Τίτλος Διδακτορικής Διατριβής: Νέος Αλγόριθμος Ομαδοποίησης και Εφαρμογές

Ονοματεπώνυμο: Εμμανουήλ Κ. Οικονομάκης

Τριμελής Συμβουλευτική Επιτροπή

Μιχαήλ Ν. Βραχάτης (Καθηγητής Τμήματος Μαθηματικών, Σχολής Θετικών Σπουδών, Πανεπιστημίου Πατρών)

Γεώργιος Σπύρου (Διευθυντής Ερευνητικής Ομάδας Βιοπληροφορικής Ινστιτούτου Νευρολογίας και Γενετικής Κύπρου)

Γεώργιος Καστής (Ερευνητής Α, Κέντρο Ερευνών Θεωρητικών και Εφηρμοσμένων Μαθηματικών Ακαδημίας Αθηνών)

Περίληψη

Η αξία ενός επιστημονικού πεδίου είναι άμεσα συνυφασμένη με τη συμβολή αυτού στη λοιπή επιστημονική κοινότητα και την κοινωνία γενικότερα. Η ομαδοποίηση δεδομένων αποτελεί ένα τέτοιο πεδίο με εφαρμογές στη βιοϊατρική, την οικονομία κ.ά. Ωστόσο, η ομαδοποίηση δεδομένων δεν παύει να εξελίσσεται και νέοι αλγόριθμοι παρουσιάζονται σε τακτικά διαστήματα. Οι αλγόριθμοι προέρχονται τόσο από την εξέλιξη των ήδη υπάρχοντων αλγορίθμων, όσο και από την ανάγκη να αντιμετωπιστούν συγκεκριμένα προβλήματα.

Η ομαδοποίηση δεδομένων βασίζεται σε πλειάδα αλγορίθμων, οι περισσότεροι από αυτούς μπορούν να διαχωριστούν σε κατηγορίες βάσει των βασικών τους χαρακτηριστικών και του τρόπου με τον οποίο αντιλαμβάνονται τις ομάδες. Οι έννοιες της απόστασης και της πυκνότητας είναι ίσως οι σημαντικότερες στο χώρο της ομαδοποίησης. Από τη μία πλευρά, οι αλγόριθμοι που βασίζονται στην απόσταση κρίνουν αν τα σημεία ενός συνόλου δεδομένων ανήκουν στην ίδια ομάδα με βάση τις μεταξύ τους αποστάσεις. Από την άλλη πλευρά, οι αλγόριθμοι που βασίζονται στην πυκνότητα συνήθως δεν εξετάζουν τα σημεία ξεχωριστά, αλλά μελετάνε τις περιοχές του συνόλου δεδομένων. Με αυτό τον τρόπο, προσδιορίζουν περιοχές αυξημένης πυκνότητας τις οποίες χαρακτηρίζουν ως ομάδες ή μέρη αυτών. Αν και η ομαδοποίηση δεδομένων αποτελείται και από πολλές επιπλέον κατηγορίες όπως αλγόριθμοι βασισμένοι στην Ασαφή Λογική, σε Γκαουσσιανές κατανομές και στην ομοιότητα, κατά βάσιν οι αλγόριθμοι αυτοί μπορούν να ενταχθούν σε κάποια από τις βασικές κατηγορίες που βασίζονται στην απόσταση ή την πυκνότητα. Τέλος, πρέπει να σημειωθεί ότι έχουν προταθεί αλγόριθμοι που συνδυάζουν και τις δύο έννοιες, συνδυάζοντας τις σε ένα βασικό μέτρο που χρησιμοποιείται ώστε να κατασκευάσει τις ομάδες. Ωστόσο, θεωρώντας την απόσταση ως βασικό κριτήριο για την ομαδοποίηση σημείων δημιουργούνται αλγόριθμοι που έμμεσα ή άμεσα συγκρίνουν όλα τα σημεία του συνόλου δεδομένων οδηγούμενοι έτσι σε αυξημένο υπολογιστικό κόστος.

Αντιθέτως, αλγόριθμοι που βασίζονται στην πυκνότητα παρουσιάζουν μία αδυναμία να αντιληφθούν τις συσχετίσεις μεταξύ σημείων που βρίσκονται σε περιοχές ίσης ή παρεμφερούς πυκνότητας.

Η συμβολή αυτής της διατριβής στο πεδίο της ομαδοποίησης δεδομένων είναι η εισαγωγή ενός αλγορίθμου που συνδυάζει τις δύο αυτές έννοιες όχι δημιουργώντας μία νέα έννοια που τις ενοποιεί σε μία, αλλά επιλέγοντας της έννοια που εξυπηρετεί το εκάστοτε βήμα του αλγορίθμου. Συγκεκριμένα, χρησιμοποιεί την πυκνότητα προκειμένου να εντοπίσει τις περιοχές αυξημένης πυκνότητας και συνεπώς τις ομάδες ή τουλάχιστον μέρη αυτών. Αντιθέτως, θέλοντας να εντοπιστούν οι πλησιέστερες ομάδες αξιοποιείται η έννοια της απόστασης.

