάρχει ο κίνδυνος να υποεκτιμήσει ο πρόεδρος τα μέλη του και να διαφωνεί μαζί τους περισσότερες φορές απ΄ όσο αν ήξερε την απόδοσή τους με μεγαλύτερα σύνολα εκπαίδευσης, όπως αυτά που είναι διαθέσιμα κατά την εξέταση της επιτροπής. Το διάγραμμα (4-8) φαίνεται να δικαιολογεί αυτή την ανησυχία για τον *k*-NN, αφού η απόδοσή του είναι αισθητά χαμηλότερη, ειδικά για λ=1, όταν εκπαιδευτεί με τα 2/3 των διαθέσιμων δεδομένων εκπαίδευσης, ενώ παρόμοια είναι και η εικόνα για τον NB, όπως παρουσιάζεται στο [Androutsopoulos et al. 2000a].

Επίσης είναι δυνατόν να αλλάξει η σχετική απόδοση των μελών για διαφορετικά μεγέθη εκπαίδευσης, δηλαδή ο πρόεδρος να μάθει να εμπιστεύεται το ένα μέλος για μια κατηγορία στιγμιοτύπων, αλλά το άλλο μέλος να είναι πιο αξιόπιστο αν δοθούν περισσότερα στιγμιότυπα εκπαίδευσης. Τέλος, στην περίπτωση που κάποια από τα μέλη προέρχονται από ασταθείς αλγόριθμους (π.χ. δέντρα απόφασης), η ένωση των V'_{ij} που σχηματίζει το σύνολο

TR'_i εκπαίδευσης του προέδρου ενδέχεται να είναι αρκετά ασυνεπής, αφού τα V'_{ij} προέρχονται από ταξινομητές εκπαιδευμένους σε διαφορετικά, έστω και επικαλυπτόμενα, σύνολα εκπαίδευσης.

5.C.II) Συσσώρευση δείγματος ελέγχου

Λόγω των πιθανών προβλημάτων της CV συσσώρευσης, δοκιμάστηκε και μια δεύτερη μεθοδολογία. Αναλυτικά τα βήματα είναι τα εξής:

1. Γίνεται διασταυρωμένη επικύρωση 10 σημείων. Στο βήμα i της διασταυρωμένης επικύρωσης:
   a. Όπως και στη CV συσσώρευση, ένα τμήμα είναι το σύνολο ελέγχου TS_i και τα υπόλοιπα 9 αποτελούν το TR_i. Πάνω στο TR_i γίνεται νέα διασταυρωμένη επικύρωση 3 σημείων. Στο βήμα j της εσωτερικής διασταυρωμένης επικύρωσης:
      i. Το TR_i χωρίζεται σε V_{ij} και $T_{ij} = TR_i - V_{ij}$, τα ίδια που χρησιμοποιήθηκαν στη CV συσσώρευση.
      ii. Τα δύο μέλη της επιτροπής εκπαιδεύονται ξεχωριστά πάνω στο T_{ij} και οι ταξινομητές C_{ij} που κατασκευάζονται εξετάζονται στο V_{ij}.
      iii. Κατασκευάζεται το V'_{ij} από το V_{ij}, όπως και στη CV συσσώρευση.
      iv. Ο πρόεδρος εκπαιδεύεται στο V'_{ij}.
      v. Οι δύο ταξινομητές C_{ij} αξιολογούνται στο TS_i. Το TS'_i κατασκευάζεται από το TS_i όπως το V'_{ij} από το V_{ij} στο βήμα (1.a.iii).
      vi. Ο πρόεδρος αξιολογείται στο TS'_i.
2. Η τελική (αποτιμημένη) ακρίβεια υπολογίζεται από το μέσο όρο της ακρίβειας των $10 \times 3 = 30$ προέδρων.

Η μεθοδολογία αυτή διαφέρει σε δύο σημεία από τη CV συσσώρευση. Το ένα είναι πως ο πρόεδρος εκπαιδεύεται πάνω σε ένα μόνο V'_{ij} και όχι στην ένωση όλων των V'_{ij} για κάθε j. Το άλλο σημείο είναι πως οι ίδιοι ταξινομητές C_{ij} χρησιμοποιούνται και για την παραγωγή των δεδομένων επιπέδου-1 (βήμα ii) και για την εκπαίδευση των μελών πριν την εξέταση στο TS_i (βήμα v). Έτσι ο πρόεδρος μαθαίνει πιο αξιόπιστα τις δυνατότητες των μελών για την εξέταση στο σύνολο ελέγχου. Ουσιαστικά, η διαδικασία συσσώρευσης βασίζεται σε μία μόνο διαμέριση του TR_i σε σύνολα εκπαίδευσης T_{ij} και επικύρωσης V_{ij}. Τόσο η εξωτερική, όσο και η εσωτερική διασταυρωμένη επικύρωση δεν αποτελούν μέρος της συσσώρευσης και γίνονται μόνο για λόγους στατιστικής αξιοπιστίας των αποτελεσμάτων, για να ελαττωθεί δηλαδή η πιθανότητα να παρατηρηθούν αποτελέσματα που οφείλονται στον τυχαίο χωρισμό σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Συνολικά προκύπτουν 30 επιτροπές, των οποίων η μέση αποτιμημένη ακρίβεια είναι και η τελική εκτιμώμενη ακρίβεια.

Ονομάζουμε αυτή την παραλλαγή *συσσώρευση δείγματος ελέγχου* (*holdout stacking*), γιατί μοιάζει με την ομώνυμη μέθοδο εκτίμησης (βλ. ενότητα (2.C.III.b)) ως προς το ότι γίνεται μία μόνο διαμέριση του TR_i. Ένα πρακτικό πλεονέκτημά της σε σύγκριση με τη CV συσσώρευση είναι πως απαιτεί λιγότερους υπολογισμούς, αφού στη λειτουργική φάση ενός

συστήματος δεν χρειάζονται οι διασταυρωμένες επικυρώσεις, και κατά συνέπεια η διαδικασία περιορίζεται στα βήματα (i)-(vi).

Το μειονέκτημα της holdout συσσώρεσης είναι πως παρέχει μικρό σύνολο εκπαίδευσης στον πρόεδρο (εδώ, το 1/3 του TR_i), οπότε καθίσταται δύσκολη η επιτυχής εκπαίδευσή του. Ουσιαστικά εμφανίζεται το εξής δίλημμα: Να δωθούν στον πρόεδρο

1) λίγα δεδομένα, αλλά πιο αξιόπιστα όσον αφορά τη συμπεριφορά των μελών (holdout συσσώρευση) ή
2) πολλά δεδομένα, αλλά πιθανόν όχι αρκετά αντιπροσωπευτικά (CV συσσώρευση) ;

