An International Journal
JSCDSS
http://www.jscdss.com
Vol.4 No.5 October 2017: 8-16
Article history:
Accepted 3 September 2017
Published online 3 September 2017

Journal of Soft Computing and Decision Support Systems

E-ISSN: 2289-8603

Knowledge Discovery and Diseases Prediction: A Comparative Study of Machine Learning Techniques

Mehrbakhsh Nilashi [a,f,*], Hossein Ahmadi [b,c,*], Leila Shahmoradi [b], Abbas Mardani [d], Othman Ibrahim [a], Elaheh Yadegaridehkordi [e]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

[b] Health Information Management Department, 5th Floor, School of Allied Medical Sciences, Tehran University of Medical Sciences, No #17, Farredanesh Alley, Ghods St, Enghelab Ave, Tehran, Iran

[c] Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

[d] Faculty of Management, Universiti Teknologi Malaysia (UTM), Skudai Johor

[e] Department of Software Engineering, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[f] Department of Computer, Lahijan Branch, Islamic Azad University, Lahijan, Iran

*** Corresponding authors email addresses**: nilashidotnet@hotmail.com; hosseinis3007@gmail.com

Abstract

The use of medical datasets has attracted the attention of researchers worldwide. Data mining techniques have been widely used in developing decision support systems for disease classification through a set of medical datasets. In this paper, we propose a predictive method for diseases prediction using machine learning techniques. The proposed method is developed through clustering, noise removal, and supervised machine learning techniques. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Neural Network (NN), Adaptive Network-Based Fuzzy Inference System (ANFIS), Support Vector Regression (SVR) and Classification and Regression Trees (CART) are used for diseases prediction task. We also use the Principal Component Analysis (PCA) for dimensionality reduction and to address multi-collinearity problems in the experimental datasets. We test our proposed method on several public medical datasets. Experimental results on Wisconsin Diagnostic Breast Cancer, StatLog, Cleveland and Parkinson's telemonitoring datasets show has potential to be used as a decision support system in healthcare.

Keywords: Machine Learning, Diseases prediction, Medical datasets, Knowledge discovery, Accuracy

1. Introduction

The potential of data mining in solving the problems of diseases classification was identified by World Health Organization (WHO) (Gulbinat, 1997). The use of data mining techniques in knowledge discovery for diseases classification has been one of the interesting and important topics addressed by the researchers (Bellazzi et al., 2008; Cauchi et al., 2015; Chen et al., 2016). Data mining is the unified name for all tools that can be used when searching for relationships and trends in large amounts of data (Han et al., 2011). Data mining is a process of discovering useful knowledge from database to build a structure (i.e., model or pattern) that can meaningfully interpret the data. Data mining techniques are pattern recognition techniques can be used to assist physicians in diagnosing and predicting diseases so they can provide the necessary treatment and prevent the impact, including the possibility of death.

Supervised learning techniques construct prediction models for classifying future events in a way consistent with historical information (Kotsiantis et al., 2007; Caruana and Niculescu-Mizil, 2006). Supervised data mining techniques have frequently appeared in various practical areas especially in health care (Zadeh et al., 2010). Classification and prediction problems have a vital role in medical decision making (Pendharkar et al., 1999). Accordingly, due to diseases diagnosis importance to mankind, several studies have been conducted on modeling procedures for their classification (Fida et al., 2011; Lahsasna et al., 2012; Palaniappan and Awang, 2008; Cauchi et al., 2015; Chen et al., 2016; Nilashi et al., 2017a; Nilashi et al., 2017b; Nilashi et al., 2017d).

Using clinical data, data-driven predictive models are increasingly being implemented. The main aim of using the data mining techniques is to find best ways to construct generalizable models using clinical data. There is a vast sea of different techniques and algorithms used in data mining especially for supervised machine learning techniques; therefore, selecting the appropriate techniques has been a challenge among researchers in developing the medical