ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

A Survey on Identification of Diabetes Risk Using Machine Learning Approaches

B.Senthil Kumar[1], Sreejith.R[2]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India[1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: Several health oriented studies used machine learning approaches for analysis, detection and prediction of health risks from different attributes of patient health records. Diabetes is one of the common and wide spread health issues in India. Diabetes mellitus type 2 or type2 diabetes is a long term metabolic disorder that is considered by high insulin defiance, lack of insulin and high blood sugar levels. Several machine learning approaches such as supervised learning, clustering and regression etc., have been proposed. This paper surveys different data mining approaches used to handle health care information's along with the result summary. This survey explores the popular and effective machine learning techniques along with its pros and cons.

KEYWORDS: Diabetes, Machine Learning, data mining, classification and prediction.

I. INTRODUCTION

Data mining is a key role in the intelligent health domain [1]. There are several software's and tools have been used to diagnose and classifies health information's based on the attributes. The huge size databases are included into this process as input. This process resulted in data collection complication. The followings are the basic information's about the diabetes and its basic causes and symptoms.

Diabetes risk Prediction Model can support medical professionals and practitioners in predicting risk status based on the clinical data records. In biomedical field data mining and its techniques plays an essential role for prediction and analyzing different type of health issues. The healthcare industry gives huge amounts of healthcare data and that need to be mined to ascertain hidden information for valuable decision selection. Determining hidden patterns and relationships may often very tough and unreliable. The health record is classified and predicted if they have the symptoms of Diabetes risk and using risk factors of disease [2]. It is indispensable to find the best fit algorithm that has greater accuracy, speedy and memory utilization on prediction in the case of Diabetes.

A. DIABETES:

Diabetes is classified into three types:

- **Type 1 Diabetes:** It is a chronic condition in which the pancreas produces little or no insulin. This type of Diabetes results from the pancreas's failure to produce enough insulin. This necessitates the individual to insert insulin or carry an insulin pump. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM). The cause of type 1 diabetes is unknown [3].
- **Type 2 Diabetes:** begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses a lack of insulin may also develop. This form was previously referred to as "non insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The primary cause is excessive body weight and not enough exercise.
- Gestational diabetes is the third main form and occurs when pregnant women without a previous history of diabetes develop high blood-sugar levels.

as a consequence of the human bodies malfunction to generate insulin, and necessitates the individual to insert insulin or carry an insulin pump. This category was previously indicated as "Insulin-Dependent Diabetes Mellitus"

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

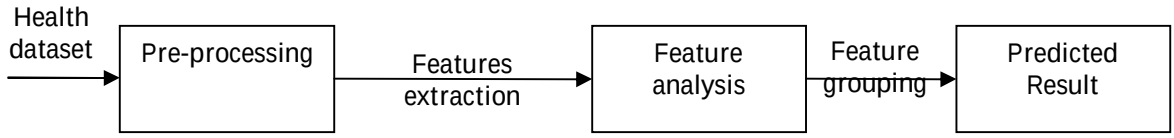
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

(IDDM). The second category of DM is recognized as "Type II DM" as a consequence of insulin confrontation, a situation in which cells are ineffective to exploit insulin appropriately, occasionally merged with an absolute insulin insufficiency. This category also called as "Non Insulin Dependent Diabetes Mellitus" (NIDDM) or "adult-onset diabetes". At last, "gestational diabetes" takes place when conceived women without an earlier

There are several researches conducted epidemiological and public health studies regarding the associations between anthropometric measurements and type2 diabetes. In general the diabetes detection, heart disease risk detection are handled using effective data mining algorithms. The common steps to detect the risk of disease are represented in fig 1.0.



.

Fig 1.0 steps involved in the health risk prediction process

The above fig 1.0 represents the basic process involved in risk prediction in health dataset. In order to predict the risk with its features, this is important to find the pertinent and appropriate features from the health dataset. The above process shows the feature selection and feature based risk calculation for diabetes and other health dataset.

