Expert Systems with Applications 41 (2014) 2239–2249

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A hybrid intelligent system for medical data classification

Manjeevan Seera [a], Chee Peng Lim [b,*]

[a] Intel Technology, Penang, Malaysia
[b] Centre for Intelligent Systems Research, Deakin University, Australia

ARTICLE INFO

Keywords:
Fuzzy Min–Max neural network
Classification and regression tree
Random forest
Hybrid intelligent systems
Medical decision support

ABSTRACT

In this paper, a hybrid intelligent system that consists of the Fuzzy Min–Max neural network, the Classification and Regression Tree, and the Random Forest model is proposed, and its efficacy as a decision support tool for medical data classification is examined. The hybrid intelligent system aims to exploit the advantages of the constituent models and, at the same time, alleviate their limitations. It is able to learn incrementally from data samples (owing to Fuzzy Min–Max neural network), explain its predicted outputs (owing to the Classification and Regression Tree), and achieve high classification performances (owing to Random Forest). To evaluate the effectiveness of the hybrid intelligent system, three benchmark medical data sets, *viz.*, Breast Cancer Wisconsin, Pima Indians Diabetes, and Liver Disorders from the UCI Repository of Machine Learning, are used for evaluation. A number of useful performance metrics in medical applications which include accuracy, sensitivity, specificity, as well as the area under the Receiver Operating Characteristic curve are computed. The results are analyzed and compared with those from other methods published in the literature. The experimental outcomes positively demonstrate that the hybrid intelligent system is effective in undertaking medical data classification tasks. More importantly, the hybrid intelligent system not only is able to produce good results but also to elucidate its knowledge base with a decision tree. As a result, domain users (i.e., medical practitioners) are able to comprehend the prediction given by the hybrid intelligent system; hence accepting its role as a useful medical decision support tool.

1. Introduction

Research in computerized intelligent systems for medical applications is an important and exciting domain. In general, a physician typically accumulates his/her knowledge based on patients' symptoms and the confirmed diagnoses. In other words, prognostic relevance of symptoms towards certain diseases and diagnostic accuracy of a patient are highly dependent on a physician's experience (Meesad & Yen, 2003). As medical knowledge and treatment therapy progress rapidly, e.g. the occurrence of new diseases and the availability of new drugs, it is challenging for a physician to keep up-to-date with all recent knowledge and development in clinical practice (Meesad & Yen, 2003). On the other hand, with the advent of computing technologies, it is now relatively easy to acquire and store a lot of information digitally, e.g. in dedicated databases of electronic patient records (Pavlopoulos & Delopoulos, 1999). As such, the deployment of computerized medical decision support systems becomes a viable approach to assisting physicians to swiftly and accurately diagnose patients (Chabat, Hansell, & Yang, 2000). Nevertheless, numerous issues have to be overcome before a useful medical decision support system can be developed and deployed, which include decision making in the presence of uncertainty and imprecision (Tsipouras, Voglis, & Fotiadis, 2007). While medical experts' knowledge and experience is important, ranging from assessing a patient's condition to making a diagnosis, advances in machine learning (Kwiatkowska, Atkins, Ayas, & Ryan, 2007) techniques have opened up the way for medical practitioners to exploit computerized intelligent systems for decision support in their workplace, e.g. surgical imagery and X-ray photography (Isola, Carvalho, & Tripathy, 2012). When treating a patient, a physician first needs to narrow down the suspected disease to the root cause (out of a list of probable causes with similar symptoms) using his/her knowledge and experience, and then confirms the diagnosis by performing a number of tests (Isola et al., 2012). Concomitantly, computerized intelligent systems can be useful in assisting the physician to arrive at an informed decision quickly, e.g. by learning from similar past cases in a large database of electronic patient records and inferring the diagnosis for the current patient with proper justifications. The advantages of using such intelligent systems include increasing diagnosis accuracy and, at the same time, reducing time and costs associated with patient treatment (Çomak, Polat, Güneş, & Arslan 2007).

* Corresponding author. Tel.: +61 3 5227 3307.
E-mail address: chee.lim@deakin.edu.au (C.P. Lim).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.022

Machine learning models have been developed to support various medical decision making tasks. As an example, intelligent classifiers have been used for prognosis, diagnosis, and screening of diabetes, breast cancer and Parkinsons disease (Luukka, 2011). A number of neural-fuzzy models have been used as classifiers for heart disease, because they are capable of learning from data samples (i.e., patient records) and generalizing beyond the training samples (Kahramanli & Allahverdi, 2009). These include fuzzy neural networks, fuzzy probabilistic neural networks, and fuzzy learning vector quantization networks (Sekar, Dong, Shi, & Hu, 2012). However, one key limitation of these models is the lack of ability to explain their predictions (Markowska-Kaczmar & Matkowski, 2006). This is the motivation of this research, whereby we attempt to devise a machine learning-based system that is able to reveal its reasoning in dealing with an input case, and to provide justification for its predictions.

