International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Application of Data Mining Techniques for Prediction of Diabetes - A Review

Santosh Rani[1], Dr. Sandeep Kautish[2]
[1]M.Tech Scholar, Department of computer Science and Engineering , Guru Kashi University, Talwandi, Bathinda, Punjab, India
[2]Professor, Department of computer Science and Engineering , Guru Kashi University, Talwandi, Bathinda, Punjab, India

ABSTRACT

During past years, the increase in use of internet and scientific information, enormous amount of data produced every minute and this caused growth in lots of repositories and databases. Health care industry also contains a wealth of database. Millions of patients' database stored in repositories every year. The health care industry ironic in information but knowledge is meager because, with valuable information unsolicited information is also stored and sometimes the valuable information is not analyzed. This unimportant information increases the difficulties of doctors for disease prediction. To solve the problem of this medical mismanagement in health care industry, various machine learning and data mining techniques are used. "Data mining is a process to analyzing the data from large databases. As it is also clear from its name Data Mining searching for valuable information in a large database". Data mining is also known as knowledge discovery. This paper reviews the application and techniques of data mining to determined how these application and techniques have established, during the past years. The main attention is to explore data mining techniques which are widely used to predict some chronic disease like diabetes and cancer, heart attack and for this various most sited research papers of highest journals are reviewed. The techniques of data mining namely, neural networks, association, classification, regression are analyzed in this review paper.

Keywords: Prediction, Association, Classification, Data mining, Regression.

I. INTRODUCTION

In recent [3] time various types of services are emerging in the society. These services are related to the different fields of society. Out of those major field is medical and, in medical disease prediction plays a significant role in health care informatics. It is very essential to diagnose the disease at an early stage. Early detection of disease helps in effective treatment at an initial stage and it is very helpful for both patients and doctors also. There is large amount of data set related to patients, suffering from various diseases all over the world, various types of organized and unorganized medical facilities are available, But due to the population explosion each facility remains scarce and maintenance of this medical dataset has become very decisive task. To overcome and catalyst the growth in this part of the applications. Various researchers are involved which are growing with different researches so that the problem of scarcity of the resources can be catered without increasing the much cost.

Data mining is a process of abstraction, in which massive amount of data is extracted from Different fields. Data is being collected from different fields of health care industry and this data contains massive amount of indispensable and dispensable information so, there is need to extract useful information from this massive amount of data for precise prediction of disease at early stage. It very difficult to read this large amount of data manually and extract useful information from this large dataset. So there is need to use new generation tools and techniques to assist human in extracting vital knowledge from massive amount of data. These tool and techniques are very useful to extract valuable knowledge from large amount of data.

Now health care industry immerging with data mining. So purpose of this paper is to review the various data mining techniques that are used in health care industry to make the data more understandable and useful in prediction of chronic disease like heart disease, diabetes etc.

1.1 Introduction to Data mining

Now a days we deal [1]with massive amount of data generated by search, surfing data, social interactions, transection level data, health care data, enormous amount of sensor data from internet of things, data accumulation from business application, data about what we listen , what we watch, and amount of data in the field of finance, telecommunication and so on. But the problem is people have no time to look at this data and that much of data that is being generated is never analyzed at all.[3] So there is a big gap between generation of data and our understanding of data. Possibly useful knowledge may hide in this data. "By using computational techniques extracting the hidden and valuable information from massive amount of raw data" is called Data Mining. It is a process of analyzing unseen patterns of data according to different perspectives in large data sets involving machine learning, data base system statics. In order to extract valuable information and to mine knowledge from the massive amount of data out there in the world, mining tools are used these mining tools cleanse and purify the raw information and convert it into understandable manner.[4] Data mining is used for discovering valuable patterns and relationship in data to make batter and quick decision. There is massive amount of data being created on a daily basis today, this term is called Big Data. To handle this large amount of data there are several data mining techniques have been developing such as association, clustering, classification, prediction, association, decision tree, sequential patterns.