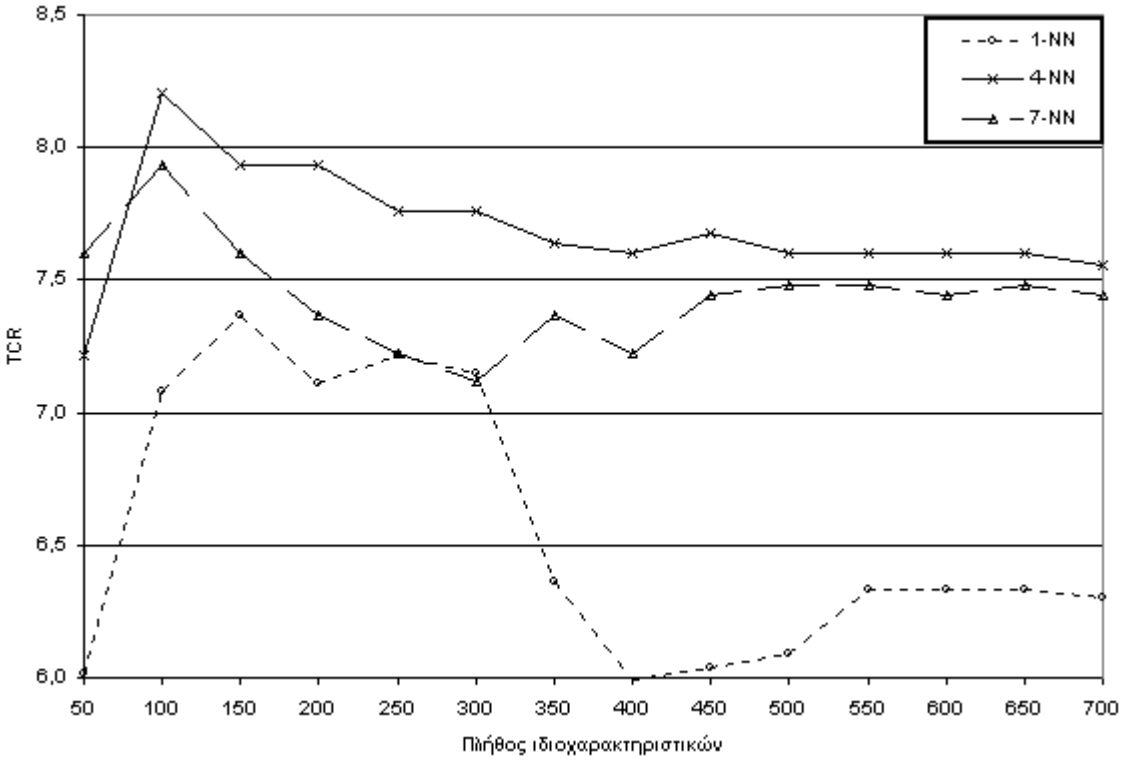
Επειδή δεν υπήρχε εύκολη απάντηση εκ των προτέρων σε αυτό το δίλημμα, δοκιμάστηκαν και οι δύο μεθοδολογίες. Στη βιβλιογραφία χρησιμοποιείται, απ'όσο είναι γνωστό, μόνο η CV συσσώρευση. Ο λόγος είναι πως τα μειονεκτήματά της ελαχιστοποιούνται αν η (εσωτερική) διασταυρωμένη επικύρωση γίνει σε αρκετά σημεία. Πράγματι, καθώς το πλήθος των σημείων αυξάνει, τα σύνολα εκπαίδευσης των μελών τείνουν να είναι τα ίδια, τόσο για την εκπαίδευση του προέδρου, όσο και για την κατάταξη νέων στιγμιοτύπων. Για παράδειγμα, αν αντί για 3 χρησιμοποιούνταν 10 σημεία διασταυρωμένης επικύρωσης (όπως στο [Ting & Witten 1999]), κάθε σύνολο επικύρωσης θα ήταν το 1/10 των συνολικών δεδομένων εκπαίδευσης. Έτσι κάθε μέλος θα εκπαιδευόταν στα υπόλοιπα 9/10, το οποίο είναι αρκετά κοντά στα 10/10 που θα διαθέτει στη φάση κατάταξης. Τα πολλά σημεία επικύρωσης όμως αυξάνουν τις απαιτήσεις σε χρόνο υπολογισμού και χώρο αποθήκευσης στη φάση εκπαίδευσης της επιτροπής. Επειδή στην εργασία ήταν απαραίτητη και η εξωτερική διασταυρωμένη επικύρωση 10 σημείων, για λόγους στατιστικής αξιοπιστίας όπως αναφέρθηκε, επιλέχθηκε το πλήθος των σημείων της εσωτερικής να περιοριστεί στο 3. Η αύξηση τους θα έδινε πιθανότατα ακόμα καλύτερα αποτελέσματα.

5.D) Πειραματικά αποτελέσματα

Στα διαγράμματα (5-1) και (5-2) παριστάνεται για λ=1 και 9, αντίστοιχα, το *TCR* της επιτροπής με holdout συσσώρευση ως συνάρτηση της διαστασιμότητας του προέδρου και για κάποιες τιμές του k. Αποτελέσματα βγήκαν και για τους υπόλοιπους k-NN μέχρι για k=10, αλλά παραλείπονται από το διάγραμμα για λόγους ευκρίνειας. Οι κύριες παρατηρήσεις που έγιναν είναι οι εξής:

- Για λ=1
  - Κορυφαία καθολικά τιμή του *TCR* για τη holdout συσσώρευση είναι το 8.44, η οποία σημειώνεται για τον 5-NN και 100 (πρωτογενή) features. Υπενθυμίζεται πως η καλύτερη επίδοση που παρατηρήθηκε από έναν μόνο ταξινομητή ήταν το 7.18 του 8-NN για 600 features (βλ. πίνακα (4.1)).
  - Το *TCR* είναι άνω του 7.11 για όλα τα k και τις διαστασιμότητες, που ήταν περίπου η βέλτιστη τιμή του απλού k-NN. Έξαίρεση ήταν οι 1-NN και 2-NN για τις περισσότερες διαστασιμότητες και ο 3-NN για 50 features.
  - Οι 1-NN και 2-NN είναι σαφώς χειρότεροι από τους υπόλοιπους και σχεδόν ταυτίζονται μεταξύ τους. Όλοι οι άλλοι κυμαίνονται περίπου μεταξύ των 7-NN και 4-NN που παριστάνονται στο διάγραμμα.

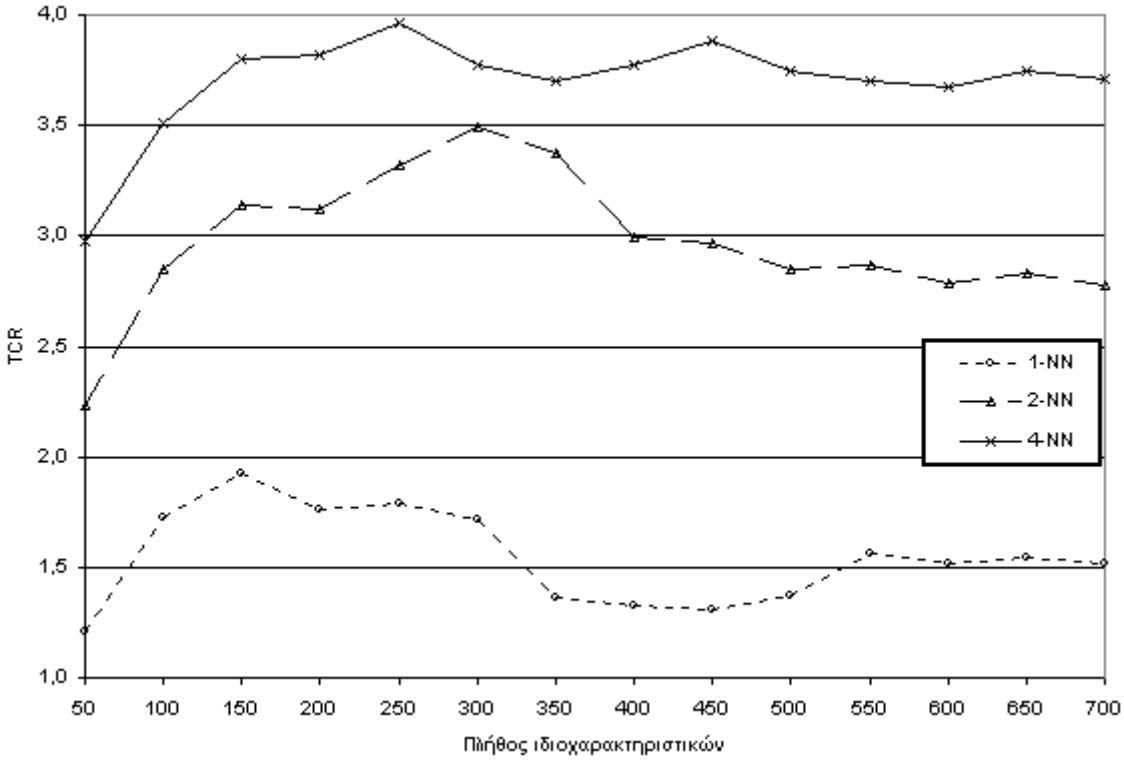
- Ο 4-NN έχει γενικά τη σταθερότερη συμπεριφορά ως προς τη διαστασιμότητα και, μαζί με τον 5-NN, έρχεται πρώτος σε απόδοση.
- Όλες οι καμπύλες κορυφώνονται μεταξύ των 100 και 200 features.



Διάγραμμα 5-1: *TCR* του holdout stacking για λ=1 (t=0.5)
για διάφορους k-NN προέδρους και μέλη τους "βέλτιστους" k-NN και NB.

- Για λ=9
  - Κορυφαία καθολικά τιμή του *TCR* είναι το 3.97, που σημειώνεται για (k=3, 200 features) και (k=4, 250 features). Η καλύτερη τιμή για ένα μεμονωμένο ταξινομητή ήταν το 3.64 του 2-NN για 700 features (βλ. πίνακα (4.1)).
  - Το *TCR* είναι άνω του 3.0 για όλα τα k και τις διαστασιμότητες, ενώ στον απλό k-NN μόνο οι 2-NN και 3-NN κατόρθωσαν να περάσουν το 3.0. Εξαίρεση ήταν ο 1-NN για όλες τις διαστασιμότητες, ο 2-NN για τις περισσότερες και οι 3-NN και 4-NN για 50 features.
  - Ο 1-NN είναι τελευταίος με πολύ μεγάλη διαφορά από τον προτελευταίο, τον 2-NN, ο οποίος με τη σειρά του ξεχωρίζει από τους υπόλοιπους.
  - Ο 4-NN εξακολουθεί να είναι αρκετά σταθερός και συνήθως καλύτερος.
  - Εκτός από τα 50 features, τα οποία είναι λίγα για όλους τα k, δεν ξεχωρίζει κάποιο διάστημα διαστασιμοτήτων που να είναι εξίσου καλό για όλους.

