Τμήμα Ψηφιακών Συστημάτων
Ακαδημαϊκό Έτος: 2021 - 2022
Εξόρυξη Γνώσης και Μηχανική Μάθησης

ΕΡΓΑΣΙΑ
Συσταδοποίηση και Κατηγοριοποίηση με το WEKA[1]

Στόχος της εργασίας είναι (α) η ομαδοποίηση των βάσει των χαρακτηριστικών και (β) η κατασκευή ενός μοντέλου κατηγοριοποίησης

Ζήτημα Α

Το τμήμα του Άμεσου Μάρκετινγκ μιας εταιρείας επιθυμεί να εντοπίσει ομάδες πελατών με βάση τα δημογραφικά στοιχεία τους που διαθέτει σε βάση δεδομένων. Ο σκοπός της έρευνας είναι στα πλαίσια του προσδιορισμού στρατηγικών Μάρκετινγκ για την προώθηση νέων προϊόντων. Να εφαρμοστεί η μέθοδος της συσταδοποίησης με την εφαρμογή του αλγορίθμου k-means για κατάλληλων ομάδων πελατών, λαμβάνοντας υπόψη την ηλικία, το εισόδημα, το φύλο, τη μόρφωση. την οικογενειακή κατάσταση, το πλήθος παιδιών και τα έτη που διαμένει ο κάθε πελάτης στην παρούσα κατοικία του. (α) Να οριστεί το πλήθος των συστάδων βάσει του SSE τεκμηριώνοντας την επιλογή αυτή με την παρουσίαση κατάλληλου διαγράμματος (πχ. στο MS Office Excel/LibreOffice Calc). (β) Να περιγραφούν αναλυτικά οι συστάδες που προκύπτουν σκιαγραφώντας τα αντίστοιχα προφιλ των ομάδων πελατών παραθετοντας και τα αντίστοιχα στατιστικά στοιχεία. (γ) Να αναφέρετε τον τρόπο με τον οποίο μετρώνται οι αποστάσεις κατά την εφαρμογή του αλγορίθμου. (δ) Να συσχετίσετε διαγραμματικά τις συστάδες που θα προκύψουν με την περιοχή (region) που διαμένει ο κάθε πελάτης.

Το σύνολο των δεδομένων έχει τα εξής γνωρίσματα:

Attribute		Coding
ID	Customer ID	
Responded	Responded on previous campaign	0-no, 1-yes
Age	Number of years	
Income		1--<25, 2-25-49, 3-50-74,4-75+
Education		1-Some high school or less 2-High school 3-Some college 4-College 5-Post-graduate
Reside	Years at current residence	
Gender		0-Male, 1-Female
Married		0-no, 1-yes
Children	Number of children	
Region	Region of residence	1-North, 2-South, 3-East, 4-West

[1] https://www.cs.waikato.ac.nz/ml/weka/

Ζήτημα Β

Ένας τραπεζικός οργανισμός σας ανέθεσε να δημιουργήσετε έναν κατηγοριοποιητή που θα χρησιμοποιεί δεδομένα που αφορούν την εξυπηρέτηση δανείων και θα προβλέπει αν κάποιο δάνειο θα εξυπηρετείται ή όχι. Στόχος είναι να επιλέξετε το καλύτερο δυνατό μοντέλο κατηγοριοποίησης. Σας ανατέθηκε η διεξαγωγή πειραματικής μελέτης που θα συγκρίνει αλγόριθμους κατηγοριοποίησης χρησιμοποιώντας τη μέθοδο επικύρωσης 10-fold-cross-validation και έτσι να επιλεγεί το καλύτερο δυνατό μοντέλο κατηγοριοποίησης. Το σύνολο δεδομένων έχει τα εξής γνωρίσματα:

age	Age in years	Numeric / scale
ed	Level of education	ordinal
employ	Years with current employer	Numeric / scale
address	Years at current address	Numeric / scale
income	Household income in thousands	Numeric / scale
debtinc	Debt to income ratio (x100)	Numeric / scale
creddebt	Credit card debt in thousands	Numeric / scale
othdebt	Other debt in thousands	Numeric / scale
default	Previously defaulted (δεν εξυπηρέτησε το δάνειο)	Class: 0-No το εξυπηρέτησε, 1-Yes δεν το εξυπηρέτησε

