www.ioirp.com

International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 3, Issue – 1. March 2018

PREDICTIVE ANALYTICS IN DIABETES RESEARCH

Munirathinam M
M.E.(CSE)
Bannari Amman Institute of Technology
Sathyamangalam, Erode
Tamil Nadu
munirathinammrm187@gmail.com

K. Premalatha
Professor / CSE
Bannari Amman Institute of Technology
Sathyamangalam, Erode
Tamil Nadu
kpl_barath@yahoo.co.in

Abstract

The remarkable advances in biotechnology and health sciences have led to a significant production of data, such as high throughput genetic data and clinical information, generated from large Electronic Health Records (EHRs). To this end, application of machine learning and data mining methods in biosciences is presently, more than ever before, vital and indispensable in efforts to transform intelligently all available information into valuable knowledge. Diabetes mellitus (DM) is defined as a group of metabolic disorders exerting significant pressure on human health worldwide. Extensive research in all aspects of diabetes has led to the generation of huge amounts of data. A wide range of machine learning algorithms were employed. Support Vector Machines (SVM) arises as the most successful and widely used algorithm. In this work, SVM is applied on DM dataset and the accuracy is measured. The experimental result shows that the accuracy obtained from SVM is 77.1%. For further analysis, ROC, Precision/Recall, Sensitivity/Specificity and Predicted/Observed curves are plotted.

Keywords: Machine learning, Data mining, Diabetes mellitus, SVM

1. INTRODUCTION

Significant advances in biotechnology and more specifically high throughput sequencing result incessantly in an easy and inexpensive data production, thereby ushering the science of applied biology into the area of big data. To date, besides high performance sequencing methods, there is a plethora of digital machines and sensors from various research fields generating data, including super-resolution digital microscopy, mass spectrometry, Magnetic Resonance Imagery (MRI), etc. Although these technologies produce a wealth of data, they do not provide any kind of analysis, interpretation or extraction of knowledge. To this end, the area of Biological Data Mining or otherwise Knowledge Discovery in Biological Data, is more than ever necessary and important. The primary objective is to delve into the rapidly accruing body of biological data and set the basis potentiating answers to fundamental questions in biology and medicine.The power and effectiveness of these approaches are derived from the ability of commensurate methods to extract patterns and create models from data. The aforementioned fact is particularly significant in the big data era, especially when the dataset can reach terabytes or petabytes of data. Consequently, the abundance of data has strengthened considerably data-oriented research in biology. In such a hybrid field, one of the most important research applications is prognosis and diagnosis related to human-threatening and/or life quality reducing diseases. One such disease is Diabetes Mellitus (DM).

Applying machine learning and data mining methods in DM research is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge. The severe social impact of the specific disease renders DM one of the main priorities in medical science research, which inevitably generates huge amounts of data. Undoubtedly, therefore, machine learning and data mining approaches in DM are of great concern when it comes to diagnosis, management and other related clinical administration aspects. Hence, in the framework of this study, efforts were made to review the current literature on machine learning and datamining approaches in diabetes research.

1.1 DATA MINING

Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term “machine learning” is identical to the term “artificial intelligence”, given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience.

Knowledge discovery in databases (KDD) is a field encompassing theories, methods and techniques, trying to make sense of data and extract useful knowledge from them. It is considered to be a multistep process (selection, reprocess, transformation, datamining, interpretation evaluation) depicted in Figure 1. The most important step in the entire KDD process is data mining, exemplifying the application of machine learning algorithms in analyzing data. A complete definition of KDD is given by Fayyad et al. KDD is the nontrivial process identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