II. DATA MINING IN HEALTHCARE MANAGEMENT

To aid healthcare organization and decision making, data mining applications can be developed to improved classification and track risk of several diseases from patient health records, this types of design, analysis and decision making processes reduces the many work in real time. Consider, in order to create enhanced diagnosis and treatment procedures, this is quite tough when comparing the decision making from scientific literature. In this case, data mining is used to avoid data analysis and clinical problems by managing effective healthcare datasets. Data mining can be used to analyse vast amount of data and statistics to search for patterns [4]. This survey brings the tools and techniques used in the health care management.

In this chapter, we provide a detailed study about the traditional data mining technique in health care application, such as heart disease analysis, type classification, and diabetes and heart risk assessment.

Data mining techniques for Heart Disease:

Healthcare data clustering is the process of segmenting text Healthcare data's into different groups based on its similarity level. The clustering is the unsupervised learning process, where it won't need any training samples for the grouping process [5]. The followings are the popular clustering algorithms are used for Healthcare data management.

There are n number of studies was carried out the heart disease identification and risk prediction using data mining. From the statistical data, the risk factors are associated and from those associations, the risks are detected. The factors such as patient's age, gender, blood pressure, food habits, cholesterol, heredity and hypertension etc., the heart disease details are collected and stored as huge training sample, from the set of data, the useful information's are extracted. In this paper [6] author detected useful patterns from the database and find the risk of heart disease. These works are focused on classification process.

The classification algorithm such as Decision tree [7], naïve bayes[8], SVM [9], and neural network algorithms[10] are included. And some bagging [] and semi supervised classifiers also used to detect heart disease risk.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Table 1.0 data mining algorithms are proposed for heart disease

Algorithm	Description	Papers used the algorithm	Results
Decision Tree	It's a tree like graph model used for classification.	Tu, et al., 2009 (J4.8 Decision Tree) Andreeva, P. 2006 Palaniappan, et al. 2007	The accuracy gained in this paper is 78.9%.
SVM	Support vector machine is a fully supervised learning process. It is a cost effective and suitable for massive data	Kangwanariyakul, et al. 2010 (linear support vector machine)	74.9%
		polynomial support vector machine	70.59%
		radial basis function kernel support vector machine	60.89%
Naïve Bayes	Naive bayes is completely based on the training samples. This is a successful classifier when the training data is huge.	Srinivas, et al. 2010 Naïve Bayes and One Dependency Augmented Naïve Bayes classifier	84.14%
Neural Network	Neural network is dependent with neural schemas, and it divided into linear, probabilistic, radial and polynomial based approaches.	Kangwanariyakul, et al. 2010	76.5%
Bagging	Bagging is the process of iterative classification, where data's are divided and used different classifier	Tu et al., 2009	81.43%

The table 1.0 shows the list of algorithms used to handle the heart disease. The table gives the basic description, paper details along with its result. The results are given in accuracy. The accuracy has been collected from the necessary paper result. The above table clearly shows the basic neural network and SVM can only give limited percentage of accuracy than others. The highest accuracy is bagging method. So the studies in the heart disease and diabetes can be extended from the above result. Several papers used different set of attributes, so the results may vary according to the dataset.

Data mining techniques for Diabetes:

As like the heart disease, the diabetes and risk is classified and predicted by various data mining techniques in the literature such as regression [11], decision tree[12], and Artificial neural networks (ANN)[13].

Algorithm	Description	Papers used the algorithm	Results
ANN (Artificial neural Network)	This follows evolutionary process.	Lee SM, Kang JO 2004	The accuracy gained in this paper is 73.52 %

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Decision Tree (C5.0)	It's a most recent version of decision tree, which has high accuracy and ability to handle missing and null values. C5.0 follows the post-pruning approach, which removes branches from a fully grown tree.	Hu FB, Manson JE,	78.27%
Regression	It's a non linear regression method for predicting diabetes risk. The main advantage of using this is, it supports categorical data.	Lai CL, Lai CL, Chien SW, Fang K. 2007	72.74%

Table 2.0 data mining algorithms are proposed for Diabetes disease

The table 2.0 shows the list of algorithms used to handle the diabetes. Because detection of diabetes is necessary, it generates several diseases such as retinopathy, neurological disorders, eye related issues and stroke etc., The table gives the basic description, paper details along with its result. The results are given in accuracy. The accuracy has been collected from the necessary paper result. The table 2.0 clearly shows the basic ANN and regression can only give limited percentage of accuracy than Decision Tree. The highest accuracy is bagging method.