Πρέπει να συγκρίνετε (με χρήση του WEKA) τους εξής αλγόριθμους κατηγοριοποίησης: (α) δέντρο αποφάσεων (C4.5), (β) κ εγγύτεροι γείτονες (ibk), (γ) Naive Bayes και τελικά να προτείνετε τον κατάλληλο κατηγοριοποιητή με την ανάλογη τεκμηρίωση. Θα πάρετε την απόφασή σας λαμβάνοντας υπ' όψιν αποκλειστικά την ακρίβεια που επιτυγχάνουν οι αλγόριθμοι κατηγοριοποίησης και όχι άλλα κριτήρια όπως, λ.χ. το υπολογιστικό κόστος. Ωστόσο, είναι σημαντικό το μοντέλο να προβλέπει σωστά τα μη εξυπηρετούμενα δάνεια (ώστε ο οργανισμός να λάβει τα κατάλληλα μέτρα). Συνεπώς, σημαντικά κριτήρια εκτός από την ακρίβεια είναι η ευαισθησία (recall) και η ορθότητα (precision) της κλάσης που αφορά τα δάνεια που δεν εξυπηρετούνται.

Ζήτημα Β - Επισημάνσεις:

1. Το precision για την κλάση Yes μετράει "πόσες προβλέψεις για μη εξυπηρέτηση δανείου είναι σωστές" ενώ το recall μετράει "από τις μη εξυπηρετήσεις πόσες προβλέφθηκαν". Φυσικά, αν ένας κατηγοριοποιητής προβλέπει πάντα "μη εξυπηρέτηση", θα υπάρχουν πολλά FP. Αυτό σημαίνει χαμηλό precision. Από την άλλη, δεν θα υπάρχει κανένα FN. Άρα θα έχουμε Recall 1. Συνεπώς, το Recall είναι σημαντικό αλλά σε καμία περίπτωση δεν πρέπει να έχουμε εξαιρετικά χαμηλό precision. Το κριτήριο που λαμβάνει υπ' όψιν και το Recall και το Precision, ονομάζεται F-measure και υπολογίζεται από το WEKA. (Χρήσιμοι σύνδεσμοι: 1, 2, 3)