1.2 Importance of Data Mining

We can simply define data mining as a process that involves searching, collecting, filtering and analyzing the data. It is important to understand that this is not the standard or accepted definition. But the above definition caters for the whole process. Large amount of data can be retrieved from various websites and databases. It can be retrieved in form of data relationships, co-relations and patterns. With the advent of computers, internet and large databases it is possible collect large amounts of data. The data collected may be analyzed steadily and help identify relationships and find solutions to the existing problems. Governments, private companies, large organizations and all businesses are after large volume of data collection for the purposes of business and research development. The data collected can be stored for future use. Storage of information is quite important whenever it is required. It is important to note that it may take long time for finding and searching for information from websites, databases and other internet sources.

1.3 How does Data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two, Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Data mining consists of five major elements:

- ✓ Extract, transform, and load transaction data onto the data warehouse system.
- ✓ Store and manage the data in a multidimensional database system.
- ✓ Provide data access to business analysts and information technology professionals.
- ✓ Analyze the data by application software.
- ✓ Present the data in a useful format, such as a graph or table.

1.4 Process and Discipline involved Data Mining

Data mining [5] is process of discovering patterns from a massive amount of data sets. It is a multidisciplinary field that assimilates various disciplines such as Data warehousing, machine learning, artificial intelligence, neural networks, data base management, data visualization spatial data analysis Probability graph theory.

1.5 Data Mining Techniques

a) Classification

Classification consists [5] [6] of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of pre-classified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

- • Classification of credit applicants as low, medium or high risk
- • Classification of mushrooms as edible or poisonous
- • Determination of which home telephone lines are used for internet access

Is the task of segmenting a diverse group into a number of similar subgroups or clusters? What distinguishes from classification is that does not rely on predefined classes. In, there are no predefined classes. The records are grouped together on the basis of self-similarity. Is often done as a prelude to some other form of data mining or modeling? For example, might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster. The [7][8][9][10]table given below describe classification techniques studied by different authors:

Table 1

Author	Techniques
Md. Maniruzzaman et.el (2017)	
Sajida Perveena(2016)	Adaboots and bagging classifications
Fedi Thabtha (2008)	Associative classification
Roger J. Marshall (2001)	Tree based classification

b) Association

Association is [5] also called relation technique. Because in association a pattern is discovered based on a relationship between items in the same transaction. The structure of neural network involves several input and hidden layers composed with an output layer. The objective of neural network is to diminish the error generated between forecast and desired output

c) Association Rules

An association rule is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form XY, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: 30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to fund all association rules that satisfy user-specified minimum support and minimum confidence constraints. The [11] given table describes some other association techniques studied by different authors:

Table 2

Author	Techniques
Bathany A. (2007)	Spider fear association

d) Regression

Regression [6] is a data mining (machine learning) technique used to fit an equation to a dataset. Regression is a data mining function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of the regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on. In [12][13][14] the table below there are some techniques studied by different authors:

Author	Technique
R.Boukezzoula et.al(2018)	fuzzy and gradual regression
Chang SikKim (2011)	superior regression
D.D.Liu et.al(2011)	Support vector Regression
G. Trepepi et.al (2008)	Leaner and logistic regression.

Clustering

Clustering in [15] data mining is used to discover pattern distribution in essential data. The aim of clustering methods to subdividing the data point s in a class in a way that points which belongs to same class are comparable to different classes. These classes are termed as cluster and number of classes can be determined by algorithm reassigned. The application of clustering is varied in the field as medical, communication, business etc. The other techniques of clustering studied by [16] [17] [18] different authors are stated in given table:

Table 3

Author	Technique
R. Mythily et.al(2015)	Clustering Models for Data Stream
Luca De Angelis et.al (2014)	Heuristic clustering
Steve Russal (1999)	Fuzzy Clustering

Neural Network

One of a typical data mining technique is neural network which can be used in early prediction of disease. It consists a tuple of neuron and each neuron stores its own data. . The structure of neural network involves several input and hidden layers composed with an output layer. The objective of neural network is to diminish the error generated between forecast and desired output. The [19] [20] [21] table given below describe other technique of neural network studied by different authors:

Table 4

Author	Technique
Adrian Horzyk (2017)	fast neural network
Basheer M. et.al(2016)	Artificial Neural Network
Ahmed R. Abas (2013)	Adaptive competitive learning neural networks

II. LITERATURE SURVEY

Pragati Shukla el al. (2018) has proposed Electronic Health Records (EHRs) [22] that are main sources of clinical health system in clinical care and contain patient's historical health data. EHRs can be used to mine the data and make the information more understandable. The system supports the doctors to predict disease more appropriately and make patients benefited. For making information more proficient different data mining techniques have applied with machine learning. The system used C4.5 algorithm and calculate batter accuracy for diabetes prediction and help quick clinical decision making.

Kavakiotis et al. (2017) Defined Diabetes mellitus (DM) as a group of metabolic disorders that exercising pressure on human health worldwide. The aim of this training is to conduct a review of the data mining, machine learning techniques and tools in the area of diabetes research. A tuple of machine learning algorithms were engaged for diabetes prediction. Some of those were characterized by supervised ones and others by unsupervised learning. Machine learning and data mining techniques were used to predict the diabetes by applying different machine algorithms based on biomarker identification. Big data approaches also used to cover the huge amount of data in present world.

Hamadi et al. (2017) has proposed in this study that, for prediction, Artificial Intelligence could be more competent compare to mathematical algorithms. In type 1 Diabetes, for precise blood glucose level prediction three layered (input layer, output layer and hidden) Artificial Neural Network is used that is inspired by human brain and neurons of human brain.12 patient's data was examined for continuous glucose monitoring (CGM) and validate the proposed method. The model predicts blood[i] glucose with minimal error. In future, large amount of patient's data and long period of time can be used for much precise result.

R.M. Khalil and Jumaili (2017) have proposed in this research that depression [23] take a huge impact on human's physical and mental health. The aim of this research to fuse the prediction of depression operation and type 2 diabetes applying machine learning techniques .Supervised machine learning construct a model based on class labels sets to imitate the reality. To give class labels classification technique has been used under testing. To use data supervised learning classifier has been used with modification. Support Vector Machine classifier has been used to attain the high accuracy. After compared four machines learning models the results demonstrate that SVM classifier gives high precision. In future other machine learning procedures can be strained for improved accuracy.

F. Mercaldo. et al. (2017) have projected a method that classify the patients, who are affected and not affected by diabetes, for this authors used a set of characteristics defined by World Health Organization.. 90% patients are affected by type 2 diabetes and rest 10% are by type 1. Type 2 diabetes can be treated with exercise, insulin and drug such as metform in. Real world data has been used in machine learning techniques and results improved from 0.770 to 0.775 by using Hoeffding Tree algorithm. In future other disease take impact to diabetes can be considered.

Han Wu. et al. (2017) have projected Hybrid prediction model based on [28]data mining for type 2 diabetes mellitus prediction in this manuscript . The main purpose of this study is to make the model adaptive and more efficient than other models. Improved K-Mean algorithm is compared with logical reassertion for more accurate results. The main reason for type 2 diabetes is failure of pancreas which causes low insulin production and high glucose availability in body. The main purpose of the model to extract the unknown hidden features from large amount of data set. The model comprises double level algorithm one is improved K- Mean algorithm other is logistic regression algorithm. The model dodges deleting immoderate original data. In future more improved K-Mean algorithm will be used that will less time consuming and less complex

Neesha jothi. et al. (2015) have studies data mining and various data mining techniques like clustering, generalization, association, visualization, pattern matching in this review paper. Data mining and health care industry have developed unfailing detection system. Author reviewed various papers involved in this field in terms of results and algorithms. 50 articles are reviewed from 2005 to 2015.Time period has considered because during the long time many techniques can arise. the review puts a light on introduction of data mining, Disciplines Involved in Data Mining, Data mining models, data mining methods, data mining tasks, various algorithms of data mining as , Anomaly Detection, K-Nearest Neighbor, clustering ,association, swarm intelligence, Decision tree, statistical, classification, Logistic Regression, Bayesian Classifier used in machine learning for health care prediction.