Τα αποτελέσματα της CV συσσώρευσης ήταν ακόμα καλύτερα. Τα αντίστοιχα διαγράμματα δεν διαφέρουν από τα (5-1) και (5-2) ως προς τη γενική μορφή τους και δεν παρουσιάζονται. Αρκούν οι ακόλουθες παρατηρήσεις:



Διάγραμμα 5-2: *TCR* του holdout stacking για λ=9 (t=0.9)
για διάφορους k-NN προέδρους και μέλη τους "βέλτιστους" k-NN και NB.

- Για λ=1
  - Κορυφαία καθολικά τιμή του *TCR* είναι το 8.6. Μάλιστα αυτή σημειώνεται για διαφόρους συνδυασμούς k και διαστασιμότητας.
  - Το *TCR* είναι άνω του 7.18, όση είναι η βέλτιστη τιμή του απλού k-NN, για όλους σχεδόν τους συνδυασμούς k και διαστασιμότητας,.
  - Η κακή εικόνα των 1-NN και 2-NN έναντι των υπολοίπων εμφανίζεται και εδώ. Κατά τα άλλα, δε μπορεί να βγεί κάποιο ασφαλές συμπέρασμα για το ποιά k και διαστασιμότητες είναι καλές.
- Για λ=9
  - Η βέλτιστη τιμή είναι λίγο πάνω από το 4.0 και σημειώνεται για όλες τις καμπύλες, πλην των 1-NN και 2-NN, στα 100 features.
  - Το *TCR* είναι άνω του 3.25 για όλα τα k και τις διαστασιμότητες, ενώ ο απλός k-NN μόνο για k=2 και για τουλάχιστον 600 features κατόρθωσε να περάσει αυτό το όριο. Εξαίρεση ήταν ο 1-NN για όλες τις διαστασιμότητες και ο 2-NN για τις περισσότερες.
  - Γενικά καλύτερος ανεξαρτήτως της διαστασιμότητας είναι ο 3-NN.

Από όλες αυτές τις παρατηρήσεις, δύο συμπεράσματα εξάγονται με ασφάλεια:

1) Η επίδοση είναι σαφώς ανώτερη για τις επιτροπές σε σχέση με καθένα μέλος ξεχωριστά. Η παραπάνω δουλειά που χρειάζεται για την εκπαίδευση των μελών και του προέδρου φαίνεται να αποδίδει. Μεταξύ των δύο μεθοδολογιών, η CV συσσώρευση δείχνει ελαφρώς καλύτερη, αν μάλιστα ληφθεί υπόψη πως η διασταυρωμένη επικύρωση έγινε

μόνο σε 3 σημεία. Περισσότερα σημεία μειώνουν την απόσταση ανάμεσα στη συμπεριφορά των μελών κατά την εκπαίδευση του προέδρου και τη φάση λειτουργίας (κατάταξης), με αποτέλεσμα ο πρόεδρος να έχει πιο ακριβή δεδομένα σχετικά με τις τελικές δυνατότητες των μελών.

2) Και για τις δύο μεθοδολογίες, οι 1-NN και 2-NN είναι σημαντικά χειρότεροι των υπολοίπων *k*-NN ως πρόεδροι της επιτροπής, τουλάχιστον έχοντας ως μέλη τους "βέλτιστους" *k*-NN και NB που δοκιμάστηκαν. Επιπλέον, για λ=1 οι δύο καμπύλες ταυτίζονται σχεδόν απόλυτα, ενώ για λ=9 υπάρχει μεγάλη απόσταση μεταξύ τους, με τον 1-NN να είναι πολύ χειρότερος από τον 2-NN.

Η μεγάλη αυτή υστέρηση των 1-NN και 2-NN από τους άλλους *k*-NN κρίθηκε ενδιαφέρουσα και τα αίτιά της διερευνήθηκαν. Όπως διαπιστώθηκε, τα περισσότερα μηνύματα που ταξινομούνται λάθος από αυτούς, αλλά όχι από τους υπόλοιπους, είναι θεμιτά μηνύματα τα οποία έχουν κοντινότερο γείτονα spam. Τα επιπλέον λάθη είναι δηλαδή $L \rightarrow S$, γεγονός που εξηγεί την πολύ κακή εικόνα του 1-NN για λ=9. Αυτός ο κοντινότερος spam γείτονας είχε ταξινομηθεί λανθασμένα ως θεμιτός και από τα δύο μέλη. Έτσι, η αιτία της αποτυχίας είναι πως ο πρόεδρος επιλέγει να αγνοήσει και τα δύο μέλη μερικές φορές, αν αυτά είχαν ταξινομήσει λάθος κάποιο στιγμιότυπο αρκετά κοντινό (="σχετικό") με το άγνωστο. Διαισθητικά, το να αγνοούνται και τα δύο μέλη είναι παρακινδυνευμένη απόφαση, έστω κι αν δικαιολογείται από το σύνολο εκπαίδευσης, δεδομένου πως σπάνια κάνουν από κοινού λάθος. Ειδικά για λ=9, σίγουρα θα συνέφερε τον πρόεδρο να μην είναι ποτέ αντίθετος και με τα δύο μέλη μαζί όταν αυτά κρίνουν ένα mail ως (πιθανότατα) θεμιτό. Στην επόμενη ενότητα παρουσιάζονται στατιστικά στοιχεία που στηρίζουν αυτό τον ισχυρισμό.

Όσον αφορά τη σύμπτωση των 1-NN και 2-NN για λ=1, αυτή εξηγείται εύκολα. Με την παρουσία των δύο αριθμητικών features, τα οποία δηλώνουν την εμπιστοσύνη των μελών ως προς το να είναι το μήνυμα spam, οι ισοπαλίες διανυσμάτων (πρακτικά οι ταυτίσεις τους) είναι σχεδόν ανύπαρκτες. Έτσι, ο 2-NN σχεδόν πάντα θεωρεί ακριβώς δύο γείτονες που βρίσκονται σε διαφορετική απόσταση μεταξύ τους, από τους οποίους αυτός που υπερισχύει είναι ο κοντινότερος, αυτός δηλαδή που θεωρεί και ο 1-NN.

Για λ=9 το πρόβλημα δεν υφίσταται τόσο έντονα, αφού για να καταταχθεί ένα μήνυμα ως spam από τον 2-NN, δεν αρκεί ο κοντινότερος γείτονας να είναι spam, αλλά πρέπει να έχει και πάνω από 9 φορές μεγαλύτερο βάρος από τον δεύτερο, αν ο τελευταίος είναι θεμιτός. Αυτός ο περιορισμός προφυλάσσει τον 2-NN σε πολλές περιπτώσεις με αποτέλεσμα τη σημαντικά υψηλότερη απόδοση. Ωστόσο, ούτε αυτός είναι αρκετός πάντοτε – τα πειράματα δείχνουν πως χρειάζονται 3 με 4 γείτονες για να μην υπάρχει σημαντική πτώση της spam ορθότητας (*SP*), ή τουλάχιστον η πτώση στην *SP* να ισοσκελίζεται από (πολλαπλάσια) αύξηση της spam ανάκλησης (*SR*).

Στατιστικά στοιχεία των προβλέψεων

Ακολούθως παρατίθενται ορισμένα στατιστικά στοιχεία πάνω στις προβλέψεις της καλύτερης και της χειρότερης επιτροπής σε σχέση με τις προβλέψεις των μελών. Λέγοντας βέλτιστη (αντ. χείριστη) επιτροπή εννοούμε το συνδυασμό k και διαστασιμότητας για τον πρόεδρο που έδωσε το υψηλότερο (αντ. χαμηλότερο) *TCR*. Με αυτό τον τρόπο ίσως φωτιστεί περισσότερο η εξάρτηση της απόδοσης μιας επιτροπής (δηλ. του προέδρου) από τις κρίσεις των μελών. Τα στοιχεία αφορούν τη CV συσσώρευση, για την οποία παρατηρήθηκαν και τα καλύτερα αποτελέσματα.

Στον πίνακα (5.2) δίνονται τα ποσοστά των επιτυχών προβλέψεων της βέλτιστης και της χείριστης επιτροπής σε περίπτωση διαφωνίας των μελών. Φαίνεται λοιπόν πως ακόμα και η χείριστη επιτροπή αποφασίζει να εμπιστευθεί το σωστό μέλος πάνω από τις μισές φορές, όταν υπάρξει διαφωνία μεταξύ τους. Επίσης παρατηρείται πως η βέλτιστη επιτροπή έχει περισσότερες επιτυχίες στον εντοπισμό των θεμιτών μηνυμάτων, ενώ η χείριστη είναι καλύτερη στον εντοπισμό των spam. Έτσι συμβαίνει το "παράδοξο" η χείριστη επιτροπή να επιλύει πιο επιτυχώς τις ισοπαλίες από τη βέλτιστη στο σύνολο των μηνυμάτων για λ=9. Αυτό όμως είναι ελάχιστα σημαντικό για την τελική της απόδοση, αφ'ενός μεν διότι οι διαφωνίες των μελών είναι αρκετά σπάνιες (λιγότερες του 3.5%, σύμφωνα με τον πίνακα 5.1) και αφ'ετέρου επειδή για λ=9 η σωστή κατάταξη των θεμιτών μηνυμάτων έχει πολλαπλάσια αξία από τη σωστή κατάταξη των spam, ενώ η χείριστη επιτροπή είναι καλύτερη μόνο στο τελευταίο.