In paper [14], author compared three prediction models for diabetes using 12 different attributes. The authors have taken C5.0 decision tree algorithm, ANN and regression algorithms. The results from the paper show that the C5.0decision tree model performed best on classification accuracy. This paper concluded and suggests future assist with optimal predictive models. This also proofs the result accuracy varies when the attributes are effectively utilized.

The paper [15] presents the knowledge about the diabetes from the web data such as like MEDLINE. The use of text mining, the reviews and the web data's are mined effectively. Form this, the research outlined the brief knowledge about the diabetes.

SVM also used for diabetes disease detection, in paper [16] proposed supervised learning method, i.e SVM, to improve the terms' discriminating power for disease detection task. This paper utilizes vector space model for text representation, this transforms the content of a Healthcare data into a hyper pane. In this study, the authors investigated numerous unsupervised and supervised classification methods with SVM and ANN algorithms. Finally the paper shows the supervised term weighting methods are good in performance.

VI. CONCLUSION

With the use of data mining approach, health data management grows tremendously. In the modern scenario, diabetes is a common disease spreads to all, and several techniques were implemented with several common and few uncommon features. With the numerous sizes in digital Healthcare data processes, the classification and prediction based on the statistical data is very tough. In such environment effective factors should be identified and that should be used for further disease analysis otherwise any type classification and prediction process is became ineffective.

this paper shows the advantages and disadvantages of several traditional classification algorithms based on different techniques. However, the techniques almost concentrated on general classification process, where the Healthcare data prediciton needs additional concentration and work to improve the following problem. The first problem is discovering appropriate features with less effort and validation on discovered features are not yet studied. And there is a need for a new system to handle the above problem in Healthcare data management disease risk prediction.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

REFERENCES

[1]. Han, j. and M. Kamber, *Data Mining Concepts and Techniques*. 2006: Morgan Kaufmann Publishers. 2. Lee, I.-N., S.-C. Liao, and M. Embrechts, *Data mining techniques applied to medical information.* Med. inform, 2000.
[2]. Obenshain, M.K., *Application of Data Mining Techniques to Healthcare Data.* Infection Control and Hospital Epidemiology, 2004.
[3]. Sandhya, J., et al., *Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques.* International Journal of Engineering and Technology, 2010. Vol.2, No.4.
[4]. Thuraisingham, B., *A Primer for Understanding and Applying Data Mining.* IT Professional IEEE, 2000. 6. Ashby, D. and A. Smith, *The Best Medicine?* Plus Magazine - Living Mathematics., 2005.
[5]. Liao, S.-C. and I.-N. Lee, *Appropriate medical data categorization for data mining classification techniques.* MED. INFORM., 2002. Vol. 27, no. 1, 59–67, .
[6]. Ruben, D.C.J., *Data Mining in Healthcare: Current Applications and Issues.* 2009.
[7]. Porter, T. and B. Green, *Identifying Diabetic Patients: A Data Mining Approach*. Americas Conference on Information Systems, 2009.
[8]. Panzarasa, S., et al., *Data mining techniques for analyzing stroke care processes.* Proceedings of the 13th World Congress on Medical Informatics, 2010.
[9]. Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA, *Data mining techniques for cancer detection using serum proteomic profiling*. Artificial Intelligence in Medicine, Elsevier, 2004.
[10]. Das, R., I. Turkoglu, and A. Sengur, *Effective diagnosis of heart disease through neural networks ensembles.* Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.
[11]. World Health Organization. 2007 7-Febuary 2011]; Available from: http://www.who.int/mediacentre/factsheets/fs310.pdf.