2. Είναι πιθανόν κάποια γνωρίσματα του συνόλου δεδομένων να μην συνεισφέρουν στην ορθή πρόβλεψη και τελικά, μπορεί να βλάπτουν ένα ή περισσότερα από τα παραπάνω μέτρα αξιολόγησης. Καλείστε να διαπιστώσετε αν συμβαίνει κάτι τέτοιο για κάθε έναν από τους κατηγοριοποιητές που θα συμπεριλάβετε στην πειραματική μελέτη. Γνωστές μέθοδοι επιλογής χαρακτηριστικών (attribute selection) είναι η infogain και η gainratio. Υπολογίζουν το κέρδος πληροφορίας (ή αλλιώς εντροπία) κάθε γνωρίσματος (Χρήσιμοι σύνδεσμοι: 1, 2) και τελικά κατατάσσουν αυτά τα γνωρίσματα βάσει κέρδους πληροφορίας. Και οι δυο προσφέρονται από το WEKA μέσω της επιλογής “select attribute”. Συνεπώς, είναι πιθανόν να προτείνετε ένα μοντέλο στην εταιρεία που χρησιμοποιεί ένα υποσύνολο των γνωρισμάτων του αρχικού συνόλου δεδομένων.
3. Να κατασκευάσετε δέντρο αποφάσεων εφαρμόζοντας τον αλγόριθμο C4.5. Στο WEKA ο αλγόριθμος αυτός ονομάζεται J48. Μπορείτε να πειραματιστείτε με διαφορετικές τιμές για τις παραμέτρους minNumObj και unpruned του J48. Το WEKA διαθέτει την επιλογή “Visualize tree” και μπορείτε να δείτε το δέντρο που κατασκευάστηκε.
4. Ο αλγόριθμος κ εγγύτερων γειτόνων στο WEKA αναφέρεται ως ibk. Δεν είναι απαραίτητο να δοκιμάσετε διαφορετικά μέτρα απόστασης πέρα από το μέτρο της Ευκλείδειας απόστασης που είναι το default στο WEKA. Η υλοποίηση ibk μπορεί να υπολογίσει αποστάσεις μεταξύ nominal γνωρισμάτων με τον εξής απλό τρόπο: Αν x=y τότε η απόσταση είναι 0 αλλιώς είναι 1 (Χρήσιμος σύνδεσμος: 1). Δεν είναι απαραίτητο να δοκιμάσετε διαφορετικές παραμέτρους για την τιμή του κ. Το WEKA μπορεί να εκτελέσει τον ibk με την καλύτερη δυνατή τιμή για το κ η οποία αποδεικνύεται βάσει της μεθόδους cross-validation (χρήση validation set, Χρήσιμος σύνδεσμος: 1). Αυτό γίνεται εφόσον επιλέξετε “cross-validate” από το παράθυρο παραμέτρων του κατηγοριοποιητή. Σε αυτή την περίπτωση το WEKA θα κάνει τόσες δοκιμές όσες η τιμή που θα βάλετε στο textbox “KNN”. Προσοχή! Η επιλογή “cross-validate” είναι διαφορετική από την επιλογή “cross-validation Folds:” που εμφανίζεται στο αρχικό παράθυρο “classify”.
5. Για τον αλγόριθμο Naive Bayes μπορείτε να χρησιμοποιήσετε την υλοποίηση “NaiveBayes” ή την υλοποίηση “NaiveBayesSimple”. Ο αλγόριθμος Naive Bayes, στην απλή του μορφή (αυτή που διδάχθηκε στο μάθημα), προϋποθέτει ότι τα γνωρίσματα είναι τύπου nomimal. Αν οι υλοποιήσεις του Naive Bayes στο WEKA μπορούν να διαχειριστούν τέτοιου είδους γνωρίσματα, δεν επιθυμούμε να εξετάσουμε αυτή την παραλλαγή του Naive Bayes. Συνεπώς πρέπει να γίνει διακριτοποίηση (discretization) των nomimal γνωρισμάτων του συνόλου δεδομένων. Αυτό μπορεί να γίνει “αυτόματα” στην περίπτωση που χρησιμοποιηθεί η υλοποίηση “NaiveBayes” επιλέγοντας “useSupervisedDiscretization”. Άλλη λύση είναι να επιλέξετε το φίλτρο “discretize” από το παράθυρο preprocess. Σε αυτή την περίπτωση μπορεί να χρησιμοποιηθεί είτε η μια υλοποίηση είτε η άλλη.
6. Το σύνολο δεδομένων είναι imbalanced. Δηλαδή, η πλειοψηφία των στιγμιοτύπων ανήκει στην κλάση No (εξυπηρέτηση). Μπορείτε να εκτελέσετε oversampling μέσω της μεθόδου SMOTE (Χρήσιμος σύνδεσμος: 1) ώστε να ενισχύσετε την άλλη κλάση με “τεχνητά” στιγμιότυπα με στόχο την αύξηση της ακρίβειας. Η μέθοδος SMOTE είναι προσβάσιμη στο WEKA από το tab preprocess -> filter -> filters -> supervised -> instances. Μπορείτε να αυξήσετε τα στιγμιότυπα της κλάσης Yes τόσο ώστε οι δύο κλάσεις να περιέχουν περίπου τον ίδιο αριθμό στιγμιοτύπων. Αν η έκδοση του weka δεν διαθέτει το SMOTE, μπορείτε να το κατεβάσετε μέσω το package manager του WEKA ο οποίος είναι διαθέσιμος στο αρχικό παράθυρο, στο μενού tools

Τελικό παραδοτέο

Το παραδοτέο είναι μια έκθεση που θα πρέπει να περιλαμβάνει παρουσίαση και σχολιασμό / ερμηνείες αποτελεσμάτων σε σχέση με τα ζητούμενα των παραπάνω ζητημάτων Α και Β. Μπορείτε να συμπεριλάβετε screenshots, πειραματικές μετρήσεις, τιμές παραμέτρων που δοκιμάσατε και οπωσδήποτε τον κατάλληλο σχολιασμό. Η παράδοση της τεχνικής έκθεσης πρέπει να γίνει έως 5 Ιουνίου 2022 αποστέλλοντας email στις διευθύνσεις: stoug@uop.gr . Ως θέμα (Subject) του email που θα στείλετε να αναγράφετε “DMML, Εργασία WEKA”.