λ	**Μηνύματα**	**Βέλτιστη επιτροπή**	**Χείριστη επιτροπή**
1	*Νόμιμα*	81,25%	68,75%
	Spam	79,66%	79,66%
	Όλα	80,00%	77,33%
9	*Νόμιμα*	75,00%	62,50%
	Spam	66,30%	86,96%
	Όλα	67,00%	85,00%

Πίνακας 5.2: Ποσοστά επιτυχών προβλέψεων της βέλτιστης και της χείριστης επιτροπής σε περίπτωση διαφωνίας των μελών, ανά σενάριο (λ) και κατηγορία μηνυμάτων.

Στους πίνακες (5.3) και (5.4) δίνονται τα ποσοστά των σωστών και λανθασμένων αποφάσεων της καλύτερης και της χειρότερης επιτροπής, αντίστοιχα, σε περίπτωση συμφωνίας των μελών, η οποία είναι και η συνήθης. Τα ποσοστά αυτά αναλύονται παραπέρα στις περιπτώσεις συμφωνίας και διαφωνίας του προέδρου με τα δύο μέλη.

Οι κύριες παρατηρήσεις που μπορούν να γίνουν βάσει αυτών των στοιχείων είναι οι εξής:

1) Η βέλτιστη επιτροπή τείνει να διαφωνεί πιο σπάνια και με τα δύο μέλη ταυτόγχρονα σε σχέση με τη χείριστη, τόσο στην περίπτωση που τα μέλη συμφωνήσουν σε κατάταξη ως spam, όσο και ως θεμιτό μήνυμα, και για τα δύο σενάρια.

2) Η μεγαλύτερη τάση ανεξαρτητοποίησης της χείριστης επιτροπής από τα μέλη είναι επωφελής για την σωστή κατάταξη των spam. Π.χ. για λ=9, στο 4.88% των περιπτώσεων διαφωνεί και δικαιώνεται και μόλις στο 0.51% διαφωνεί και αποτυγχάνει. Και η βέλτιστη επιτροπή έχει συνολικά κέρδος από τη διαφωνία με τα δύο μέλη, αλλά μικρότερο: 2.57% επιτυχίες έναντι 0.77% αποτυχίες.
3) Η κατάσταση αντιστρέφεται πλήρως όσον αφορά τη σωστή κατάταξη των θεμιτών μηνυμάτων. Εδώ τα δύο μέλη δεν κάνουν ποτέ λάθος από κοινού (στα συγκεκριμένα πειράματα), οπότε θα συνέφερε μια επιτροπή να μην είναι ποτέ αντίθετη αν τα δύο μέλη αποφανθούν πως ένα μήνυμα είναι θεμιτό. Αυτό συμβαίνει με τη βέλτιστη επιτροπή, η οποία σχεδόν ποτέ δε διαφωνεί (0.04%). Αντίθετα, η χείριστη επιτροπή διαφωνεί και χάνει το 0.88% των περιπτώσεων για λ=1 και το 1.08% για λ=9, χωρίς ποτέ να βγει κερδισμένη από τη διαφωνία της. Ειδικά για λ=9, η απώλεια παραπάνω από 1% θεμιτών μηνυμάτων αποδεικνύεται μοιραία για την τελική απόδοση της επιτροπής, πέφτοντας χαμηλότερα και από αυτή των δύο μέλων.

λ	**Μηνύματα**	**Σωστή απόφαση**		**Λανθασμένη απόφαση**	
		Συμφωνεί	**Διαφωνεί**	**Συμφωνεί**	**Διαφωνεί**
1	*Νόμιμα*	99,87%	0,00%	0,08%	0,04%
	Spam	90,28%	0,71%	9,00%	0,00%
	Όλα	98,44%	0,11%	1,42%	0,04%
9	*Νόμιμα*	99,88%	0,00%	0,08%	0,04%
	Spam	86,63%	2,57%	10,03%	0,77%
	Όλα	98,03%	0,36%	1,47%	0,14%

Πίνακας 5.3: Αποφάσεις της βέλτιστης επιτροπής σε περίπτωση συμφωνίας των μελών, ανά σενάριο (λ) και κατηγορία μηνυμάτων.

λ	**Μηνύματα**	**Σωστή απόφαση**		**Λανθασμένη απόφαση**	
		Συμφωνεί	**Διαφωνεί**	**Συμφωνεί**	**Διαφωνεί**
1	*Νόμιμα*	99,04%	0,00%	0,08%	0,88%
	Spam	89,34%	3,08%	6,64%	0,95%
	Όλα	97,59%	0,46%	1,06%	0,89%
9	*Νόμιμα*	98,84%	0,00%	0,08%	1,08%
	Spam	86,89%	4,88%	7,71%	0,51%
	Όλα	97,17%	0,68%	1,15%	1,00%

Πίνακας 5.4: Αποφάσεις της χείριστης επιτροπής σε περίπτωση συμφωνίας των μελών, ανά σενάριο (λ) και κατηγορία μηνυμάτων.

5.Ε) Σύγκριση καλύτερων επιδόσεων

Στους πίνακες (5.5) και (5.6) συνοψίζονται τα καλύτερα αποτελέσματα από όσα πειράματα έχουν γίνει στη συλλογή Ling-Spam για καθεμιά από τις τεχνικές που χρησιμοποιήθηκαν, για λ=1 και 9, αντίστοιχα[*]. Και για τα δύο αυτά σενάρια, η CV συσσώρευση πέτυχε το υψηλότερο *TCR*, με μικρή διαφορά ακολουθεί η holdout συσσώρευση, τρίτος έρχεται ο *k*-NN και τελευταίος ο NB.

Αλγόριθμος	**Spam Ανάκληση (*SR*)**	**Spam Ορθότητα (*SP*)**	***TCR***
NB	84,43%	97,41%	5,60
k-NN	88,60%	97,39%	7,18
Holdout stacking	91,71%	96,45%	8,44
CV stacking	89,63%	98,66%	8,60

Πίνακας 5.5: Βέλτιστες επιδόσεις των διενεργηθέντων πειραμάτων για λ=1 ανά αλγόριθμο μάθησης.

Αλγόριθμος	**Spam Ανάκληση (*SR*)**	**Spam Ορθότητα (*SP*)**	***TCR***
NB	80,90%	98,32%	3,10
k-NN	81,93%	98,79%	3,64
Holdout stacking	84,23%	98,85%	3,98
CV stacking	84,84%	98,80%	4,08

Πίνακας 5.6: Βέλτιστες επιδόσεις των διενεργηθέντων πειραμάτων για λ=9 ανά αλγόριθμο μάθησης

Για λ=1, η holdout συσσώρευση είναι πρώτη ως προς την *SR*, υπερβαίνοντας κατά 2.1% περίπου τη CV συσσώρευση, 3.1% τον *k*-NN και 7.3% τον NB. Από την άλλη, έχει τη χειρότερη *SP*, 2.2% μικρότερη ως προς τη CV συσσώρευση και 1.25% περίπου μικρότερη από τους *k*-NN και NB. Η CV συσσώρευση πετυχαίνει εδώ την υψηλότερη *SP* και τη δεύτερη καλύτερη *SR*. Για λ=9, οι δύο επιτροπές έχουν παρόμοια *SR*, αφήνοντας πίσω τον *k*-NN και τον NB, ενώ και οι τέσσερις τεχνικές έχουν πολύ υψηλή *SP*.

Γενικά πάντως, η holdout συσσώρευση φαίνεται να αποτελεί ανταγωνιστική λύση ως προς τη CV συσσώρευση για πρακτικές εφαρμογές, καθώς απαιτεί υποπολλαπλάσιο χρόνο εκπαίδευσης, χωρίς ιδιαίτερα χαμηλότερη ακρίβεια κατάταξης.

Από τα αποτελέσματα αυτά φαίνεται επίσης πως η διόρθωση της spam ορθότητας που συμβαίνει για λ=9 είναι αρκετά μικρότερη από τη μείωση της spam ανάκλησης που προκαλεί η επιφυλακτικότητα του φίλτρου. Αυτό οφείλεται καταρχήν στο ότι η *SP* είναι ήδη πολύ υψηλή για λ=1 και τα περιθώρια βελτίωσής της είναι μικρά. Ένας ακόμα λόγος είναι πως η κατηγορία των θεμιτών μηνυμάτων έχει ούτως ή άλλως προβάδισμα στο να επιλεγεί για ένα αυθαίρετο μήνυμα, καθώς αποτελεί την κλάση πλειοψηφίας (“προκαθορισμένη” κλάση –

"default" class). Μεγαλύτερες τιμές του λ ενισχύουν ακόμα περισσότερο αυτό το προβάδισμα, κάνοντας διστακτικότερο τον ταξινομητή να μπλοκάρει κάποιο μήνυμα με επακόλουθο την αισθητή πτώση της *SR*. Ειδικά όσον αφορά τη CV συσσώρευση, η *SP* βελτιώνεται ελάχιστα (λιγότερο από 0.15%) σε σχέση με την χειροτέρευση της *SR* κατά 4.8% περίπου. Αν αυτό δεν είναι συμπτωματικό για τη συγκεκριμένη συλλογή αλλά συμβαίνει γενικότερα για ένα σύστημα, στην πράξη είναι ασύμφορη η αντιστοίχηση του λ=9 σε κάποιο σενάριο χρήσης. Σε μια τέτοια περίπτωση θα πρέπει να αντιστοιχήσουμε το λ=1 στο σενάριο αν η σημειωθείσα *SP* κρίνεται αποδεκτή, διαφορετικά πρέπει να επιλέξουμε αρκετά μεγαλύτερο λ, έτσι ώστε να επιτευχθεί η ελάχιστη επιδιωκόμενη ορθότητα, αποδεχόμενοι όμως και την, ενδεχομένως σημαντική, υποβάθμιση της ανάκλησης που αναπόφευκτα θα επέλθει.

