جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Bouchra EL OUASSIF

soutiendra publiquement sa thèse de Doctorat en Informatique

Le Samedi 11 Septembre 2021 à 11h00
au Grand Amphi à l'ENSIAS

Intitulé de la thèse

Ensembles methods for Breast Cancer classification

UNIVERSITE MOHAMMED V - RABAT ★ ENSIAS

Devant le Jury composé de :

Président :

Pr. Rachid OULAD HAJ THAMI, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :

Pr. Ali IDRI, PES, ENSIAS, Université Mohammed V de Rabat

Rapporteurs :

Pr. Azeddine ZAHI, PES, FST, Université Sidi Mohamed Ben Abdellah de Fès

Pr. Abdelaziz BERRADO, PES, EMI, Université Mohammed V de Rabat

Pr. Fatima OUZAYD, PH, ENSIAS, Université Mohammed V de Rabat

Examinateur :

Pr. Imade BENELALLAM, PH, INSEA, Rabat

Invité :

Pr. Mohamed HOSNI, PESA, ENSAM, Université Moulay Ismail de Meknès

Ensembles methods for Breast Cancer classification

Abstract: Data mining (DM) or Data Analytics is a set of techniques that allows to analyzing data from different perspectives and summarizing it into useful information. It is the process of finding correlations or patterns in large historical datasets. It can be applied in almost any field ranging from business to education, then to medicine. Data mining has been increasingly used in medicine, especially in oncology. Breast cancer (BC) becomes the most common cancer among females worldwide and the leading cause of death in developed countries. Many studies have attempted to apply Data mining techniques to detect survivability of cancers in human beings. Nevertheless, this research area has not yet reached a consensus on the best technique that can perform better in all circumstances. To tackle this challenge, ensemble-based classification has been recently investigated as a new solution and consists on classifing patients by combining through a combination rule more than one single classification technique.

This thesis aims to: First, perform a systematic mapping and review study to analyze and synthesize studies on the application of data mining techniques in breast cancer. Second, we develop and evaluates heterogeneous ensembles based on three well-known machine learning techniques (Support Vector Machines (SVMs), Multilayer Perceptron (MLP), and Decision Trees (DTs)) and we investigate three parameters tuning techniques: Grid Search (GS), Particle Swarm Optimization (PSO) and the default parameters of the Weka Tool to tune the parameters settings of those single techniques. Third, we evaluate homogeneous ensembles whose members are four variants of the SVM classifier. The four SVM variants used four different kernels: Linear Kernel, Normalized Polynomial Kernel, Radial Basis Function Kernel, and Pearson VII function based Universal Kernel. A MLP classifier is used for combining the outputs of the base classifiers to produce a final decision. Fourth, a method of selection ensemble members based on accuracy and diversity is proposed in order to obtain better classification performance, we evaluate and compare ensemble members' selection based on accuracy and diversity with ensemble members' selection based on accuracy only. A comparison with ensembles without member selection was also performed. Ensemble performance was assessed in terms of accuracy, recall and precision. Q statistics diversity measure was used to calculate the classifiers diversity. The experiments were carried out on three well-known available BC datasets from online repositories. Seven classifiers were included in our experiments. The majority voting combination rule was used to combine the output of the classifiers.

Results show that classification techniques was the most investigated task of DM in BC with a number of 176 studies published between 2000-2018. It was found that of the nine classification techniques investigated, artificial neural networks, support vector machines and decision trees were the most frequently used. We also found that artificial neural networks, support vector machines and ensemble classifiers performed better than the other techniques, with median accuracy values of 95%, 95% and 96% respectively. As for the investigation of the parameters tuning in ensemble based BC classification, the overall results obtained suggest that using GS or PSO techniques for single techniques provide more accurate classification. Moreover, in general, ensembles generate more accurate classification than their single techniques regardless of the optimization techniques used. Also, heterogeneous ensembles based on optimized single classifiers generate better results than the Uniform Configuration of Weka (UC-WEKA) ensembles; and PSO and GS slightly have the same impact on the performances of ensembles. The findings concerning the evaluation of homogenous ensembles ensembles whose members are four variants of the SVM classifier, suggest that ensembles provided a very promising performance compared to its base, and that there is no SVM ensemble with a combination of kernels that have better performance in all datasets. As for the method of selection ensemble members, which takes into account accuracy and diversity of the classifiers, the empirical results suggest that investigating both accuracy and diversity to select ensemble members often led to better performances, and in general, selecting ensemble members using accuracy and/or diversity led to better ensemble performance than constructing ensembles without members' selection.

Keywords: Breast Cancer, Machine Learning, Ensemble Breast Cancer, accuracy.