Applied Computing and Informatics (2016) 12, 117–127

Saudi Computer Society, King Saud University

Applied Computing and Informatics

(http://computer.org.sa)
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Multi Filtration Feature Selection (MFFS) to improve discriminatory ability in clinical data set

S. Sasikala [a,*], S. Appavu alias Balamurugan [b], S. Geetha [c]

[a] Anna University, Tamil Nadu, India
[b] K.L.N. College of Information Technology, Tamil Nadu, India
[c] Thiagarajar College of Engineering, Tamil Nadu, India

Received 13 June 2013; revised 21 March 2014; accepted 29 March 2014
Available online 5 April 2014

KEYWORDS

Medical data mining;
Biomedical classification;
Variance coverage factor;
Principal Component Analysis;
Multi Filtration Feature Selection

Abstract Selection of optimal features is an important area of research in medical data mining systems. In this paper we introduce an efficient four-stage procedure – feature extraction, feature subset selection, feature ranking and classification, called as Multi-Filtration Feature Selection (MFFS), for an investigation on the improvement of detection accuracy and optimal feature subset selection. The proposed method adjusts a parameter named "variance coverage" and builds the model with the value at which maximum classification accuracy is obtained. This facilitates the selection of a compact set of superior features, remarkably at a very low cost. An extensive experimental comparison of the proposed method and other methods using four different classifiers (Naïve Bayes (NB), Support Vector Machine (SVM), multi layer perceptron (MLP) and J48 decision tree) and 22 different medical data sets confirm that the proposed MFFS strategy yields promising results on feature selection and classification accuracy for medical data mining field of research.

1. Introduction

Data mining application in medicine has proved to be a successful strategy in the areas of medical services including prediction of usefulness of surgical procedures, clinical tests, medication procedures, and the discovery of associations among clinical and diagnosis data [37]. The applicability of data mining for healthcare applications is increasingly gaining importance. The availability of diverse-natured medical data for diagnosis and prognosis and of pervasive data mining techniques to process these data offers medical data mining a distinctive place to truly assist and impact patient care.

Due to proliferation of synergized information from enormous patient repositories, there is a paradigm shift in the insight of patients, clinicians and payers from qualitative analysis of clinical data to demanding a better quantitative visualization of information based on all supporting medical data.

* Corresponding author. Tel.: +91 9443831534.
E-mail addresses: nithilannsasikala@yahoo.co.in (S. Sasikala), app_s@yahoo.com (S. Appavu alias Balamurugan), sgeetha@tce.edu (S. Geetha).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2014.03.002
2210-8327

For instance, the physicians can evaluate the diagnostic information of many patients with identical conditions. In the same way, they can verify their findings too, with the conformity of peer physicians working on similar cases in other parts of the world. The patterns that are discovered denote valuable knowledge that helps medical discoveries, for example discovering that a certain combination of features may help in better, and more accurate diagnosis of a particular disease. Accurate diagnosis of diseases and subsequently, providing efficient treatment, form an important part of valuable medical services given for patients in a health-care system.

The unique characteristics of medical databases that pose challenges for data mining are the privacy-sensitive, heterogeneous, and voluminous data. These data may have valuable information which awaits extraction. The required knowledge is found to be encapsulated in/as various regularities and patterns that may not be apparent in the raw data. Extracting such knowledge has proved to be priceless for future medical decision making. Feature selection is crucial for analysing various dimensional bio-medical data. It is difficult for the biologists or doctors to examine the whole feature-space obtained through clinical laboratories at one time. In machine learning, all the computational algorithms recommend only few significant features for disease diagnosis. Then these recommended significant features may help doctors or experts to understand the biomedical mechanism better with a deeper knowledge about the cause of disease and provide the fastest diagnosis for recovering the infected patients as early as possible.

Feature selection methods [12] tend to identify the features most relevant for classification and can be broadly categorized as either subset selection methods or ranking methods. The former type returns a subset of the original set of features which are considered to be the most important for classification. Ranking methods sort the features according to their usefulness in the classification task. Most of the classifiers, irrespective of the application domain, uses the ranking strategy to select the final feature subset, in an ad hoc manner. Feature selection, as a pre-processing step to machine learning, is prominent and effective in dimensionality reduction, by removing irrelevant and redundant data, increasing learning accuracy, and improving result comprehensibility. Feature selection algorithms generally fall into two broad categories, the filter model and the wrapper model [37].The filter model depends on general characteristics of the training data to select some features without involving any learning algorithm. The filter model assesses the relevance of features from data alone, independent of classifiers, using measures like distance, information, dependency (correlation), and consistency. The filter method is further classified into Feature Subset Selection (FSS) and Feature Ranking (FR) methods. The wrapper model needs one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. For each of the generated new subset of features, the wrapper model is supposed to learn the hypothesis of a classifier. It has a propensity to find features better suited to the predetermined learning algorithm resulting in superior learning performance, but it also tends to take more computation time and is economically more expensive than the filter model [37]. Whenever dealing with a large number of features, the filter model is usually chosen due to its high accuracy [9]. The hybrid model takes the advantages of the two previous models, and uses an independent measure to identify the best subsets for a given cardinality and applies a mining algorithm to select the best subset among all best subsets across different cardinalities. However, the ensemble of a filter based model with another filter based model, once for subset selection and again for ranking proves to be a promising approach, for medical data mining. The ensemble is brought about in a fashion so as to reduce the number of features and also to enhance the classification accuracy.

The objective of this research work is aimed at showing that the selection of more significant features from the available raw medical dataset helps the physician to arrive at an accurate diagnosis. The primary focus is on aggressive dimensionality reduction so as to end up with increase in the prediction accuracy. The features are subjected to a double filtration process, at the end of which, only the features that increase the accuracy, and form the subset with the lowest cardinality, with their corresponding rank, are obtained. The method employs an efficient strategy of ensemble feature correlation with ranking method. The empirical results show that the proposed Multi Filtration Feature Selection (MFFS) embedded classifier model achieves remarkable dimensionality reduction in the 22 medical datasets obtained from the UCI Machine Learning repository [10] and Kentridge repository [13].

2. Related work

Numerous works have been carried out in the field of dimensionality reduction for medical diagnosis. The following section presents the summary of those works, highlighting the strengths and weaknesses of each method.

It could be observed that the naive Sequential Forward Feature Selection (SFFS) (pure wrapper approach) [5] is impractical for feature subset selection from a large number of samples of high-dimensional features. Hence Gan et al. [4] proposed the Filter-Dominating Hybrid Sequential Forward Feature Selection (FDHSFFS) algorithm for high dimensional feature subset selection. This method proved to be fast but demanded huge computational complexity. Another variant of the SFFS method called improved F-score and Sequential Forward Search (IFSFS) was proposed by Xie and Wang [36] for feature selection to diagnose erythemato-squamous disease. This method was designed so as to improve the F-score and measured the discrimination between more than two sets of real numbers instead of measuring between only two sets of real numbers. The method's applicability to other medical data sets was not reported and hence it was a very specific system targeted at the diagnosis of erythemato-squamous disease only.

Another category of feature selection methods used Mutual Information score. Vinh et al. [32] proposed a novel feature selection method based on the normalization of this well-known mutual information measurement and utilized the information measurement to estimate the potential of the features. The method could not eclipse the strongly correlated features impact on the classification results. Correlated features may be accounted for redundancy and hence a single representative feature from that subset may be selected for further processing.

An incremental learning algorithm in which the most informative features are learnt at each step, is proposed by