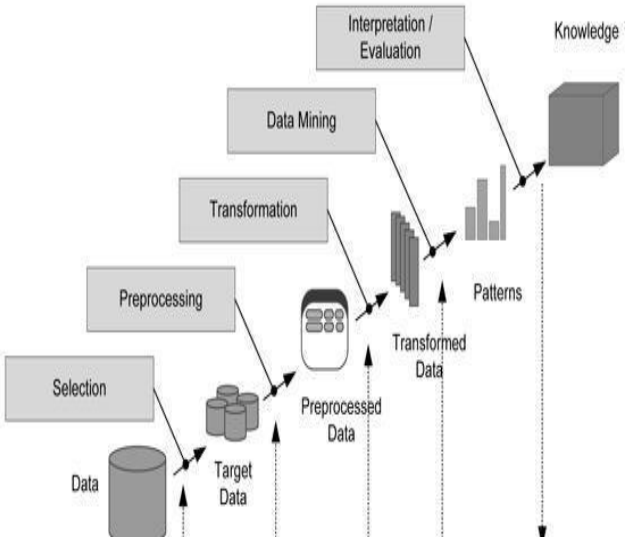

Figure 1 KDD Process

1.2 CATEGORIES OF MACHINE LEARNING TASKS

Machine learning tasks are typically classified into three broad categories. These are:

a) Supervised learning, in which the system infers a function from labelled training data,

b) Unsupervised learning, in which the learning system tries to infer the structure of unlabelled data,

c) Reinforcement learning, in which the system interacts with a dynamic environment.

1.2.1 Supervised Learning

In supervised learning, the system must "learn" inductively a function called target function, which is an expression of a model describing the data. The objective function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values of the function, i.e. its domain, are called instances.

Each case is described by a set of characteristics (attributes or features). A subset of all cases, for which the output variable value is known, is called training data or examples. In order to infer the best target function, the learning system, given a training set, takes into consideration alternative functions, called hypothesis and denoted by h. In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models try to predict distinct classes, such as e.g. blood groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as k-Nearest Neighbours (K-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

1.2.2 Unsupervised Learning

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels.

Association Rule Learning

Association Rule Mining appeared much later than machine learning and is subject to greater influence from the research area of databases. It was proposed in the early 1990s by Rakesh Agrawal as a market basket analysis, in which the aim was to find correlations in the objects of a database. Based on the shopping cart example, association rules are of the form $\{X1, \ldots, Xn\} \rightarrow Y$, which means that if you find all of X1, …, Xn in a cart it is possible to find Y. The most well-known association rule discovery algorithm is Apriori, proposed in 1994 by Rakesh Agrawal.

Clustering

Clusters are informative patterns occurring through Clustering i.e. the separation of a whole dataset into groups of data, so that instances belonging to the same group are as similar as possible and instances belonging to different groups differ as much as possible.

1.2.3 Reinforcement learning

The term Reinforcement Learning is a general term given to a family of techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward. It is important to mention that the system has no prior knowledge about the behaviour of the environment and the only way to find out is through trial and failure (trial and error). Reinforcement learning is mainly applied to autonomous systems, due to its independence in relation to its environment.

1.4 PROBLEM STATEMENT

DM is rapidly emerging as one of the greatest global health challenges of the 21st century. The treatment of Diabetes Mellitus through employment of machine learning and data mining techniques are essential.

1.5 OBJECTIVE

Applying machine learning and data mining methods in DM research is a key approach to utilizing large volumes of available diabetes-related data for extracting knowledge. Techniques are identified to find better prediction accuracy and reduction of computational time.

1.6 DATASET

Diabetes Mellitus (DM) is used to collect for data into huge amount of diabetes records. In this record is required for large number of attributes and labels.

Number of Attributes: 9

Number of records: 769

1.7 PERFORMANCE MEASURE

The following performance measures are used to find the performance of SVM.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{P} \quad (2)$$

$$\text{Specificity} = \frac{TN}{N} \quad (3)$$

$$\text{Error Rate} = \frac{FP+FN}{P+N} \quad (4)$$

Where $TP, TN, FP \text{ and } FN$ represent number of true positives, number of true negatives, number of false positive and number of false negative.

2. LITERATURE REVIEW

2.1 DIABETES MELLITUS

Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells.

