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

Κατηγοριοποίηση Δεδομένων Βιοψίας Μαστού

Αγγελική Β. Παλιούρα

Επιβλέπων: **Διονύσης Κάβουρας,** Καθηγητής ΤΕΙ-Α

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2014

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κατηγοριοποίηση Δεδομένων Βιοψίας Μαστού

Αγγελική Β. Παλιούρα

Α.Μ.: ΠΙΒ081

ΕΠΙΒΛΕΠΩΝ: **Διονύσης Κάβουρας,** Καθηγητής ΤΕΙ-Α

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Διονύσης Κάβουρας,** Καθηγητής Α-ΤΕΙ

Μανώλης Σαγκριώτης, Καθηγητής ΕΚΠΑ

Παντελής Ασβεστάς, Επίκουρος Καθηγητής Α-ΤΕΙ

Νοέμβριος 2014

ΠΕΡΙΛΗΨΗ

Αντικείμενο της συγκεκριμένης εργασίας είναι η ανάπτυξη και εφαρμογή αλγορίθμων αναγνώρισης προτύπων για την κατηγοριοποίηση δεδομένων βιοψίας μαστού. Τα δεδομένα περιλαμβάνουν 569 πρότυπα κυτταρολογικής βιοψίας μαστού (357 καλοήθη και 212 κακοήθη) με 30 χαρακτηριστικά κυτταρικών πυρήνων που προέρχονται από τη βάση δεδομένων «Breast Cancer Wisconsin Diagnostic».

Από το αρχικό πλήθος των 30 χαρακτηριστικών, επιλέχθηκε το υποσύνολο εκείνο με την καλύτερη διαχωριστική ικανότητα ύστερα από εφαρμογή των μεθόδων σειριακής οπίσθιας επιλογής (Sequential Backward Selection) και κατάταξης χαρακτηριστικών (rank-features criterion). Για την ταξινόμηση των δειγμάτων σε 2 κλάσεις (καλοήθειες ή κακοήθειες) υλοποιήθηκαν ο ταξινομητής Πλησιέστερου Γείτονα, το Πιθανοκρατικό Νευρωνικό Δίκτυο και τα Διανύσματα Υποστήριξης Απόφασης. Η απόδοσή τους αποτιμήθηκε με 2 διαφορετικές μεθόδους αξιολόγησης: Leave-One-Out και External Cross Validation.

Η ακρίβεια του προτεινόμενου συστήματος στην ταξινόμηση κυτταρικών πυρήνων στις 2 κατηγορίες είναι 97% με την εφαρμογή Διανυσμάτων Υποστήριξης Απόφασης και ως μέθοδο επιλογής χαρακτηριστικών την Sequential Backward Selection.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Αναγνώριση Προτύπων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: χαρακτηριστικά, επιλογή χαρακτηριστικών, ταξινομητής, καρκίνος μαστού, κυτταρολογική βιοψία

ABSTRACT

The purpose of the present thesis is the study and implementation of classification algorithms for breast cancer data set that comes from Fine Needle Aspiration. The data set includes 569 patterns (357 benign and 212 malignant) and 30 features of cellular cores that emanate from the «Breast Cancer Wisconsin Diagnostic» database.

From the initial set of 30 features, the optimal subset of features was derived employing the Sequential Backward Selection method and ranking features with rank-features criterion. The Nearest Neighbor, the Probabilistic Neural Network and the Support Vector Machine classifiers were implemented for the classification of nuclei in two classes (benign or malignant). Their accuracy was evaluated through 2 different evaluation methods: Leave-One -Out and External Cross Validation.

The success rate of proposed system to classify cell nuclei in two categories is 97% using the Support Vector Machine and Sequential Backward Selection as feature selection method.

SUBJECT AREA: Pattern Recognition

KEYWORDS: features, feature selection, classifier, breast cancer, fine needle aspiration