ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μελέτη αλγορίθμων βελτιστοποίησης μη-ντετερμινιστικών ταξινομητών σε ιατρικά δεδομένα

Μαρία Π. Τουτουντζή

Επιβλέπων: Σταύρος Περαντώνης, Διευθυντής Ερευνών

ΑΘΗΝΑ

ΜΑΡΤΙΟΣ 2011

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μελέτη αλγορίθμων βελτιστοποίησης μη-ντετερμινιστικών ταξινομητών σε ιατρικά δεδομένα

Μαρία Π. Τουτουντζή

Α.Μ.: ΠΙΒ 08028

Επιβλέπων: Σταύρος Περαντώνης, Διευθυντής Ερευνών, Ερευνητής Α'

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Σταύρος Περαντώνης,** Διευθυντής Ερευνών

Ηλίας Μανωλάκος, Αναπληρωτής Καθηγητής

Σέργιος Πετρίδης, Ερευνητής Δ'

Μάρτιος 2011

ΠΕΡΙΛΗΨΗ

Στον τομέα της ιατρικής, το κόστος μιας λάθος διάγνωσης μπορεί να έχει βαρύτατες συνέπειες. Η εφαρμογή αλγορίθμων μηχανικής μάθησης για εύρεση βέλτιστων ταξινομητών μπορεί να διαδραματίσει σημαντικό ρόλο συμβάλλοντας στην ελαχιστοποίηση των λαθών αυτών. Η εργασία αυτή έχει δύο σκέλη.

Το πρώτο είναι η διερεύνηση της χρησιμότητας που μπορούν να έχουν σε ιατρικά δεδομένα οι μη-ντετερμινιστικοί ταξινομητές. Οι ταξινομητές αυτοί επιτρέπουν την πρόβλεψη, για κάθε δείγμα, περισσότερων από μια κλάσεις. Δεδομένου ότι η πραγματική κλάση πρέπει να συμπεριλαμβάνεται στις προβλέψεις και ότι ο αριθμός των προβλεφθείσων κλάσεων πρέπει να είναι όσο το δυνατόν μικρότερος, αυτό το είδος ταξινομητών μπορεί να θεωρηθεί ως διαδικασία ανάκτησης πληροφορίας (information-retrieval IR). Ειδικότερα μελετήθηκε ο αλγόριθμος του del Coz για την εύρεση βέλτιστων τέτοιων ταξινομητών, ο οποίος χρησιμοποιεί τις εκ των υστέρων πιθανότητες για να υπολογίσει το υποσύνολο των κλάσεων με τη χαμηλότερη αναμενόμενη απώλεια.

Το δεύτερο σκέλος της εργασίας είναι η διερεύνηση τρόπων συνδυασμένης χρήσης μη-ντετερμινιστικών ταξινομητών με ταξινομητές μετά-επιπέδου, όπως οι ταξινομητές με χρήση ψηφοφορίας και οι ταξινομητές συσσωρευμένης εκμάθησης. Συγκεκριμένα, μελετήθηκαν τρεις διαφορετικοί τρόποι συνδυασμού ντετερμινιστικών και μη ντετερμινιστικών ταξινομητών στο βασικό και στο μετά-επίπεδο αντίστοιχα. Η διερεύνηση αυτή γίνεται για πρώτη φορά στην βιβλιογραφία. Στα πλαίσια της εργασίας, γίνεται εκτενής πειραματική συγκριτική αξιολόγηση των μη-ντετερμιστικών αλγόριθμων με και χωρίς ταξινομητές μετά-επιπέδου, σε σχέση με τους αντίστοιχους ντετερμινιστικούς, σε ιατρικά δεδομένα.

Τα αποτελέσματα αναδεικνύουν τόσο την χρησιμότητα των μη-ντετερμινιστικών ταξινομητών σε ιατρικά δεδομένα όσο και την εν γένει χρησιμότητα της συνδυασμένης χρήσης μη-ντετερμινιστικών ταξινομητών με ταξινομητές μετά επιπέδου

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Αναγνώριση Προτύπων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μη-ντετερμινιστικοί ταξινομητές, μετά-ταξινομητές, ιατρικά δεδομένα, αλγόριθμος, βελτιστοποιήση

ABSTRACT

In medicine the cost of a misdiagnosis can have important consequences. For the minimization of errors, the application of algorithms of mechanic learning for finding optimal classifiers can be particularly useful. This work aims at the investigation of usefulness of algorithms for finding optimal non-deterministic classifiers.

Namely, we will study the non-deterministic classifiers. These classifiers allow the prediction more than one categories for a given set of samples. Since the true class should be included in the prediction and that the number of classes predicted should be as small as possible, this type of classifiers can be considered as an information-retrieval (IR) process. In particular we will study an algorithm for the creation of optimal classifiers of this type, according to information-retrieval metrics. Taking into consideration an entry from the set of samples, the algorithm uses posterior probabilities in order to calculate the subset of classes with the lowest expected loss. The final objective is to raise percentage of the predictions that include the true class of the sample against their deterministic ones. Comparative experiments for the evaluation of the algorithm against the equivalent deterministic show the usefulness of the studied approach as well as its effect on meta-level. Finally we introduce a new method to measure the algorithm's effects in which meta-level's classifiers attempt to classify non-deterministic data that were produced by the algorithm.

SUBJECT AREA: Pattern Recognition

KEYWORDS: non-deterministic classification, meta-classification, medical data, algorithm, optimization