T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the

Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L).

DM progression is strongly linked to several complications, mainly due to chronic hyperglycemia. It is well-known that DM covers a wide range of heterogeneous pathophysiological conditions. The most common complications are divided into micro-and macro-vascular disorders, including diabetic nephropathy, retinopathy, neuropathy, diabetic coma and cardiovascular disease. Due to high DM mortality and morbidity as well as related disorders, prevention and treatment attracts broad and significant interest. Insulin administration is the main treatment for T1D, although insulin is also provided in certain cases of T2D patients, when hyperglycemia cannot be controlled through diet, weight loss, exercise and oral medication. Current medication targets primarily a) saving one's life and alleviating the disease symptoms, and b) prevention of long term diabetic complications and/or elimination of several risk factors, thereby increasing longevity.

3. PREDICTIVE ANALYTICS IN DIABETES RESEARCH

3.1 SUPPORT VECTOR MACHINES

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. Figure 2 shows sample example.

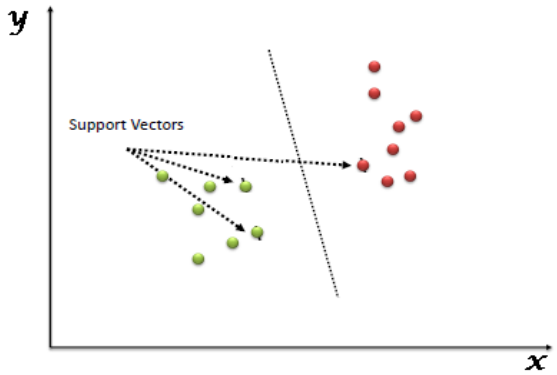

Figure 2 Sample Example for finding the hyper-plane

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

In SVM, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin. In

Figure 3, The margin for hyper-plane C is high as compared to both A and B. Hence, the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin, then there is high chance of miss-classification.

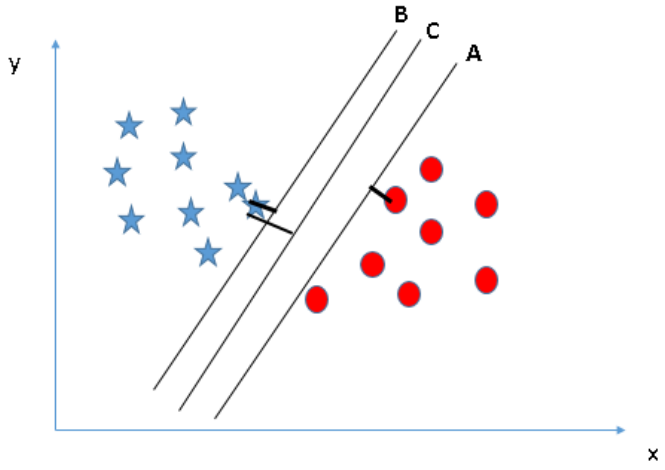

Figure 3 SVM hyper-plane identification

To construct an optimal hyperplane, SVM employs an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, SVM models can be classified into four distinct groups:

- Classification SVM Type 1 (also known as C-SVM classification)
- Classification SVM Type 2 (also known as nu-SVM classification)
- Regression SVM Type 1 (also known as epsilon-SVM regression)
- Regression SVM Type 2 (also known as nu-SVM regression)

3.2 Pros & Cons of Support Vector Machines

Pros

- Accuracy
- Works well on smaller cleaner datasets
- It can be more efficient because it uses a subset of training points

Cons

- Isn't suited to larger datasets as the training time with SVMs can be high
- Less effective on noisier datasets with overlapping classes

3.3 Predictive Analytics using SVM

Figures 4, 5 and 6 show the ROC curve, Precision/Recall Curve, Sensitivity/ Specificity and Predicted vs Observed Model for DM dataset.