Έλεγχος στατιστικής σημαντικότητας

Στην ενότητα (2.C.III.b) θίχτηκε το ζήτημα της σύγκρισης διαφορετικών συστημάτων ως προς τις επιδόσεις τους και ο ρόλος που παίζει η στατιστική στην εκτίμηση της σημαντικότητας των διαφορών στις εν λόγω επιδόσεις. Στα πλαίσια της εργασίας έγινε έλεγχος στατιστικής σημαντικότητας των διαφορών μεταξύ των καλύτερων ταξινομητών κάθε μεθόδου που δοκιμάστηκε (πίνακες 5.5-5.6) και τα αποτελέσματα του ελέγχου παρατίθενται στους πίνακες (5.7) και (5.8) για λ=1 και 9, αντίστοιχα.

Υπόθεση	**Εμπιστοσύνη**	**Στατιστικά σημαντική ;**
k-NN > NB	87,80%	ΟΧΙ
CV stacking > NB	*98,71%*	*ΝΑΙ*
CV stacking > *k*-NN	92,72%	ΟΧΙ
CV stacking > Holdout stacking	57,56%	ΟΧΙ

Πίνακας 5.7: Έλεγχος στατιστικής σημαντικότητας των διαφορών των αλγορίθμων ως προς την WAcc για λ=1 (μονόπλευρη κατά ζεύγη δοκιμή-t / όριο σημαντικότητας p=0,95).

Υπόθεση	**Εμπιστοσύνη**	**Στατιστικά σημαντική ;**
k-NN > NB	82,60%	ΟΧΙ
CV stacking > NB	89,37%	ΟΧΙ
CV stacking > *k*-NN	74,24%	ΟΧΙ
CV stacking > Holdout stacking	56,24%	ΟΧΙ

Πίνακας 5.8: Έλεγχος στατιστικής σημαντικότητας των διαφορών των αλγορίθμων ως προς την WAcc για λ=9 (μονόπλευρη κατά ζεύγη δοκιμή-t / όριο σημαντικότητας p=0,95)

[*] Με προεπεξεργασία που περιλαμβάνει λημματοποίηση και δεν κάνει χρήση λίστας απομάκρυνσης λέξεων (stop-list).

Η μέθοδος ελέγχου που χρησιμοποιήθηκε ήταν η *κατά ζεύγη δοκιμή-t* (*matched-pair t-test*). Η δοκιμή-t αποδίδει την πιθανότητα δυο δείγματα να προέρχονται από δύο ίδιους πληθυσμούς με την ίδια μέση τιμή, υποθέτοντας πως οι πληθυσμοί ακολουθούν την κατανομή t. Η τελευταία χρησιμοποιείται στον έλεγχο υποθέσεων σε σύνολα δεδομένων λίγων δειγμάτων και προσεγγίζει την κανονική κατανομή καθώς το πλήθος των δειγμάτων τείνει στο άπειρο. Η κατά ζεύγη εκδοχή της μεθόδου είναι κατάλληλη όταν υπάρχει ένα-προς-ένα αντιστοιχία μεταξύ των τιμών των δειγμάτων. Στη συγκεκριμένη περίπτωση, η αντιστοιχία αφορά την επίδοση δύο υπό σύγκριση αλγορίθμων στο ίδιο τμήμα (fold) της διασταυρωμένης επικύρωσης, δηλαδή όταν έχουν εκπαιδευθεί πάνω στο ίδιο σύνολο εκπαίδευσης και εξεταστεί πάνω στο ίδιο σύνολο ελέγχου. Επίσης, η εκτίμηση έγινε βάσει της μονόπλευρης κατανομής t, διότι αυτό που ελέγχεται είναι αν ο κατά μέσο όρο καλύτερος ταξινομητής είναι στατιστικά σημαντικά καλύτερος από τον άλλο και όχι αν απλά ανήκει σε διαφορετική κατανομή από αυτόν, οπότε θα χρησιμοποιούνταν δίπλευρη. Ως μέτρο επίδοσης δεν επιλέχθηκε το *TCR*, λόγω της μη γραμμικότητάς του ως προς τις προβλέψεις των ταξινομητών [Yeh 2000], αλλά η αποτιμημένη ακρίβεια WAcc (βλ. (3.C)) η οποία είναι γραμμική.Σε αυτούς τους πίνακες φαίνεται πως δυστυχώς όλες σχεδόν οι διαφορές μεταξύ των "βέλτιστων" ταξινομητών θεωρούνται ασήμαντες στατιστικά, αν ως κατώφλι εμπιστοσύνης θεωρηθεί το σύνηθες 95% (ισοδύναμα 5% κατώφλι απόρριψης της μηδενικής υπόθεσης). Μοναδική εξαίρεση είναι η υπεροχή της συσσώρευσης (και των δύο εκδοχών της) έναντι του NB για λ=1. Η πλέον ασήμαντη διαφορά είναι μεταξύ των δύο παραλλαγών της συσσώρευσης, ενώ ούτε ο *k*-NN κατάφερε να αναδειχθεί σημαντικά καλύτερος του NB.

Η τελική απόφαση για την υπεροχή του ενός ή του άλλου αλγορίθμου γίνεται ακόμα πιο δύσκολη εξαιτίας της δυσκολίας για την επιλογή του βέλτιστου ταξινομητή σε κάθε περίπτωση. Δεδομένου πως η βέλτιστη ρύθμιση των παραμέτρων είναι τόσο πιο δύσκολη, όσο πιο πολλές και αλληλοεξαρτώμενες είναι αυτές, είναι λογικό να αναμένεται πως η ρύθμιση του NB στην πράξη θα βρίσκεται πιο κοντά στη βέλτιστη του απ' όσο η αντίστοιχη για τον *k*-NN. Και αυτό γιατί για τον *k*-NN πρέπει να επιλεχθεί το *k* και οι μέθοδοι αποτίμησης χαρακτηριστικών και γειτόνων, επιπλέον της αναπαράστασης και της μεθόδου επιλογής χαρακτηριστικών που αποτελεί παράμετρο για κάθε σύστημα. Ακόμα κι αν δεχθούμε πως η συνάρτηση πληροφοριακού κέρδους (*IG*) είναι η καλύτερη επιλογή για αποτίμηση των features ανεξαρτήτως των υπολοίπων παραμέτρων, όπως και η $1/d^3$ για αποτίμηση γειτόνων, οι καλύτερες επιλογές για τη διαστασιμότητα και το *k* όπως εκτιμήθηκαν από τα αποτελέσματα των πειραμάτων στο κεφάλαιο 4 δεν πείθουν ως οι πραγματικά βέλτιστες για το πρόβλημα γενικά, και όχι απλά για τη συγκεκριμένη συλλογή και διαμέριση που έγινε κατά τη διασταυρωμένη επικύρωση. Εξάλλου, όσες περισσότερες παραμέτρους έχει ένας αλγόριθμος, τόσο περισσότερο ευάλωτος είναι στο πρόβλημα του υπερταιριάσματος με τα δεδομένα εκπαίδευσης (overfitting). Έτσι στην πράξη δεν αποκλείεται το προβάδισμα του *k*-NN έναντι του NB να είναι μικρότερο απ' όσο υποδεικνύει η σύγκριση των δύο καλύτερων ρυθμίσεων που εντοπίστηκαν.

Η δυσκολία κατασκευής του βέλτιστου ταξινομητή εμφανίζεται εντονότερα όσον αφορά τις επιτροπές ταξινομητών και το stacking ειδικότερα, αφού οι επιλογές εδώ είναι ακόμα

περισσότερες, όπως περιγράφηκε στην ενότητα (5.Α). Το θετικό στοιχείο εδώ όμως είναι πως, με την εξαίρεση των 1-NN και 2-NN που αναλύθηκε, οι παραχθέντες ταξινομητές ήταν συνήθως καλύτεροι κατά μέσο όρο από κάθε μέλος ξεχωριστά, ανεξαρτήτως της διαστασιμότητας και του k του προέδρου. Επομένως, αν και είναι πολύ δύσκολο να κατασκευαστεί η βέλτιστη επιτροπή, υπάρχει μεγάλη πιθανότητα να σχηματίσουμε μία μέτρια, αποτελεσματικότερη όμως σε σύγκριση με κάθε μέλος χωριστά. Υπενθυμίζεται ξανά πως η CV συσσώρευση έγινε με διασταυρωμένη επικύρωση σε 3 μόνο σημεία – στην πράξη περισσότερα σημεία ενδείκνυνται για υψηλότερες επιδόσεις της επιτροπής.