A.dutt. et al. (2015) has studied various types of educational clustering algorithms as applied in Education data mining in this review paper. Educational data mining concentrate on analyzing data set up by numerous disparate systems to develop model for enlightening learning experience. Knowledge discovery database is also used in educational context. Because of multi-level hierarchy educational data mining techniques can be different from standard data mining techniques. The main purpose of this study is to review the different clustering algorithms applied in educational data mining and it studies the various techniques applied in educational data set. Educational data mining change noise data into valuable information, coming from different educational sources that could be useful for learning and strategic learning achievements.

Huang et al. (2014) prolonged unsupervised and semi supervised chores based on multiple regularization so that the applicability of EMSs can be extended impressively. Both algorithms show learning capability of ELMs and handle multiclass clustering and unseen data directly. Both algorithms put into the combined framework that provide perspective understanding .By combining ELM and SVM a new algorithm is proposed named ESVM, PSVM that provide high accuracy then basic ELMs. Compare to unsupervised machine learning this algorithm required less training time and US-ELM have calculated on real world clustering tasks that provides more accuracy then basic ELM.

III. COMPARITIVE ANALYSIS

Author name	Year	Technique	Constraints
Praghti Shukla	2018	Electronic Health[22] Records (EHRs) are used for managing the Patient's health record Electrically. Data Mining, C4.5 and clinical decision support system have been used for early Disease prediction.	Current paper has implemented C4.5 and Decision support System. Have not taken time series based data Processing Considerations.
F.Mercaldo. et.al	2017	For classifying diabetes affected patients this paper used Deep learning, , machine learning techniques with hoeffding tree algorithms	Current paper only considered diabetic patients' data but Other disease that take same impact on health and caused for diabetes symptoms can also be considered.

R.M.Khalil	2017	For prediction of depression [23]operation in type 2 diabetes this paper apply various supervised machine learning techniques as Optimization, F-CMEAN, K-MEAN Probabilistic Neural Network (PNN), SVM(support vector machine)	Diabetes can be associated with other disease for batter prediction.
A. Horzyk and J A.Starzyk	2017	Pulsing model of neurons are used in fast self-organized neural network and such networks are associated temporal storage properties. Associative and computational complexity are main techniques used in his model	Current paper proposed spiking neurons and linear approximation of associative pulsing neurons but other properties of APN with mini-column concept can also be used to increase the memories.
Hang Wu et.al	2017	To predict type two diabetes this [24] paper proposed Hybrid Prediction Model by using K-Mean and logistic regression algorithm of data mining.	Current paper proposed Hybrid Prediction model that is very time consuming and complex but consists less amount of data, large amount of data can be taken with less complexity and minimum time consuming approach.
TakouaHAMDIa	2016	For continues glucose monitoring in type 1 diabetes this paper applied and Artificial Neural Network, Clarke error grid analysis for diabetes prediction.	Current paper has applied the technique only on 12 patients but ANN can be applied on a large number of patients and longtime database of patients.
S.K.Anbananthen .et.al	2005	This paper applied the Artificial Neural network in data mining techniques with decision tree to overcome the “black box” nature of ANN.	Current paper applied decision tree algorithm but clustering and association algorithms can also be implemented for more precise results with ANN

IV. CONCLUSION

From the above study it is clear that data mining is the most important aspect as far as today’s environment is concerned. Trillion of bytes are produced every hour. From such data analysis cannot be drawn out easily. For better analysis various data mining techniques are required. These techniques are being suitable for data mining for purposes like health care, telecommunication, Adhar card etc. these fields can use data mining for prediction purposes like customer or client behavior, client communication pattern, health issues etc. This research area can has large scope for its enhancement in different fields. So that society can take benefit from these kind of analysis.

V. FUTURE WORK

Various association and cluster based techniques are being used for data mining purposes. These techniques put the data in systematic order for better analysis. Further this work can be enhanced by using time series based prediction models. We can apply this model for health care services for prediction system.

VI. REFERENCES

[1]. https://en.wikipedia.org/wiki/Data_mi ning

[2]. Jerome H. Friedman, "Data mining and Statistic: What's Conection?"