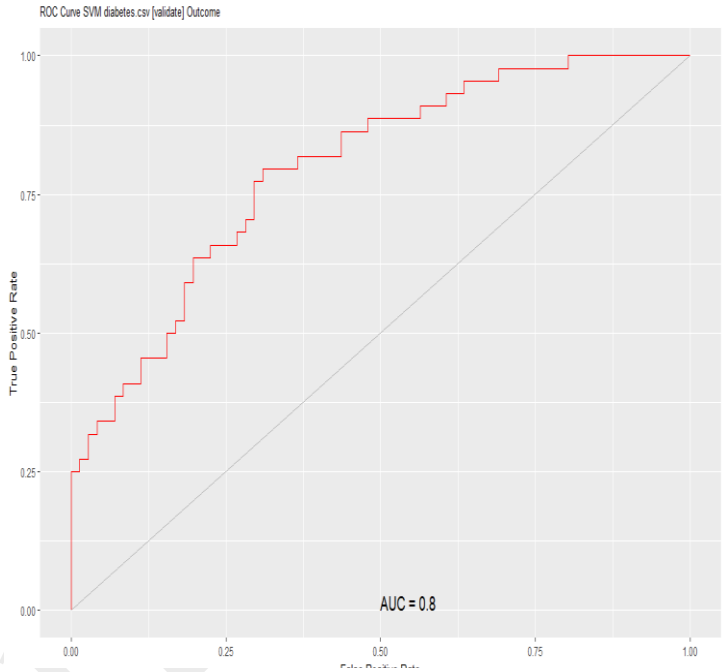

Figure 3 ROC Curve

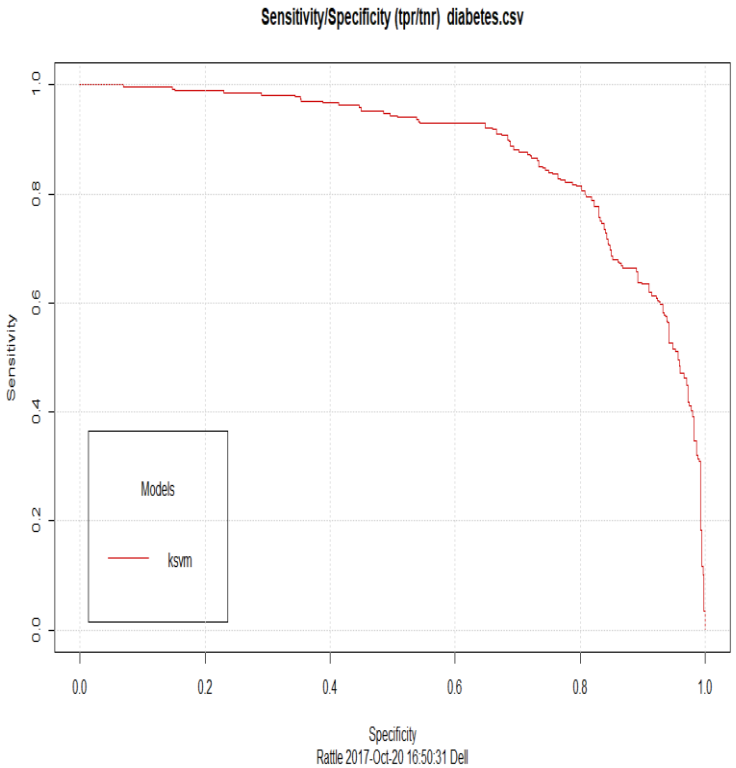

Figure 4 Sensitivity/Specificity Curve

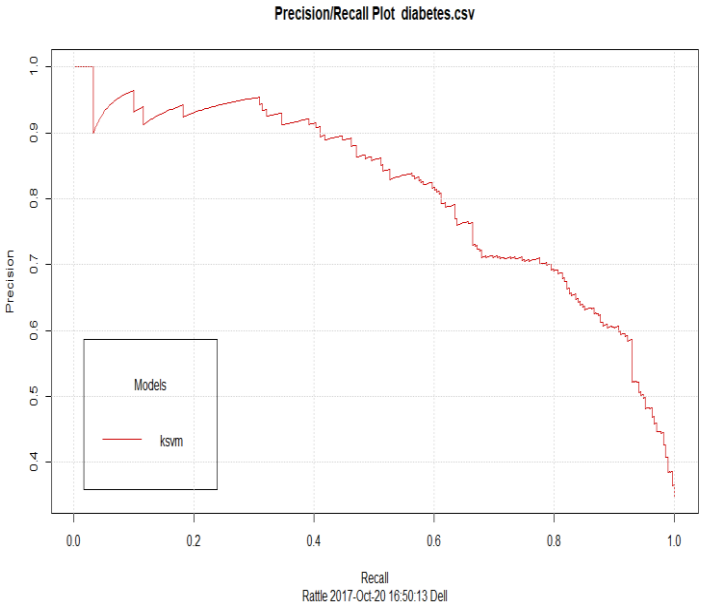

Figure 5 Precision/Recall Curve

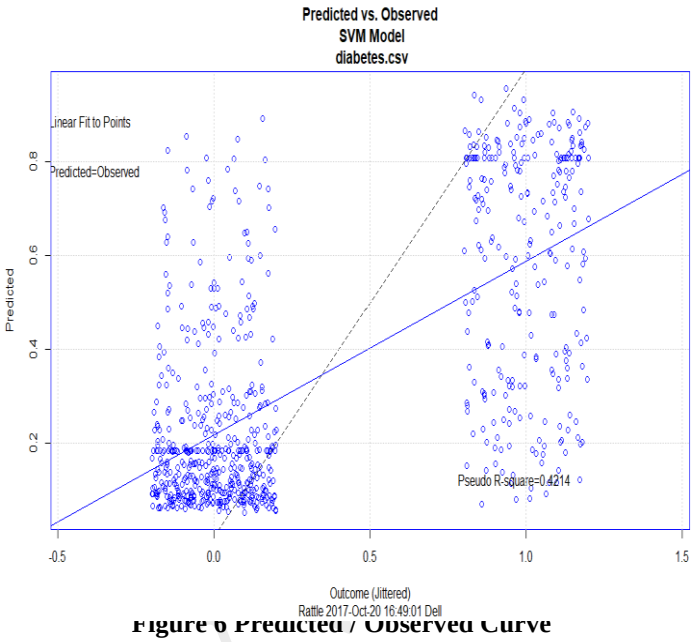

Figure 6 Predicted / Observed Curve

5. CONCLUSION

DM is rapidly emerging as one of the greatest global health challenges of the 21st century. To date, there is a significant work carried out in almost all aspects of DM research and especially biomarker identification and prediction-diagnosis. The advent of biotechnology, with the vast amount of data produced, along with the increasing amount of Electronic Health Records is expected to give rise to further in-depth exploration toward diagnosis and treatment of DM through employment of machine learning and data mining techniques in enriched datasets that include clinical and biological information. In this work, SVM is used to identify the classification accuracy of DM. The accuracy obtained by SVM is 77.1%

REFERENCES

1 Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996;17:37–54.

2 Han J, Kamber M, Pei J. Data mining: concepts and techniques. The Morgan Kaufmann series in data management systems; 2011.

3 Mattmann CA. Computing: a vision for data science. Nature Jan 24 2013; 493(7433):473–5. http://dx.doi.org/10.1038/493473a.

4 Marx V. Biology: the big challenges of big data. Nature Jun 13 2013;498(7453): 255–60. http://dx.doi.org/10.1038/498255a.

5 Mitchell T. Machine learning. McGraw Hill0-07-042807-7; 1997 2.

6 Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd Ed.). Prentice Hall. ISBN 978-0137903955.

7 Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd Ed.). Prentice Hall. ISBN 978-0137903955.

8 Wilson RA, Keil FC. The MIT encyclopaedia of the cognitive sciences.MIT Press;1999