Journal of Applied Biomedicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Applied Biomedicine

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

Review article

A survey on applying machine learning techniques for management of diseases

Enas M.F. El Houby

National Research Centre, Systems and Information Department, Engineering Division, Cairo, Egypt

ARTICLE INFO

Article history:
Received 30 September 2017
Received in revised form 2 December 2017
Accepted 11 January 2018
Available online xxx

Keywords:
Data mining
K-nearest neighbour
Decision tree
Artificial neural network
Associative classification

ABSTRACT

During the past years, the increase in scientific knowledge and the massive data production have caused an exponential growth in databases and repositories. Biomedical domain represents one of the rich data domains. An extensive amount of biomedical data is currently available, ranging from details of clinical symptoms to various types of biochemical data and outputs of imaging devices. Manually extracting biomedical patterns from data and transforming them into machine-understandable knowledge is a difficult task because biomedical domain comprises huge, dynamic, and complicated knowledge. Data mining is capable of improving the quality of extracting biomedical patterns.

In this research, an overview of the applications of data mining on the management of diseases is presented. The main focus is to investigate machine learning techniques (MLT) which are widely used to predict, prognose and treat important frequent diseases such as cancers, hepatitis and heart diseases. The techniques namely Artificial Neural Network, K-Nearest Neighbour, Decision Tree, and Associative Classification are illustrated and analyzed. This survey provides a general analysis of the current status of management of diseases using MLT. The achieved accuracy of the various applications ranged from 70% to 100% according to the disease, the solved problem, and the used data and technique.

Introduction

The past few years have witnessed an increasing of data in all fields and of almost all types. Biomedical data specifically have increased dramatically in the past years because of the exponential growth of knowledge in biomedical domain. The amounts of data generated by healthcare transactions are too complex and huge to be processed and analyzed by traditional methods. This data often hide valuable knowledge. Biomedical researchers face a problem of finding important knowledge from this huge amount of data. So health informatics is a rapidly growing field that is concerned with applying computer science and information technology to medical and health data (Brin, 1998; Huang et al., 2006).

Health informatics is the field of information science concerned with the analysis, use and dissemination of medical data and information through the application of computers to various aspects of health care and medicine (National Library of Medicine, 2017). Health informatics is defined as "all aspects of understanding and promoting the effective organization, analysis, management, and use of information in health care." It involves the use of informatics in the discovery and management of new knowledge relating to health and disease (American Medical Informatics Association, 2017). Appropriate computer-based systems and efficient analytical methodologies can help to discover important hidden knowledge from huge medical databases. So in present era, data mining is becoming popular in healthcare field.

Data mining (DM) provides the methodology and technology to transform mounds of data into useful information for decision making. Data mining is defined as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database". It is the core step of a broader process, called knowledge discovery in databases. This process includes the application of several pre-processing methods aimed at facilitating the application of the DM algorithm and post processing methods aimed at refining and improving the discovered knowledge (Freitas, 2003).

Building prediction models from various medical data sources is possible using a knowledge discovery in data or DM approach based on different MLTs and the prediction accuracy of the resulted intelligent systems could even reach high accuracy. It is the process of finding correlations or patterns among different fields in large medical databases. Application of DM in the medical field can be used to analyze and find hidden patterns inside patients' datasets or medical databases.

E-mail addresses: enas_mfahmy@yahoo.com, em.fahmy@nrc.sci.eg (E.M.F. El Houby).

https://doi.org/10.1016/j.jab.2018.01.002

Please cite this article in press as: E.M.F. El Houby, A survey on applying machine learning techniques for management of diseases, J. Appl. Biomed. (2018), https://doi.org/10.1016/j.jab.2018.01.002

Motivated by the world-wide increasing mortality of cancer, hepatitis and heart disease patients each year and the availability of huge amount of patients' data that could be used to extract useful knowledge, researchers have been using DM to help health care professionals in the management of these disease (Helma et al., 2000). In this research, an overview of the current researches being carried out using Artificial Neural Network (ANN), K-Nearest Neighbour (K-NN), Decision Tree (DT), and Associative Classification (AC) for diagnosis, prognosis and treatment of these diseases is presented. The rest of this research is organized as follows: First an overview of the techniques and diseases which are focused in this study is introduced in the *Background* section. The used methodology and extracted data from literature review is presented in the Materials and methods section. The results analysis of extracted data is presented in the results section. And finally the conclusions and discussion are presented in the Discussion and Conclusion section.