6) ΑΝΑΚΕΦΑΛΑΙΩΣΗ

Στην εργασία αυτή μελετήθηκε η εφαρμογή του αυτόματου φιλτραρίσματος μηνυμάτων ηλεκτρονικού ταχυδρομείου, η οποία επιχειρεί να αντιμετωπίσει το πρόβλημα της συνεχώς αυξανόμενης χρήσης του ηλεκτρονικού ταχυδρομείου ως μέσο διαφήμισης προϊόντων και υπηρεσιών, δίχως τη συγκατάθεση του χρήστη. Η συγκεκριμένη εφαρμογή αποτελεί μία ειδική περίπτωση προβλήματος αυτόματης κατηγοριοποίησης κειμένου (ΑΚΚ), τεχνολογική περιοχή που έχει δώσει ένα πλήθος από διαφορετικές εφαρμογές και αναμένεται να επεκταθεί ακόμα περισσότερο εμπορικά στο άμεσο μέλλον. Η ΑΚΚ έχει δεχθεί και συνεχίζει να δέχεται μεγάλη επίδραση από το χώρο της μηχανικής μάθησης, ενός επιστημονικού και τεχνολογικού πεδίου που έχει σκοπό τη δημιουργία αυτόματων συστημάτων ικανών να μαθαίνουν επαγωγικά μέσα από ένα σύνολο δεδομένων εκπαίδευσης. Στο πρώτο μέρος της εργασίας έγινε μια εισαγωγική παρουσίαση της ΑΚΚ, της μηχανικής μάθησης και της εφαρμογής της δεύτερης στην πρώτη.

Στη συνέχεια μοντελοποιήθηκε το πρόβλημα του φιλτραρίσματος ηλεκτρονικών μηνυμάτων. Προτάθηκαν τρία ενδεικτικά σενάρια χρήσης του φίλτρου, διαφοροποιούμενα μεταξύ τους ως προς το βαθμό επιφυλακτικότητας κατάταξης ενός μηνύματος ως ανεπιθύμητου. Εισάχθηκαν μέτρα απόδοσης με βάση το κόστος λανθασμένων κατατάξεων ανά σενάριο χρήσης. Αυτά τα μέτρα χρησιμοποιήθηκαν για την αξιολόγηση των παραχθέντων ταξινομητών, τόσο αυτών που προέκυψαν με βάση τον απλοϊκό ταξινομητή Μπαίυζ (NB) σε προηγούμενα πειράματα, όσο και με βάση τον αλγόριθμο των *k*-κοντινότερων γειτόνων (*k*-NN) και τη συσσώρευση ταξινομητών (stacking) που διερευνήθηκαν στην εργασία.

Ένα σύστημα αυτόματης κατηγοριοποίησης κειμένου περιλαμβάνει ένα πλήθος από σχεδιαστικές παραμέτρους που επηρεάζουν την τελική του απόδοση. Πλην των παραμέτρων προεπεξεργασίας και αναπαράστασης των μηνυμάτων που θεωρήθηκαν δεσμευμένες στα πειράματα (λημματοποίηση, απουσία stop-list, επιλογή όρων με βάση το *IG*), τέσσερις παράμετροι ερευνήθηκαν κατά τον πειραματισμό με τον *k*-NN: η μέθοδος αποτίμησης των χαρακτηριστικών, η συνάρτηση αποτίμησης των γειτόνων με βάση την απόσταση, το μέγεθος της γειτονίας *k* και η διαστασιμότητα του προβλήματος. Για τις δύο πρώτες μπορεί να υποστηριχθεί πως προέκυψαν σχετικά ασφαλή συμπεράσματα ανεξαρτήτως των υπολοίπων παραμέτρων. Μεταξύ των τριών μέτρων αποτίμησης χαρακτηριστικών που δοκιμάστηκαν, το *IG* είναι γενικά προτιμότερο (τουλάχιστον για δυαδικά features), όπως και μια συνάρτηση αποτίμησης των γειτόνων που μειώνει δραστικά τη σημασία των απομακρυσμένων γειτόνων, όπως η $1/d^3$. Τα συμπεράσματα αυτά ενισχύθηκαν και μέσω ποιοτικών αναλύσεων που παρατέθηκαν. Αντίθετα, δεν προέκυψαν εξίσου ασφαλή πορίσματα ως προς το βέλτιστο μέγεθος γειτονίας και τη βέλτιστη διαστασιμότητα των στιγμιοτύπων. Στην πράξη, μια διαδικασία διασταυρωμένης επικύρωσης που θα εκτιμά τις βέλτιστες τιμές αυτών των παραμέτρων για τα μηνύματα ενός συγκεκριμένου χρήστη και για το επίπεδο ασφαλείας που

αυτός επιθυμεί είναι ίσως αναγκαία πριν την έναρξη λειτουργίας ενός πραγματικού συστήματος.

Τα πειράματα με τις επιτροπές ταξινομητών κατέδειξαν τις υψηλές δυνατότητες αυτής της τεχνικής. Δίχως στην ουσία να επιχειρηθεί εδώ κάποια βελτιστοποίηση, τα αποτελέσματα έδειξαν βελτίωση σε σχέση με την απόδοση κάθε μέλους της ξεχωριστά. Αν και από τα διενεργηθέντα πειράματα η διαφορά δεν κρίνεται στατιστικά σημαντική, υπάρχουν βάσιμες ελπίδες για ακόμα υψηλότερες επιδόσεις αν δοκιμαστούν εναλλακτικές επιλογές ως προς το σχηματισμό της επιτροπής.

6.Α) Προοπτικές

Τα αποτελέσματα των πειραμάτων σε αυτή την εργασία κάνουν εμφανή τα πλεονεκτήματα της χρήσης τεχνικών μηχανικής μάθησης στην κατασκευή αυτόματων συστημάτων κατηγοριοποίησης υψηλής ακρίβειας. Με spam ανάκληση της τάξης του 90% και ορθότητα άνω του 98% αντιμετωπίζουν σε μεγάλο βαθμό το πρόβλημα της μαζικής αποστολής spam μηνυμάτων, και το σημαντικότερο, με ελάχιστο κόπο και ανάγκη δεξιοτεχνίας από την πλευρά του χρήστη, σε αντίθεση με την προσέγγιση της χειρωνακτικής κατασκευής κανόνων βάσει λέξεων και φράσεων-κλειδιών που συνήθως χρησιμοποιείται μέχρι σήμερα.

Ωστόσο, αυτό δε σημαίνει πως είναι ανέφικτη η περαιτέρω βελτίωση του φίλτρου. Οι εναλλακτικές επιλογές που δε δοκιμάστηκαν στην εργασία είναι πολλές και αρκετές απ'αυτές προβάλλουν ως πολλά υποσχόμενες. Ακολούθως αναφέρονται μερικές κατευθύνσεις που αξίζει να διερευνηθούν, χωρίς βέβαια να είναι οι μοναδικές:

- Συμμετοχή ευριστικών χαρακτηριστικών στην αναπαράσταση των μηνυμάτων. Στο [Sahami et al. 1998] προτείνεται ένα σύνολο ευριστικών features, αποτελούμενο από ένα υποσύνολο χειρωνακτικά επιλεγμένων φράσεων-κλειδιών (π.χ. "be over 21", "FREE!", κ.τ.λ.) και ορισμένα μη θεματικά features, όπως η περιοχή (domain) του αποστολέα (π.χ. αν είναι `.edu` ή `.com`), το αν το γράμμα στάλθηκε προσωπικά στον αποστολέα ή σε λίστα αλληλογραφίας, το αν υπάρχουν συνημμένα (τα περισσότερα spam δεν έχουν), το ποσοστό των μη αλφαριθμητικών και κεφαλαίων χαρακτήρων στον τίτλο του μηνύματος (λόγω του ότι πολλά spam μηνύματα έχουν τίτλους όπως "$$$ BIG MONEY $$$"), κ.ο.κ. Η προσθήκη αυτών των ειδικών για τη συγκεκριμένη περιοχή (domain specific) features έδωσε σαφώς καλύτερα αποτελέσματα απ'ό,τι αν στην αναπαράσταση συμμετέχουν απλά μεμονωμένες λέξεις του μηνύματος. Το μειονέκτημα εδώ είναι πως το σύστημα παύει να είναι πλέον πλήρως αυτόματο και γίνεται ημιαυτόματο, αν ο χρήστης μπορεί να επιλέγει την αναπαράσταση χειρωνακτικά, ενώ αν τα ευριστικά αυτά features αποτελούν σταθερές του συστήματος, επανεμφανίζεται το πρόβλημα της δυναμικής προσαρμογής του φίλτρου στα μεταβαλλόμενα με το χρόνο χαρακτηριστικά των spam (αν π.χ. αυτά πάψουν να περιέχουν πολλούς μη αλφαριθμητικούς χαρακτήρες ή αρχίσουν να στέλνονται μαζί με, παραπλανητικές ή μη, επισυνάψεις).
- Χρήση αριθμητικών features (π.χ. *tfidf* – βλ. ενότητα (2.C.I)) αντί δυαδικών που δηλώνουν απλά την παρουσία ή απουσία ενός όρου.

