ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
Τηλ.: 265100-8817, Φαξ: 265100-8890 Τ.Θ. 1186, Ιωάννινα 45110

ΠΑΡΟΥΣΙΑΣΗ
ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

ΗΜΕΡΟΜΗΝΙΑ: Τετάρτη, 18 Ιουνίου 2014

ΩΡΑ: 15:00

ΑΙΘΟΥΣΑ: Αίθουσα Σεμιναρίων (ισόγειο Ι1-Ι2)
Κτήριο Τμήματος Μηχανικών Η/Υ και Πληροφορικής

ΟΜΙΛΗΤΗΣ: **Γρηγόριος Τζώρτζης**

Θ έ μ α

«Τεχνικές Ομαδοποίησης Δεδομένων Βασισμένες σε Πίνακες Ομοιότητας»

Επταμελής Εξεταστική Επιτροπή

1. Αριστείδης Λύκας, Καθηγητής του Πανεπιστημίου Ιωαννίνων (**Επιβλέπων**)
2. Κωνσταντίνος Μπλέκας, Επίκουρος Καθηγητής του Πανεπιστημίου Ιωαννίνων
3. Ευάγγελος Καρκαλέτσης, Ερευνητής Α' του ΕΚΕΦΕ «Δημόκριτος»
4. Μιχαήλ Βαζιργιάννης, Καθηγητής του Οικονομικού Πανεπιστημίου Αθηνών
5. Χριστόφορος Νίκου, Αναπληρωτής Καθηγητής του Πανεπιστημίου Ιωαννίνων
6. Αναστάσιος Τέφας, Επίκουρος Καθηγητής του Αριστοτελείου Πανεπιστημίου Θεσ/νίκης
7. Μιχαήλ Τίτσιας, Λέκτορας του Οικονομικού Πανεπιστημίου Αθηνών

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
Τηλ.: 265100-8817, Φαξ: 265100-8890 Τ.Θ. 1186, Ιωάννινα 45110

Περίληψη

Η παρούσα διατριβή μελετά το πρόβλημα της ομαδοποίησης (clustering), που έχει ως στόχο τον διαχωρισμό ενός συνόλου δεδομένων σε ομάδες (clusters), χωρίς τη χρήση επίβλεψης, ώστε τα δεδομένα που ανήκουν στη ίδια ομάδα να είναι όμοια μεταξύ τους και ανόμοια με αυτά των άλλων ομάδων, βάσει ενός μέτρου ομοιότητας/ανομοιότητας. Συγκεκριμένα, η διατριβή επικεντρώνεται στην παρουσίαση μεθόδων ομαδοποίησης που αφορούν τρεις βασικούς θεματικούς άξονες: α) την ομαδοποίηση δεδομένων για τα οποία έχουμε διαθέσιμο μόνο τον πίνακα εγγύτητας και όχι τα ίδια τα δεδομένα (proximity-based clustering), β) την μάθηση με πολλαπλές όψεις (multi-view learning), όπου για τα ίδια δεδομένα έχουμε στη διάθεσή μας πολλαπλές αναπαραστάσεις (όψεις) που προέρχονται από διαφορετικές πηγές ή/και διαφορετικούς χώρους χαρακτηριστικών και γ) την μάθηση με πολλαπλούς πυρήνες (multiple kernel learning), όπου ταυτόχρονα με την ομαδοποίηση θέλουμε να μάθουμε και τον κατάλληλο πυρήνα (kernel) για τα δεδομένα. Συνήθως ο πυρήνας παραμετροποιείται ως ένας συνδυασμός από δοθέντες πυρήνες (basis kernels) και στοχεύουμε στην μάθηση κατάλληλων τιμών για τις παραμέτρους του συνδυασμού.