Background

Data mining in healthcare is field of high importance attempts to solve real world health problems such as a deeper understanding of medical data and the prediction of diseases (Liao and Lee, 2002). Researchers are using DM in the prediction of several diseases such as diabetes, stroke, cancer, and heart disease. Several DM techniques or the so called MLT such as Naïve Bayes, Decision Tree, neural network, genetic algorithm, and support vector machine when used showed different levels of accuracies.

Cancer, Hepatitis, and cardiovascular diseases are among the most serious and diverse diseases. The amount of data coming from instrumental and clinical analysis of these diseases is quite large and therefore the development of tools to facilitate management of these diseases is of great importance. Also since that the MLTs which have been used in biomedical domain are found to be too many and applied on many diseases. So, in this research the focus will be on the application of some important techniques which are ANN, K-NN, DT, and AC to the field of management of these diseases. In the next subsections, an overview of these adopted diseases and the selected MLTs will be given.

The adopted diseases

The application of data mining on different diseases is rapidly spreading, in this review Hepatitis, Cancers and heart diseases will be tackled as widely spread diseases all over the world and increasing mortality diseases according to World Health Organization (WHO).

Heart or cardiovascular diseases (CVDs)

Heart disease is the leading cause of death in the world over the past 10 years. An estimated 17.5 million people died from CVDs in 2012, representing 31% of all global deaths. An estimated 7.4 million deaths were due to coronary heart disease and 6.7 million were due to stroke. Over three quarters of CVD deaths take place in low- and middle-income countries. Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol. Cardiovascular disease is caused by disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks), raised blood pressure (hypertension), and heart failure (WHO, 2017b).

Hepatitis

Hepatitis disease is from the most dangerous disease and cause of death in the world specially Hepatitis C. It is a contagious liver disease that can range in severity from a mild illness lasting a few weeks to a serious, lifelong illness. The hepatitis C virus is usually spread when blood from an infected person enters the body of a susceptible person. Every year, 3–4 million people are infected with the hepatitis C virus. About 150 million people are chronically infected and are at risk of developing liver cirrhosis and/or liver cancer. More than 350 000 people die from hepatitis correlated liver diseases every year. There are 6 genotypes of hepatitis C and they may respond differently to treatment (WHO, 2017c).

Cancer

Cancer figures among the leading causes of morbidity and mortality worldwide, with approximately 14 million new cases and 8.2 million cancer related deaths in 2012. The number of new cases is expected to rise by about 70% over the next 2 decades. Among men, the 5 most common sites of cancer diagnosed in 2012 were lung, prostate, colorectum, stomach, and liver cancer. Among women the 5 most common sites diagnosed were breast, colorectum, lung, cervix, and stomach cancer. It is expected that annual cancer cases will rise from 14 million in 2012 to 22 within the next 2 decades.

Cancer is a generic term for a large group of diseases that can affect any part of the body. One defining feature of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries, and which can then invade adjoining parts of the body and spread to other organs, the latter process is referred to as metastasizing which is the major cause of death from cancer (WHO, 2017a).

Machine learning techniques (MLTs)

Machine learning is to train the system over a large databases, where the applied MLT can be used to generate extraction patterns or build a model and use the generated patterns or model to make predictions in the future for unknown cases. The data set used to learn the model is known as the training data set. The records making up the training set are referred to as training samples and are randomly selected from the sample population. The model is built collectively from the training data set. Since the value or class label of each training sample is provided, this step is known as supervised learning (Han et al., 2006).

The data set used to measure the quality of the model is known as the test data set. Test data set is used to estimate the predictive accuracy of the model. Test samples are randomly selected and are independent of the training samples. The accuracy of a model on a given test set is the percentage of test samples that are correctly predicted by the model. (Han et al., 2006). If the accuracy of the model is considered acceptable, the model can be used to classify or predict future data records or objects for which the class label or value is unknown. Fig. 1 shows the different steps of learning.

The different MLTs can be compared and evaluated by the following criteria:

- Speed: The computation costs involved in generating and using the model.
- Scalability: The ability to construct the model efficiently given large amount of data.
- Interpretability: The level of understanding that is provided by the model.
- Predictive accuracy: The ability of the model to correctly predict the class label or value of new or previously unseen data (Han et al., 2006), it can be calculated using the following formula.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Please cite this article in press as: E.M.F. El Houby, A survey on applying machine learning techniques for management of diseases, J. Appl. Biomed. (2018), https://doi.org/10.1016/j.jab.2018.01.002