- Μείωση της διαστασιμότητας μέσω εναλλακτικών του *IG* συναρτήσεων φίλτρου. Για παράδειγμα, στο [Mladenić 1998a] υποστηρίζεται πως ένα μέτρο που χρησιμοποιείται συχνά στην περιοχή της ανάκτησης πληροφορίας (ΑΠ),ο *λόγος πιθανοτήτων (odds ratio)*, και κάποια άλλα μέτρα που βασίζονται σε αυτό είναι πιο επιτυχή του *IG* σε μια εφαρμογή πρόβλεψης των συνδέσμων (links) που θα ακολουθήσει ένας χρήστης κατά την περιήγησή του στο Διαδίκτυο, με αλγόριθμο μάθησης τον NB.
- Διαφορετικός αλγόριθμος μάθησης. Ένας πολλά υποσχόμενος αλγόριθμος είναι οι *μηχανές διανυσμάτων υποστήριξης* (*support vector machines- SVM*). Οι SVM είναι μία σύγχρονη τεχνική μηχανικής μάθησης, καλά θεμελιωμένη πάνω στην υπολογιστική θεωρία της μάθησης (computational learning theory) ([Kearns & Vazirani 1994]). Στο [Joachims 1998] επιχειρηματολογείται θεωρητικά και πειραματικά η άποψη πως οι SVM ταιριάζουν ιδιαίτερα σε εφαρμογές κατηγοριοποίησης κειμένου. Στα πλεονεκτήματά τους συγκαταλέγεται η ικανοποιητική ακρίβεια γενίκευσης σε χώρους μεγάλης διαστασιμότητας (χιλιάδων features) η οποία απαλείφει την ανάγκη επιλογής όρων (αν δεν είναι πρόβλημα ο απαιτούμενος αποθηκευτικός χώρος), η ανθεκτικότητά τους σε θορυβώδη δεδομένα και ο αυτόματος συντονισμός των παραμέτρων του αλγορίθμου, χωρίς την ανάγκη εκτίμησης των βέλτιστων τιμών μέσω της αξιολόγησης πάνω σε σύνολα επικύρωσης. Επιπλέον, ο χρόνος εκπαίδευσής τους είναι μικρότερος σε σχέση με άλλους ανταγωνιστικούς αλγορίθμους ανάλογης αποτελεσματικότητας, όπως οι επιτροπές προώθησης (boosting).
- Διαφορετικές επιλογές για επιτροπές συσσώρευσης (stacking). Για παράδειγμα, περισσότερα μέλη προερχόμενα από αλγορίθμους με πολύ διαφορετική επαγωγική προδιάθεση, όπως δέντρα απόφασης, νευρωνικά δίκτυα, κ.α., θα μπορούσαν να χρησιμοποιηθούν μαζί με τους NB και *k*-NN. Επίσης, διαφορετική θα μπορούσε να ήταν η επιλογή του προέδρου της επιτροπής. Στο [Ting & Witten 1999] υποστηρίζεται πειραματικά πως μεταξύ του C4.5 (δέντρα απόφασης) ([Quinlan 1993]), του NB, του IB1 (βασισμένος στη μνήμη αλγόριθμος) ([Aha et al. 1991]) και του MLR (προσαρμογή ενός αλγορίθμου γραμμικής παλινδρόμησης ελαχίστων τετραγώνων), μόνο ο τελευταίος είναι κατάλληλος για πρόεδρος της επιτροπής στις περισσότερες περιοχές (domains) μάθησης που δοκιμάστηκε.
- Καλύτερη πολιτική κατηγοριοποίησης, βασισμένη τόσο στο λ όσο και στον αλγόριθμο μάθησης. Όπως αναφέρθηκε στην ενότητα (3.C), η πολιτική κατηγοριοποίησης που υιοθετήθηκε είναι βέλτιστη αν οι υπολογιζόμενοι βαθμοί βεβαιότητας $W_S(d_j)$ και $W_L(d_j)$ είναι καλές εκτιμήσεις των $P(C = spam \mid d_j)$ και $P(C = legitimate \mid d_j)$, αντίστοιχα. Θα ήταν ενδιαφέρον αν μπορούσε να προκύψει θεωρητικά σχέση μεταξύ των βαθμών βεβαιότητας, όπως υπολογίζονται από κάποιο αλγόριθμο μάθησης (π.χ. *k*-NN) και των παραπάνω πιθανοτήτων. Επίσης, στην περίπτωση που η εκτίμηση των πιθανοτήτων δεν είναι καλή, θα είχε ενδιαφέρον η βέλτιστη επιλογή της $f(\lambda)$, ή ισοδύναμα του κατωφλίου κατάταξης *t*.

Αυτές οι κατευθύνσεις, καθώς και άλλες που παραλείφθηκαν, είναι σίγουρο πως θα εξερευνηθούν τα επόμενα χρόνια, παράλληλα με τη διαφαινόμενη εξέλιξη εμπορικών συστημάτων κατηγοριοποίησης μηνυμάτων ηλεκτρονικού ταχυδρομείου βασισμένων στη μηχανική μάθηση, προορισμένων όχι μόνο για το φιλτράρισμα spam μηνυμάτων αλλά και για την αυτόματη ιεραρχική οργάνωση των μηνυμάτων των χρηστών σε κατηγορίες ([Koller & Sahami 1997]). Η πρόοδος στον ερευνητικό τομέα, έτσι, θα ελαχιστοποιήσει το πρόβλημα των spam e-mails από τεχνικής πλευράς. Ωστόσο, αυτό ίσως δεν είναι αρκετό. Σε περιπτώσεις που απαιτείται απόλυτη ορθότητα όσον αφορά την κατάταξη θεμιτών μηνυμάτων, όπως για λ=999, φαίνεται να είναι απαραίτητη η υιοθέτηση κανονιστικών διατάξεων, όπως η χρήση λιστών έμπιστων και ανέμπιστων αποστολέων, καθώς μοιάζει αδύνατο να υπάρξει εγγύηση πως ένας ταξινομητής δε θα κάτατάξει ποτέ λάθος ένα θεμιτό μήνυμα. Ο συνδυασμός τεχνικών και κανονιστικών μεθόδων είναι αυτός που θα αντιμετωπίσει ικανοποιητικά το πρόβλημα, σύμφωνα με τις απαιτήσεις του κάθε χρήστη.

ΑΝΑΦΟΡΕΣ

Aha, D.W., Kibler D., and Albert M.K. 1991. Instance-based learning algorithms. *Machine Learning*, Vol. 6, pp. 37-66.

Androutsopoulos, I., Koutsias J., Chandrinos K.V., Paliouras G., and Spyropoulos C.D. 2000a. An Evaluation of Naive Bayesian Anti-Spam Filtering. In *Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)*, Barcelona, Spain, pp. 9-17

Androutsopoulos, I, Koutsias, J, Chandrinos, K.V., and Spyropoulos, C.D. 2000b. An experimental comparison of naïve Bayesian and keyword-based anti-Spam Filtering with encrypted personal e-mail messages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, Athens, Greece, pp. 160–167.

Bailey, T., and Jain A.K. 1978. A Note on Distance-Weighted k-Nearest Neighbor Rules. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-8 (4), pp. 311-313.

Becker B., Kohavi, R., Sommerfield D. 1997. Visualizing the Simple Bayesian Classifier. *KDD-97 Workshop on Issues in the Integration of Data Mining and Data Visualization*.

Box, G., Hunter W., and Hunter J. 1978. *Statistics for experimenters*. John Wiley & Sons.

Breiman, L. 1996. Stacked regressions. *Machine Learning*, 24, 49-64.

Cohen, W.W. 1995. Learning to classify English text with ILP methods. In L. De Raedt Ed., *Advances in inductive logic programming*. Amsterdam, NL: IOS Press.

Cohen, W.W., and Singer Y. 1999. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems* 17 ,2,141 –173.

Cranor, L.F. and LaMacchia, B.A. 1998. Spam! *Communications of ACM*, 41(8):74–83.

Daelemans W., Zavrel J., van der Sloot K., and van den Bosch A. 2000. TiMBL: Tilburg Memory Based Learner, version 3.0, Reference Guide. ILK, Computational Linguistics, Tilburg University. http:/ilk.kub.nl/~ilk/papers/ilk0001.ps.gz.