Αρχικά προτείνεται μια μέθοδος για την αντιμετώπιση του γνωστού προβλήματος της αρχικοποίησης (initialization problem), από το οποίο πάσχει ο αλγόριθμος *k*-means. Συγκεκριμένα, τροποποιούμε το κριτήριο (objective function) του *k*-means έτσι ώστε να δίδεται μεγαλύτερη έμφαση στην ελαχιστοποίηση των ομάδων που στην τρέχουσα επανάληψη εμφανίζουν μεγάλη διακύμανση (intra-cluster variance). Κατ' αυτόν τον τρόπο ο χώρος λύσεων σταδιακά περιορίζεται σε ομάδες που εμφανίζουν παρεμφερή διακύμανση, το οποίο επιτρέπει στη μέθοδό μας να εντοπίζει σε συστηματική βάση λύσεις καλύτερης ποιότητας σε σχέση με τον *k*-means. Επιπλέον, παρουσιάζεται η προσαρμογή της μεθόδου ώστε να μπορεί να εφαρμοστεί για ομαδοποίηση με πίνακα ομοιότητας τροποποιώντας το κριτήριο του αλγορίθμου kernel *k*-means.

Στη συνέχεια, η διατριβή εστιάζεται στο πρόβλημα της ομαδοποίησης με πολλαπλές όψεις. Η βασική συνεισφορά στο αντικείμενο αυτό σχετίζεται με την ανάθεση βαρών στις όψεις, τα οποία μαθαίνονται αυτόματα και τα οποία αντικατοπτρίζουν την ποιότητα των όψεων. Οι υπάρχουσες προσεγγίσεις θεωρούν όλες τις όψεις εξίσου σημαντικές, κάτι που μπορεί να οδηγήσει σε σημαντική μείωση της απόδοσης εάν υπάρχουν εκφυλισμένες όψεις (π.χ. όψεις με θόρυβο) στο σύνολο δεδομένων. Ειδικότερα, παρουσιάζονται για το ανωτέρω πρόβλημα δύο διαφορετικές μεθοδολογίες. Στην πρώτη περίπτωση αναπαριστούμε τις όψεις μέσω κυρτών μικτών μοντέλων (convex mixture models) λαμβάνοντας υπόψη τις διαφορετικές στατιστικές ιδιότητές τους και παρουσιάζουμε έναν αλγόριθμο με βάρη στις όψεις και έναν χωρίς βάρη. Στην δεύτερη περίπτωση αναπαριστούμε την κάθε όψη μέσω ενός πίνακα ομοιότητας (kernel matrix) και μαθαίνουμε ένα συνδυασμό με βάρη από τους πίνακες αυτούς. Το προτεινόμενο μοντέλο διαθέτει μία παράμετρο που ελέγχει την αραιότητα των βαρών, επιτρέποντας την καλύτερη προσαρμογή του συνδυασμού στα δεδομένα.

Η τελευταία ενότητα της διατριβής αφορά στην ομαδοποίηση με πολλαπλούς πυρήνες, όπου συνήθως το κριτήριο που βελτιστοποιείται είναι το εύρος (margin) της λύσης, όπως είναι γνωστό από τον ταξινομητή SVM (support vector machine). Στην προσέγγιση που προτείνεται, βελτιστοποιείται ο λόγος μεταξύ του εύρους και της διακύμανσης (intra-cluster variance) των ομάδων, λαμβάνοντας έτσι υπόψη τόσο τον διαχωρισμό (separability) τους όσο και το πόσο συμπαγείς (compactness) είναι οι ομάδες, το οποίο δύναται να οδηγήσει σε καλύτερες λύσεις. Έχει δειχθεί ότι το εύρος από μόνο του δεν επαρκεί ως κριτήριο για την μάθηση του κατάλληλου πυρήνα, καθότι μπορεί να γίνει αυθαίρετα μεγάλο μέσω μίας απλής κλιμάκωσης (scaling) του πυρήνα. Αντιθέτως, το κριτήριο που προτείνουμε είναι αμετάβλητο (invariant) σε κλιμακώσεις του πυρήνα και, επιπλέον, το ολικό του βέλτιστο είναι αμετάβλητο ως προς τον τύπο της νόρμας που εφαρμόζεται στους περιορισμούς (constraints) των παραμέτρων του πυρήνα. Τα πειραματικά αποτελέσματα επιβεβαιώνουν τις ιδιότητες του κριτηρίου μας, καθώς και τις αναμενόμενες βελτιωμένες επιδόσεις ομαδοποίησης.