Performance Analysis of Some Competent Learners on Medical Data:
Using GA-Based Feature Selection Approach

Amit Kumar, Birla Institute of Technology, Mesra, India

Bikash Kanti Sarkar, Birla Institute of Technology, Mesra, India

ABSTRACT

Research in disease diagnosis is a challenging task due to inconsistent, class imbalance, conflicting, and the high dimensionality of medical data sets. The excellent features of each data set play an important role in improving performance of classifiers that may follow either iterative or non-iterative approaches. In the present study, a comparative study is carried out to show the performance of iterative and non-iterative classifiers in combination with genetic algorithm (GA)-based feature selection approach over some widely used medical data sets. The experiment assists to identify the clinical data sets for which feature reduction is necessary for improving performance of classifiers. For iterative approaches, two popular classifiers, namely C4.5 and RIPPER, are chosen, whereas k-NN and naïve Bayes are taken as non-iterative learners. Fourteen real-world medical domain data sets are selected from the University of California, Irvine (UCI Repository) for conducting experiment over the learners.

KEYWORDS

Accuracy, Classifier, Extraction, Feature, Prediction

INTRODUCTION

Data mining is a technology for finding useful patterns, meaningful information or knowledge from raw data. In fact, classification (a *supervised* learning task of data mining) is used for extracting useful knowledge over a *large volumetric* data by predicting the *class values* based on the *relevant* attribute values. Therefore, designing an accuracy-based classification model is one of the most important tasks of data mining research. Among the multitudinous classified models, the most widely used models are namely *tree based model*, *rule based model*, *probabilistic model* and *evolutionary algorithms* (Bremermann, Rogson, & Salaff, 1966).

At the present date, data mining in medical domain greatly contributes in discovery of disease diagnosis, and provides the domain users (i.e., medical practitioners) with valuable and previously unavailable knowledge to enhance diagnosis and treatment procedures for various diseases. A number of tools have been made to assist medical practitioners in their

DOI: 10.4018/IJKBO.2021010103

clinical decisions. Few of them are cited here. Seera and Lim (2014) designed an efficient hybrid classification system for medical data using fuzzy min-max neural network. An efficient feature selection based classification technique for health care systems is proposed by using machine learning (Selvakuberan, Kayathiri, Harini, & Devi, 2011). By applying both artificial and fuzzy neural network, Kahramanli and Allahverdi (2008) developed a hybrid system for the diagnosis of diabetes and heart diseases. Lee and Wang (2011) designed a fuzzy expert system to diagnosis diabetes. Kalaiselvi and Nasira (2014) constructed a new system for diabetes diagnosis and cancer prediction using Adaptive Neuro Fuzzy Inference System (ANFIS). Chen and Tan (2012) proposed a prediction model for type-2 diabetes based on different element levels in blood and chemo metrics. A systematic review on the effectiveness of computer based clinical decision support systems is presented on practitioner performances and patient outcomes (Garg, Adhikari, McDonald, Rosas- Arellano, Devereaux, Beyene, Sam, & Haynes, 2005). At the same time, another systematic review on identifying critical features for success in enhancing the clinical practices is presented using decision support systems (Kawamoto, Houlihan, Balas, & Lobach, 2005).

The affirm trend says that these systems have widely been used in clinical diagnosis, prediction and risk forecasting for different diseases. However, it is true that medical data sets contain large number of features but all the features do not necessarily contribute to the classification performance rather some of these may affect the performance of the learners. Also, adopting less number of features reduces the construction time of any learning model. Further, in diagnosis point of view, using less number of excellent features assists greatly the medical professionals. Obviously, effective feature selection scheme is the essential solution in this respect.

The present study focuses on conducting a comparison study showing performances of some well-known iterative and non-iterative approaches over medical data sets. In this purpose, the study first adopts a GA (in-built in WEKA 3.7.2 package) over each data set to select an optimal set of features, and then each selected learner (namely *C4.5*, *RIPPER*, *k*-NN and *Naïve Bayes*) is applied to train over the data set. Finally, the learned knowledge is applied over test data for evaluation.

Why GA is Used for Feature Selection?

It is true that genetic algorithm is a heuristic search optimization technique. It has capability to find the optimal solution(s) from a large search space. In particular, GA is well- suited for NP-hard optimization problem. Certainly, feature selection is a NP-hard optimization problem. Thus, GA can be appropriately implemented at this problem too.

Why Medical Data Sets are Here Chosen?

It may be stated that, being in natural domain, medical data sets are highly *imbalanced*, *voluminous, conflicting* and *complex* in nature. Existence of *missing values* is also a vital problem for natural domain data sets. Therefore, finding accurate learned model for medical data sets is a challenging task. That is why the present study priorities in this research.

The paper is organized as follows. The *INTRODUCTION* Section introduces the importance of data mining algorithms for *classification tasks* and its *applications* in the medical domain. *BACKGROUND* Section briefly discusses the introduction of *feature selection*, *genetic search algorithms*, *C4.5*, *RIPPER*, *k*-NN and *naïve Baysian* classifiers. *METHODOLOGY* Section describes *GA-based hybrid model*, whereas experiments and results are provided in *RESULTS & ANALYSIS* Section. Finally, conclusions are summarized in *CONCLUSION* Section.