Dagan, I., Karov Y., and Roth D. 1997. Mistake-driven learning in text categorization. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing* (Providence, US, 1997), pp.55 –63.

Deerwester, S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R. 1990. Indexing by Latent Semantic Indexing. *Journal of the American Society for Information Science* 41, 6, 391 –407.

Dietterich, T. G. 1997. Machine Learning Research: Four Current Directions. *AI Magazine* 18 (4), 97-136.

Dietterich, T.G., and Bakiri G. 1995. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2, 263-286.

Dietterich, T.G., and Kong E.B. 1995. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Department of Computer Science, Oregon State University, Corvallis, Oregon.

Domingos, P. and Pazzani M. 1996. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Proceedings of the 13th International Conference on Machine Learning*, pages 105–112, Bari, Italy.

Drucker, H. D. ,Wu D. and Vapnik V. 1999. Support vector machines for spam categorization. *IEEE Transactions On Neural Networks*, 10(5).

Duda, R.O., and Hart, P.E. 1973. Bayes Decision Theory. Chapter 2 in *Pattern Classification and Scene Analysis*, pp. 10–43. John Wiley.

Dudani, S.A. 1976. The Distance-Weighted k-Nearest Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-6, pp. 325-327.

Efron, B., and Tibshirani R. 1993. *An introduction to the bootstrap*. Chapman & Hall.

Forsyth, R.S. 1999. New directions in text categorization. In A. Gammerman Ed*., Causal models and intelligent data management*, pp.151 –185. Heidelberg, DE:Springer.

Freund, Y. and Schapire R.E. 1995. A decision-theoretic generalization of on-line learning and an application to boosting, in *Proceedings of the Second European Conference on Computational Learning Theory*, Springer-Verlag, pp. 23-37.

Friedman, J., Bentley J., and Finkel R. 1977. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software,* 3(3), 209-226.

Fuhr, N. 1985. A probabilistic model of dictionary-based automatic indexing. In *Proceedings of RIAO-85, 1st International Conference "Recherche d'Information Assistee par Ordinateur"* (Grenoble, FR, 1985), pp.207 – 216.

Gale, W.A., Church K.W., and Yarowsky D. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26, 5, 415 –439.

Hall, R.J. 1998. How to Avoid Unwanted Email. *Communications of ACM*, 41(3):88–95.

Hearst, M.A. 1991. Noun homograph disambiguation using local context in large corpora. In *Proceedings of the 7th Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary* (Oxford, UK, 1991), pp.1–22.

Gómez Hidalgo, J.M., Maña Lσpéz, M., Puertas Sanz, E. Combining Text and Heuristics for Cost-Sensitive Spam Filtering. *Fourth Computational Natural Language Learning Worksho*p*, CoNLL-2000,* Lisbon, September 14, 2000.

Hull, D.A., Pedersen J.O., and Schütze H. 1996. Method combination for document filtering. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval* (Zürich, CH, 1996), pp.279 –288.

Joachims, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning* (Nashville, US, 1997), pp.143 – 151.

Joachims, T.1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, DE, 1998), pp.137 – 142.

Kearns, M., and Vazirani U. 1994. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, Massachusetts.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-1995)*, Morgan Kaufmann, pp. 1137–1143.

Kohavi, R. and Bauer E. 1998. An empirical comparison of voting classification algorithms: Bagging, Boosting, and variants. *Machine Learning*, vv, 1-38 (1998).

Kohavi, R., and John G.H. 1998. The wrapper approach. Book chapter in *Feature Selection for Knowledge Discovery and Data Mining* (Kluwer International Series in Engineering and Computer Science), Huan Liu and Hiroshi Motoda, editors.

Koller, D., and Sahami, M. 1997. Hierarchically classifying documents using very few words. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pp. 170-178. Morgan Kaufmann.

Lam, W., Ruiz M.E., and Srinivasan P. 1999. Automatic text categorization and its applications to text retrieval. *IEEE Transactions on Knowledge and Data Engineering*.

Larkey, L.S. 1999. A patent search and classification system. *In Proceedings of DL-99, 4th ACM Conference on Digital Libraries* (Berkeley, US, 1999), pp.179 –187.

Larkey, L.S., and Croft W.B. 1996. Combining classifiers in text categorization. In *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval* (Zürich, CH, 1996), pp.289 –297.

LeBlanc M., and Tibshirani R.1993. Combining estimates in regression and classification. Technical Report 9318, Department of Statistics, University of Toronto.

Lewis, D.D. 1992. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval* (Kobenhavn, DK, 1992), pp.37–50.

Lewis, D.D. 1998. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, Germany, 1998), pp. 4-15.

Lewis, D.D., and Ringuette M. 1994. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. (Las Vegas, US, 1994), pp.81 –93.

Li, Y.H., and Jain A.K. 1998. Classification of text documents. *The Computer Journal* 41, 8, 537 –546.

McCallum, A.K., Rosenfeld R., Mitchell T.M., and Ng A.Y. 1998. Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of ICML-98, 15th International Conference on Machine Learning* (Madison, US, 1998), pp.359 –367.

McLeod J.E.S., Luk A., and Titterington D.M. 1987. A Re-Examination of the Distance-Weighted k-Nearest Neighbor Classification Rule, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-17 (4), pp. 689-696.

Mitchell, T.M. 1997. *Machine Learning*. McGraw-Hill.

Mladenić, D. 1998a. Feature subset selection in text learning. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, DE, 1998), pp.95 –100.

Mladenić, D. 1998b. Turning Yahoo! into an automatic Web page classifier. In *Proceedings of ECAI-98, 13th European Conference on Artificial Intelligence* (Brighton, UK, 1998), pp.473 –474.

Parmanto, B., Munro P.W., and Doyle H.R. 1996. Improving committee diagnosis with resampling techniques. In Touretzky, D. S., Mozer, M. C., & Hesselmo, M. E. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 8, pp. 882-888 Cambridge, MA. MIT Press.

Pomerleau, D.A. 1989. ALVINN: An autonomous land vehicle in a neural network. Technical Report CMU-CS-89-107. Pittsburgh, PA: Carnegie Mellon University.

Quinlan J.R. 1986. Induction of decision trees. *Machine learning*, 1(1), 81-106.

Quinlan J.R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Robertson, S.E., and Harding P. 1984. Probabilistic automatic indexing by learning from human indexers. *Journal of Documentation* 40, 4, 264 –270.

Ruiz, M.E., and Srinivasan P. 1999. Hierarchical neural networks for text categorization. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, US, 1999), pp.281 –282.

Sable, C.L., and Hatzivassiloglou V. 1999. Text-based approaches for the categorization of images. In *Proceedings of ECDL-99, 3rd European Conference on Research and Advanced Technology for Digital Libraries* (Paris, FR, 1999), pp.19 –38.

Sahami, M., Dumais S., Heckerman D., and Horvitz E. 1998. A Bayesian Approach to Filtering Junk E-Mail. *Learning for Text Categorization – Papers from the AAAI Workshop*, pages 55–62, Madison Wisconsin. AAAI Technical Report WS-98-05.

Salton, G., and Buckley C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24, 5, 513–523.

Salton, G., and McGill. M.J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Schapire, R. E. 1990. The strength of weak learnability. *Machine Learning* 5(2), 197-227.

Schapire, R. E., Freund Y., Bartlett P. & Lee W. S. 1997. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Machine Learning: Proceedings of the Fourteenth International Conference*, D.Fisher, ed., Morgan Kaufmann, pp. 322-330.

Schapire, R.E., and Singer Y. 2000. BoosTexter: a boosting-based system for text categorization. *Machine Learning*. Forthcoming.

Schapire, R.E., Singer Y., and Singhal A. 1998. Boosting and Rocchio applied to text filtering. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval* (Melbourne, AU, 1998), pp.215 –223.

Schütze, H., Hull D.A., and Pedersen J.O. 1995. A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval* (Seattle, US, 1995), pp.229 –237.

Sebastiani, F. 1999. Machine learning in automated text categorisation: a survey. Technical Report, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Number IEI-B4-31-1999.

Smyth, P., and Wolpert D. 1997. Stacked density estimation. *Advances in Neural Information Processing systems*.

Tesauro, G. 1995. Temporal difference learning and TD-gammon. *Communications of the ACM*, 38(3), 58-68.

Ting, K.M., and Witten I.H. 1999. Issues in Stacked Generalization. *Journal of Artificial Intelligence Research*, 10 (1999), 271-289.

Tzeras, K., and Hartmann S. 1993. Automatic indexing based on Bayesian inference networks. In *Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval* (Pittsburgh, US, 1993), pp.22 –34.

Wolpert, D. 1992. Stacked generalization. *Neural Networks*, 5 (2), 241-260.

Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval* 1, 1-2, 69 – 90.

Yang, Y., and Liu X. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, US, 1999), pp.42 –49.

Yang, Y., and Pedersen J.O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning* (Nashville, US, 1997), pp.412 – 420.

Yeh, A. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics* (Saarbrücken, DE, 2000), pp. 947-953.