Chapter 17
A Comparative Study on Medical Diagnosis Using Predictive Data Mining: A Case Study

Seyed Jalaleddin Mousavirad
University of Kashan, Iran

Hossein Ebrahimpour-Komleh
University of Kashan, Iran

ABSTRACT

Medical diagnosis is a most important problem in medical data mining. The possible errors of a physician can reduce with the help of data mining techniques. The goal of this chapter is to analyze and compare predictive data mining techniques in the medical diagnosis. To this purpose, various data mining techniques such as decision tree, neural networks, support vector machine, and lazy modelling are considered. Results show data mining techniques can considerably help a physician.

INTRODUCTION

Data mining techniques have been successfully employed in the various biomedical domains. Diagnosis of disease is one of the most important issues in this domain. Medical diagnosis is a difficult and visual task which is often carried out by an expert. An expert commonly takes decisions by evaluating the current test results of a patient or the expert compares the patient with other patients with some condition by referring to the previous decisions(Polat & Güneş, 2006). Therefore, medical diagnosis is a very difficult task for an expert. For this reason, in recent years, data mining techniques have been considered to design automated medical diagnosis systems. With the help of automated medical diagnosis systems, the possible error experts can dramatically reduce.

Data mining is a computational process to find hidden patterns in large datasets. One of the main steps in data mining is model building using predictive data mining techniques. Predictive data

DOI: 10.4018/978-1-4666-6086-1.ch017

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

mining techniques learn from past experience and apply it to future situations. Classification is a method of predictive data mining. Classification algorithms build a model to predict class labels in the given data. There are various algorithms to classification such as support vector machine, neural networks, decision tree, and nearest neighbor classifier. Performance of classification algorithms depend on the characteristics of the data. There is no classification algorithm that works best on all problems (no-free-launch principle). Various empirical tests should be done to find the best classification algorithm on a dataset.

The goal of this chapter is to analysis and compare various classification algorithms on medical diagnosis. For this purpose, a case study on diagnosis of breast cancer is considered. Breast cancer is a type of cancer originating from breast tissues. It is reported that breast cancer was the second one among the most diagnosis cancers and includes 22.9% of all cancers in women(Moftah et al., 2013). It is one of the major causes of death all over the world and more than 1.2 million women will be diagnosed with breast cancer each year worldwide(Tondini, Fenaroli, & Labianca, 2007). Such a disease needs effective and accurate diagnosis to ensure quick and effective treatment.

In this study, various classification algorithms with various parameters are compared for diagnosis of breast cancer. The rest of this chapter is organized as follows: first, various classification algorithms is considered. Then, a general configuration for medical diagnosis will be introduced. In this configuration, after feature extraction, features are reduced with feature selection and reduction algorithms then, a classification algorithm is applied on new feature subset.

LITERATURE REVIEW

This chapter is referred to the application of data mining in medical diagnosis, in particular predictive data mining methods. Data mining methods have been applied to a variety of medical diagnosis in order to improve the process of medical diagnosis. Artificial neural networks (ANN) such as Multilayer perceptron (MLP), Probabilistic Neural Network (PNN), Radial Basis function (RBF), and learning vector quantization (LVQ) have been widely used in disease diagnosis. Temurtas (2009) compared MLP, PNN and LVQ for thyroid disease diagnosis. In another work, MLP was applied for diagnosis of hepatitis disease(Bascil & Temurtas, 2011); in the subsequent work, these authors used PNN for hepatitis diagnosis(Bascil & Oztekin, 2012). PNN has also been used to Mesothelioma's disease(Er, Tanrikulu, Abakay, & Temurtas, 2012). In addition, evolutionary neural networks and ensemble of neural networks have been applied in some works(Abbass, 2002; Das, Turkoglu, & Sengur, 2009b).

Support vector machine is another classifier which was frequently used for disease diagnosis such as Thyroid(Dogantekin, Dogantekin, & Avci, 2011), hepatitis(Çalişir & Dogantekin, 2011; Chen, Liu, Yang, Liu, & Wang, 2011; Sartakhti, Zangooei, & Mozafari, 2011), breast cancer(Chen, Yang, Liu, & Liu, 2011), and heart disease(Yan & Shao, 2003). In addition, tree based classification algorithms and discriminant analysis can be seen in the literature(Jerez-Aragonés, Gómez-Ruiz, Ramos-Jiménez, Muñoz-Pérez, & Alba-Conejo, 2003; Kabari & Nwachukwu, 2012; Pandey, Pandey, Jaiswal, & Sen, 2013; Vlahou, Schorge, Gregory, & Coleman, 2003).

Fusion of classifiers is one method to improve the classification performance. These methods were applied to disease diagnosis. Das et al. (2009b) applied ensemble of neural networks for heart disease. In another works, the same authors used ensemble of neural networks for valvular heart disease (Das, Turkoglu, & Sengur, 2009a). Ozcift (2012) employed forest ensemble to improve computer-aided diagnosis of Parkinson disease.

Table 1 provides a review on data mining applications in medical diagnosis.

32 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/a-comparative-study-on-medical-diagnosis-using-predictive-data-mining/109989

Related Content

Finding Associations in Composite Data Sets: The CFARM Algorithm
M. Sulaiman Khan, Maybin Muyeba, Frans Coenen, David Reidand Hissam Tawfik (2013). *Developments in Data Extraction, Management, and Analysis (pp. 162-186).*
www.irma-international.org/chapter/finding-associations-composite-data-sets/70797

Visualization in Learning: Perception, Aesthetics, and Pragmatism
Veslava Osinska, Grzegorz Osinskiand Anna Beata Kwiatkowska (2016). *Big Data: Concepts, Methodologies, Tools, and Applications (pp. 493-526).*
www.irma-international.org/chapter/visualization-in-learning/150180

TBSGM: A Fast Subgraph Matching Method on Large Scale Graphs
Fusheng Jin, Yifeng Yang, Shuliang Wang, Ye Xueand Zhen Yan (2018). *International Journal of Data Warehousing and Mining (pp. 67-89).*
www.irma-international.org/article/tbsgm/215006

Ensemble PROBIT Models to Predict Cross Selling of Home Loans for Credit Card Customers
Hualin Wang, Yan Yuand Kaixia Zhang (2008). *International Journal of Data Warehousing and Mining (pp. 15-21).*
www.irma-international.org/article/ensemble-probit-models-predict-cross/1803

Kernal Width Selection for SVM Classification: A Meta-Learning Approach
Shawkat Aliand Kate A. Smith (2005). *International Journal of Data Warehousing and Mining (pp. 78-97).*
www.irma-international.org/article/kernal-width-selection-svm-classification/1760