ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

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

Σταματίνα Γ. Πηλιχού

Ολοκληρωμένο σύστημα ταξινόμησης εικόνων μαστογραφίας

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η μελέτη και η ανάπτυξη ενός υπολογιστικού συστήματος υποβοήθησης διάγνωσης για τη διάκριση εικόνων μαστογραφίας σε τρεις κατηγορίες: φυσιολογικές εικόνες, εικόνες με καλοήθη μορφώματα και εικόνες με κακοήθη μορφώματα.

Για τον σχεδιασμό και την υλοποίηση του συστήματος χρησιμοποιήθηκαν 295 εικόνες μαστογραφίας από τη βάση δεδομένων Mini-MIAS, καταλήγοντας σε 299 περιοχές ενδιαφέροντος (ROI), από τις οποίες 207 αποτελούνται από φυσιολογικό ιστό, 53 περιλαμβάνουν καλοήθη μορφώματα και 39 περιλαμβάνουν κακοήθη μορφώματα.

Στη συνέχεια, υλοποιήθηκαν τα ακόλουθα τέσσερα βασικά στάδια: προ-επεξεργασία, εξαγωγή χαρακτηριστικών, επιλογή χαρακτηριστικών και ταξινόμηση. Κατά το στάδιο της προ-επεξεργασίας έγινε αύξηση της αντίθεσης της εικόνας και αποκοπή του ROI. Έπειτα ακολούθησε η εξαγωγή 145 χαρακτηριστικών από κάθε ROI, τα οποία αποτελούνται από χαρακτηριστικά υφής $1^{ης}$ και $2^{ης}$ τάξης της αρχικής εικόνας, καθώς και χαρακτηριστικά υφής $1^{ης}$ τάξης από μετασχηματισμό με μάσκες Laws και φίλτρα Gabor. Τα χαρακτηριστικά αυτά αποτέλεσαν είσοδο σε ένα σύστημα αναγνώρισης προτύπων που σχεδιάστηκε ώστε να κατηγοριοποιεί σε φυσιολογική, καλοήθη ή κακοήθη τη μαστογραφική περιοχή ενδιαφέροντος. Υλοποιήθηκαν οι ταξινομητές: k-Nearest Neighbors (k–NN), Probabilistic Neural Network (PNN), με Gaussian, Exponential και Reciprocal πυρήνα, Naïve Bayes και Support Vector Machine με γραμμικό και Gaussian ακτινικής βάσης πυρήνα. Για κάθε ταξινομητή, βρέθηκε ο βέλτιστος συνδυασμός 10 χαρακτηριστικών με χρήση της μεθόδου Sequential Floating Forward Selection (SFFS). Το προτεινόμενο σύστημα αξιολογήθηκε με τη μέθοδο 10-Fold της μεθόδου Internal Cross Validation (ICV). Η απόδοση του συστήματος σε «άγνωστα» δεδομένα εκτιμήθηκε με τη μέθοδο External Cross Validation (ECV).

Το προτεινόμενο σύστημα παρουσίασε τη μεγαλύτερη ακρίβεια ταξινόμησης (92,4%) για τη διάκριση μιας περιοχής σε φυσιολογική ή παθολογική, χρησιμοποιώντας τον ταξινομητή SVM με γραμμικό πυρήνα και τη μεγαλύτερη ακρίβεια ταξινόμησης (71,2%) για τη διάκριση μιας περιοχής σε καλοήθη ή κακοήθη, χρησιμοποιώντας τον

ταξινομητή PNN με Reciprocal πυρήνα. Συγκρίνοντας τα 10 καλύτερα χαρακτηριστικά για κάθε ταξινομητή, διαπιστώθηκε ότι τα χαρακτηριστικά λοξότητα και λοξότητα εικόνας μετασχηματισμένης από την $7^{η}$ μάσκα του Laws εμφανίζονται σε περισσότερους από 5 ταξινομητές στο στάδιο της διάκρισης παθολογικών και μη παθολογικών εικόνων. Στο στάδιο της διάκρισης μεταξύ καλοήθων και κακοήθων μορφωμάτων, τα πιο συχνά εμφανίσιμα χαρακτηριστικά ήταν η τυπική απόκλιση, η λοξότητα εικόνας μετασχηματισμένης από την $4^{η}$ μάσκα του Laws και η λοξότητα εικόνας μετασχηματισμένης από την $8^{η}$ μάσκα του Laws, υποδεικνύοντας τη διαχωριστική ικανότητά τους. Η ακρίβεια ταξινόμησης «άγνωστων» δεδομένων με την ECV και χρήση των πέντε (2+3) καλύτερων χαρακτηριστικών υπολογίστηκε στο 72,2% ± 5,7% με χρήση του ταξινομητή PNN με Reciprocal πυρήνα και στο 75,2% ± 4,3% με χρήση του ταξινομητή SVM με γραμμικό πυρήνα.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Ανάλυση Εικόνας, Αναγνώριση Προτύπων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: καρκίνος του μαστού, εικόνα μαστογραφίας, εξαγωγή χαρακτηριστικών, επιλογή χαρακτηριστικών, ταξινόμηση

Integrated mammographic image classification system

Abstract

The aim of this thesis is the development of a computer-aided diagnosis (CAD) system for the discrimination between normal, benign and malignant mammographic images using image analysis and pattern recognition algorithms.

For the design and implementation of the proposed system 295 mammographic images were used from the database Mini-MIAS, resulting to 299 regions of interest (ROI). From the 299 ROIs, 207 were normal, 53 were benign and 39 were malignant masses.

Four basic steps of image processing and analysis were implemented: pre-processing, feature extraction, feature selection and classification. At the pre-processing step, the contrast of the image was enhanced and the ROI was cropped out. The next step involved feature extraction, during which 145 features were extracted from each ROI. The features comprised 1st and 2nd order textural features and Laws masks and Gabor filter transform features. These features were used as input into a pattern recognition system that was designed to predict ROI's category (normal, benign or malignant). The implemented classifiers were: k-Nearest Neighbors (k-NN), Probabilistic Neural Network (PNN), with Gaussian, Exponential and Reciprocal kernel, Naïve Bayes and Support Vector Machine with linear and Gaussian radial basis kernel. For each classifier, the optimum 10 feature combination was found, using the method Sequential Floating Forward Selection (SFFS). The proposed system was evaluated by the 10-Fold Cross Validation. The system's performance in 'unknown' data was evaluated using the External Cross Validation method.

The proposed system presented the highest normal-abnormal classification accuracy (92.4%), using the SVM classifier with linear kernel and the highest benign-malignant classification accuracy (71.2%), using the PNN classifier with Reciprocal kernel, with SFFS as feature selection method and the 10-fold CV as the evaluation method. Comparing the top 10 features for each classifier, it was noticed that the features skewness and skewness of transformed image by the 7th mask of Laws were the most frequent in the normal-abnormal classification stage, and the features standard deviation, skewness of transformed image by the 4th mask of Laws and skewness of transformed image by the 8th mask of Laws were the most frequent in the benign-malignant classification stage. The "unknown" data classification accuracy was estimated 72.2% ± 5.7% using the PNN classifier with Reciprocal kernel and 75.2% ± 4.3% using the SVM classifier with linear kernel by the External Cross Validation method using the above five (2+3) most significant features.

SUBJECT AREA: Image Analysis, Pattern Recognition

KEYWORDS: breast cancer, mammographic image, feature extraction, feature selection, classification