In machine learning, neural networks have significant advantages for medical decision support applications (Downs, Harrison, Kennedy, & Cross, 1996). Compared with expert systems, neural networks avoid the time-consuming and demanding knowledge acquisition process by directly learning complex association between input symptoms and target diseases from data samples i.e., patient records (Hayes-Roth, Waterman, & Lenat, 1983). In addition to learning, neural networks possess other useful properties, which include handling incomplete or missing data as well as filtering noise, uncertainty or imprecision (Downs et al., 1996). In view of the salient features of neural networks, the Fuzzy Min–Max (FMM) neural network is investigated for developing a usable and useful medical decision support tool in this paper. In order to further strengthen the FMM network for medical applications, the Classification and Regression Tree (CART) and Random Forest (RF) models are incorporated to produce a hybrid intelligent system. The proposed hybrid model serves as an extension of our previous work (Seera, Lim, Ishak, & Singh, 2012), which was focused on an offline model (i.e. FMM-CART) for undertaking fault detection and diagnosis problems. Here, the motivation is to extend the hybrid model such that it is equipped with the necessary characteristics for undertaking medical decision support tasks. While CART has the advantage of rule extracting in the form of a tree structure, it is less flexible in performing incremental learning from data samples. While FMM has the advantage of one-pass training with incremental learning properties, it lacks the capability of producing rules to explain its predictions. On the other hand, RF has the benefit of forming an ensemble of CART whereby the best tree can be identified to produce high prediction accuracy. Therefore, the hybrid model, i.e. FMM-CART-RF, has three distinctive capabilities, *viz,* learning incrementally from data samples (owing to FMM), explaining its predicted outputs (owing to CART), and achieving high classification performances (owing to RF). This is the key contribution of this research.

From the perspective of decision making, we commonly seek a second opinion (and sometimes even more) before making important decisions, especially one that has medical implications (Polikar, 2006). Then, different opinions are weighed and combined based on a thought process before the final decision is made. In machine learning, such ensemble concept is utilized too in developing a highly accurate model. In particular, bagging algorithms are useful for constructing an ensemble of decision trees, and one such variant is RF. RF is an effective ensemble method for data mining (Zhang, Zulkernine, & Haque, 2008). It has shown good results in many applications, which include automatic intrusion detection systems (Zhang et al., 2008) and medical data classification (Wu, Ye, Liu, & Ng, 2012). RF has some important characteristics such as providing useful internal estimates of strength, correlation, and variable importance, while producing high accuracy in comparison with many standard classification models (Wu et al., 2012). As a result, RF is exploited in this paper to form a group of diverse CART models so that the best tree can be chosen for classification and rule extraction purposes.

The hybrid model proposed in this research has two important practical implications in the domain of medical decision support. Firstly, the ability to provide explanation and justification for the prediction is of paramount importance, in order to convince domain users (i.e., medical practitioners) with the outcome given by a computerized decision support system. This ability is essential in safety critical applications, such as medical diagnosis and prognosis, whereby domain users need to understand, and be convinced of, how the computerized system arrives at such a prediction (Economou, Goumas, & Spiropoulos, 1996). The elucidated rules in the form of a decision tree from the hybrid model is, therefore, important in practice, whereby the rules could serve as a source of second opinions in medical diagnostic situations (Kovalerchuk, Vityaev, & Ruiz, 2000). Secondly, accuracy of a decision support system is very crucial in medical applications. As stated in Luukka (2011), a high false negative rate of a screening system would increase the risk of patients by depriving them from getting the necessary medical attention, while a high false alarm rate would cause unnecessary worry and stress in patients as well as increase the demand on medical resources. One the other hand, as indicated in Kinney (2003), a decision support system with high specificity and variable sensitivity could save medical costs and improve scheduling of vestibular patients in an otolaryngology clinic. Besides that, Huang, Yang, King, and Lyu (2006) also recognized the usefulness of machine learning models in reducing cost and saving time for undertaking medical diagnostic tasks. As shown in the experimental study, the proposed hybrid model not only is able to achieve high accuracy, sensitivity, and specificity rates, but also to provide explanation for its predictions in the form of a decision tree; hence demonstrating its usefulness as a decision support system in practical environments.

The organization of this paper is as follows. Literature reviews related to intelligent models for medical applications as well as rule extraction methods are detailed in Section 2. In Section 3, the dynamics of the proposed FMM-CART-RF model is presented. The experimental study, results, and discussion using a number of benchmark medical data sets are presented in Section 4. Finally, conclusions and suggestions for further work are presented in Section 5.

2. Literature Review

The literature review presented in this section is divided into two main parts: *(i)* intelligent systems for medical applications; *(ii)* rule extraction methods the emphasis on medical applications. In machine learning, supervised learning is a commonly used method for tackling medical problems. The task of identifying the smallest sets of genes and constructing a highly accurate classification model of cancers from microarray data was attempted using a fuzzy neural network and the Support Vector Machine (SVM) (Wang, Chu, & Xie, 2007). Important genes were chosen using a feature importance ranking scheme. The identified genes were then fed to the learning algorithm for classification. Cascio et al. (2006) utilized a supervised neural network for classification of breast cancer. A set of representative features was extracted and used for detecting masses in mammographic images. Among various supervised learning models, the Multi-Layer Perceptron (MLP)-based neural networks are commonly used in different application domains. In the medical area, MLP was used to classify patterns of scoliosis spinal deformity using features from the central axis curve of spinal deformity (Lin, 2008). Osareh, Shadgar, and Markham (2009) used an MLP neural network with a genetic-based algorithm for