TOPIC:AI BASED DIABETES PREDICTION SYSTEM

Abstract

Data mining and machine learning have become a vital part of different disease detection and prevention. One of them is diabetes. The purpose of this paper is to evaluate data mining methods and their performances that can be used for analysing the collected data about the diabetes. We identified the most appropriate data mining methods to analyse the data by comparing them theoretically and practically. Some attributes of this dataset are: Age, Body Mass Index, Insulin, Glucose, etc. Methods are applied on these data to determine their effectiveness in analysing and preventing diabetes. Evaluations on the data showed that the method with a higher performance is "Decision Tree". This was achieved by some performance measures, such as the number of instances correctly classified, accuracy, precision, recall and F-measure, that has brought better results compared to other methods. We come to the conclusion that the data mining methods and machine learning contribute to the predictions on the possibility of occurrence of the diabetes.

1. Introduction

Diabetes is a disease that is increasingly affecting the world even the most developed countries. Diabetes by the nature of its development as a globally problematic disease requires maximum commitment from medical staff, patients, family and society. Diabetes is a disease with high social, health and economic costs. Diabetes is a chronic disease characterized by an increase in glucose or blood sugar levels because the body cannot produce insulin or its production is insufficient, or insulin is not able to act on the cells of the organism. Medics still do not know exactly why such a thing is happening and they have called the cause: x syndrome. Historically diabetes treatment has been done by fighting the symptoms and not the cause. According to the World Health Organization, Diabetes affects about 5% of the world's population and the number of patients is constantly increasing .In developed countries, diabetes and the largest number of diabetics are found in people over 65 years of age. Whereas in developing countries where our country is part of the largest number of diabetics is found in the age of 45-64 years, but in recent years type 2 diabetes is more commonly encountered also in the age of 30-40 years. The availability of historical data naturally leads to the application of data mining techniques for pattern discovery. The goal is to find rules that help understand diabetes and make it easier to diagnose it sooner. Prevention of diabetes is of great interest in the field of medicine. The use of data mining accelerates data analysis, and analysts can examine existing data to identify patterns and trends of diabetes.

2. Using Data Mining and Machine Learning in Medicine

Medicine is the science and practice of establishing the diagnosis, prognosis, treatment, and prevention of disease. Medicine encompasses a variety of health care practices evolved to maintain and restore health by the prevention and treatment of illness.This is one of the most important areas when applying data mining techniques can produce significant results.

With data mining techniques, doctors will be able to predict illnesses effectively and they will be better equipped to manage potential high-risk candidates. The high volume of diseases data and the complexity of the relationships between them have made medicine an appropriate field for applying data mining techniques. Data mining can be used to examine many large datasets involving a large set of variables beyond what a single analyst or doctor, or even an analytical team can. Like any other problem solving method, the task of data mining begins with a problem definition. The identification of the data mining problem enables the determination of the data

mining process and the modeling technique. Machine learning is a subfield of data science that deals with algorithms able to learn from data and make accurate predictions. Data mining gives health organizations the opportunity to learn about disease trends etc. By using data mining methods and machine learning algorithms we improve diabetes analysis and we help to reduce and prevent it.

3. Data and Methodology

We compare theoretically and practically data mining methods to discover the most appropriate method for our data. The methods were compared by applying machine learning algorithms to concrete data in the WEKA "Waikato Environment for Knowledge Analysis" environment. The implemented algorithms are: Simple Logistic, Multilayer Perceptron, Logistic, Naive Bayes, Bayes Net, SMO, C4.5.

4. *Classification*

Classification is a data mining technique that categorizes data in order to assist in more accurate predictions and analysis. It is one of the data mining methods that aims to analyze very large datasets. It is used to derive patterns that accurately define the important data classes within the data set. Classification techniques predict the target classes for each of the present data instance.. Classification algorithms attempt to detect relationships between attributes that would make it possible to predict the result. They analyze the input and produce a prediction. The classification task of data mining is generally used in healthcare industries .

4.1. Naïve Bayes

Bayesian classification represents a supervised learning method as well as a statistical classification method. The Naive Bayes Classifier technique is based on the Bayesian theorem and is used especially when the dimensionality of the inputs is high.

Naive Bayes is a strong and powerful Predictor.

4.2. Support Vector Machine

SVM classifier is a supervised learning Algorithm based on statistical learning theory Introduced by Vepnik (Vapnik, 1995). The Main idea behind this method is to determine a Hyperplane that optimally separates two classes Using training dataset. SVM is a set of related Supervised learning method used in medical Diagnosis for classification and regression.Support Vector Machine (SVM) model is the Representation of examples defined as points in Space that are mapped so that the examples ofThe different categories can be divided by a Clear gap that is as large as possible.SVM Also supports regression and classification Techniques and can handle multiple continuous And categorical variables. The efficiency of SVM-based classification is not directly Dependent on the dimension of the classified Entities. This algorithm achieves high Discriminative power by using special nonlinear Functions called kernels to transform the input Space into a multidimensional space. It can Be seen that the choice of kernel function and Best value of parameters for particular kernel is Critical for a given amount of data .It also Normalizes all attributes by default.

4.3. The decision tree

Decision tree model has a tree structure, which can describe the process of classification instances based on features. It splits the data in the database into subsets based on the values of one or

more fields. This process will be repeated for each subgroup recursively until all instances are a node in a single class. The result of the decision tree is a tree-shaped structure that describes a series of decisions given at each step. Decision trees are easy to interpret and understand. They provide white box structure for each provided dataset and can be combined with any other data mining techniques. The typical algorithms of decision tree are ID3, C4.5, CART and so on. In this study, we used the C4.5 algorithm.

4.4. Artificial Neural Network

Neural networks are an area of Artificial Intelligence (AI), where based on the inspiration we have from the human brain. Applying neural network techniques, a program can learn from the examples and create an internal set of rules for classifying different inputs. All processes of a neural network are performed by this group of neurons or units.

5. Association Rules and Regression

Association Rule is one of the most important canonical tasks in data mining and probably one of the most studied techniques for pattern discovery. Association rules are if/then statements that help to uncover relationships between unrelated data in a database, relational database or other information repository.Association Rules identify the arguments found together with a given, event or record: "the presence of one set of arguments brings the presence of another set".

6. Experimental Results

To conduct this study we used WEKA software based on the approach and familiarity with its use. WEKA is an open source tool for data mining, which allows users to apply pre-processing algorithms but it does not provide assistance in terms of which one to apply. However, since different data mining algorithms have different requirements regarding the dataset, some preprocessing is applied by default inside some of the algorithms. Data preprocessing includes cleaning, instance selection, normalization, transformation, feature extraction, selection, etc. Data preprocessing affects the way in which outcomes of the final data processing can be interpreted. WEKA software package has different programs for different techniques and algorithms.