Ωστόσο, η συμβολή μίας μεθόδου εξαρτάται από την δυνατότητα να «προσφέρει» και σε άλλα επιστημονικά πεδία. Η ομαδοποίηση δεδομένων έχει βρει πρόσφορο πεδίο εφαρμογής στην ιατρική, συγκεκριμένα στη βιοϊατρική, στην προσωποποιημένη ιατρική και στην επεξεργασία ιατρικής εικόνας. Αναλυτικότερα, η ομαδοποίηση δεδομένων αξιοποιείται στη βιοϊατρική και στην προσωποποιημένη ιατρική με στόχο να εντοπίσει συσχετίσεις μεταξύ γονιδίων, ασθενειών και φαρμακευτικών ουσιών. Ειδικότερα στην περίπτωση του καρκίνου και συγκεκριμένα του καρκίνου του μαστού, η περιπλοκότητα των συσχετίσεων που τυχόν υπάρχουν έχουν τέτοιο επίπεδο περιπλοκότητας ώστε η ανάδειξη αυτών να εμφανίζει ακόμα πολύ μεγάλα περιθώρια βελτίωσης.

Συνεπώς, η εφαρμογή της ομαδοποίησης σε προβλήματα των χώρων αυτών αποτελεί μία πρόκληση για τους νέους αλγορίθμους ομαδοποίησης και η παρούσα διατριβή μελετά την εφαρμογή της ομαδοποίησης σε προβλήματα του καρκίνου του μαστού, μελετώντας τη δυναμική του νέου αλγορίθμου που εισάγεται σε αυτή. Για την περίπτωση της βιοϊατρικής, ερευνάται η συμβολή του αλγορίθμου στη ομαδοποίηση συσχετίσεων γονιδίων με βάση την συνέκφραση και τη διαφορική τους έκφραση. Στόχος αυτής της μελέτης είναι ο προσδιορισμός μονοπατιών γονιδίων που να περιγράφουν τα στάδια του καρκίνου του μαστού. Τα αποτελέσματα της εργασίας στο συγκεκριμένο πρόβλημα δικαιολογούν την επιλογή αυτή καθώς πολλά από τα μονοπάτια που προσδιορίστηκαν επιβεβαιώνονται από την υπάρχουσα βιβλιογραφία.

Title of the Thesis: A new clustering algorithm and applications

Full name: Emmanouil K. Ikonomakis

Advisor Committee

Michael N Vrahatis (Professor of the Department of Mathematics, School of Natural Sciences, University of Patras)

George Spyrou (Bioinformatics ERA Chair, The Cyprus Institute of Neurology and Genetic)

Genetic Kastis (Research director, Mathematics Research Center, Academy of Athens)

Abstract

The value of a scientific field is directly interwoven with its contribution to the rest of the scientific community and the society in general. Data clustering constitutes such a field with applications in biomedicine, finances etc. However, data clustering does not cease to evolve and new algorithms are presented regularly. The algorithms originate from developing already existing algorithms and addressing specific problems as well.

Data clustering is comprised on a large number of algorithms, most of which can be separated in groups based on their major characteristic and the way the clusters are perceived. The notions of distance and density are probably the most important in data clustering.

On the one side, algorithms based on distance determine whether the points of a dataset belong to the same cluster based on their distances. On the other side, algorithms based on density usually do not process the point separately, but investigate the areas of the dataset. In this way, they determine areas of high density which describe clusters or at least parts of those. Although data clustering is comprised by more groups of algorithms based on Fuzzy Logic, Gaussian distribution and similarities, those algorithms can be integrated in one of the main groups based on distance or density. Finally, it must be mentioned that algorithms combining those notions have also be proposed. Those combine the notions of distance and density in a single measure used in order to create the clusters. Though, by considering distance as the base criterion for clustering points, algorithms arise that compare all points of the dataset indirectly or not. This leads to an increased computational cost. On the other side, algorithms based on density present a weakness to comprehend the relationships between points in areas of similar density.

The contribution of this thesis in the field of data clustering is the introduction of an algorithm which combines those two notions by selecting the notion which better serves each algorithm step. Hence, it avoids combining the notions into one and therefore potentially restricting the capabilities of each the density and distance. More precisely, it uses the density in order to determine areas of high density and hence the clusters or at least parts of those. Conversely, in order to determine the closest clusters the notion of distance is utilized.

Nevertheless, the contribution of an algorithm depends on the possibility to contribute to other scientific fields as well. Data clustering has been successfully applied in medicine and more specifically in biomedicine, in personalized medicine and medical image processing. In more detail, data clustering is employed in biomedicine and personalized medicine in an attempt to detect interactions between genes, diseases and pharmaceutics. Especially for the case of cancer and even more specifically in the case of breast cancer, the complexity of the interactions that may exist is such that their detection has significant improvement levels.

Therefore, applying data clustering to problems of the above mentioned fields presents a challenge for new clustering algorithms and this thesis examines the application of clustering algorithms on problems related to breast cancer. This is achieved by investigating the potential of the algorithm proposed in the thesis. The clustering algorithm was applied on interactions between genes based on their co-expression and their differential expression. This study aims to determine gene pathways describing the stages of breast cancer. The results of this work on this problem justify its usage as many of the pathways determined are confirmed by the existing literature.