[3]. M. Durairaj, V. Ranjani, "Data Mining Applications In Healthcare Sector:" , INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 2, ISSUE 10, OCTOBER 2013 ISSN 2277-8616 29 IJSTR©2013 www.ijstr.org

[4]. Sunny Sharma, "A Study on Data Mining Horizons", International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 02, Issue 04; April - 2016 ISSN: 2455-1457]

[5]. Charu C. Aggarwal and Philip S. Yu, "Data Mining Techniques for Associations, Clustering and Classification", N. Zhong and L. Zhou (Eds.): PAKDD'99, LNAI 1574, pp. 13-23, 1999. c_Springer-Verlag Berlin Heidelberg .

[6]. Andreas Buja , Young-Seop Lee, "Data Mining Criteria for Tree-Based Regression and Classifiaction" ACM , 2001 1-5811-391-x/01/08

[7]. Md. Maniruzzaman et.el , "Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm"S0169-2607 (17)30282-1

[8]. Fadi Thabtah (2008) "Mining the data from a hyperheuristic approach using associative classification", Science Direct ,Expert Systems with Applications 34 (2008) 1093-1101

[9]. Sajida Perveena, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Elsevier, Procedia Computer Science 82 (2016) 115 - 121

[10]. Roger J. Marshall "The use of classification and regression trees in clinical epidemiology", 2001 Elsevier Science Inc., Journal of Clinical Epidemiology 54 (2001) 603-609,

[11]. Bathany A "Evaluating implicit spider fear associations using the Go/No-go Association Task" Elsevier, une 2007, Pages 156-167, volume 38, issue 2

[12]. lRedaBoukezzoula et.al "From fuzzy regression to gradual regression: Interval-based analysis and extensions" Elsevier , May 2018, Pages 18-40, volume 441.

[13]. G. Trepepi et.al (2008)"Linear and logistic regression analysis", Elsevier, 1 April 2008, Pages 806-810, volume 73.issue 7.

[14]. Chang SikKim and SungroLee ,"Spurious regressions driven by excessive volatility" Elsevier December 2011, Pages 292-297, volume 113, issue 3. Support vector regression

[15]. Cheng-Fa Tsai et.al "A New Data Clustering Approach for Data Mining in Large Databases" , Proceedings of the International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN.02) 1087-4089/02 $17.00 © 2002 IEEE

[16]. R. Mythily et.al "Clustering Models for Data Stream Mining", Procedia Computer Science 46 (2015) 619 - 626, Elsevier B.V

[17]. Luca De Angelis and Jose G. Dias b "Mining categorical sequences from data using a hybrid clustering Method" European Journal of Operational Research 234 (2014) 720-730, Science Direct Method" , 0377-2217/$ - see front matter _ 2013 Elsevier B.

[18]. Steve Russell and Weldon Lodwick "Fuzzy Clustering in Data Mining for Telco Database Marketing Campaigns" , 0-7803-521 1 - 4/99/$10.00 0 1999 IEEE

[19]. D.D.Liu et.al "The hourly average solar wind velocity prediction based on support vector regression method" volume: 413, Issue: 4, June 2011)

[20]. Basheer M. Al-Maqaleh, "Forrecasting using artificial Neural Network and Statistic Models" I.J. Education and Management Engineering, 2016, 3, 20-32.

[21]. Adrian Horzyk and Janusz A.Starzyk (2017) Fast neural nerwork

[22]. Adrian Horzyk and Janusz A, "Fast Neural; Network Adaption with Associative Pulsing Neurons", 978-1-5386-2726-6/17/$31.00 ©2017 IEEE. Ahmed R. Abas (2013) Adaptive competitive learning neural networks

[23]. Pragati Shukla, Simran Lal, Gauri Kumbhar, Susmita Kulkarni, Mrs. V. V. Waykule, "DISEASE STATUS PREDICTION AND IDENTIFICATION",IJAREST,Volume 5, Issue 1,pp:10-17,2018.

[24]. Raid M. Khalil,Adel Al-Jumaily,"Machine Learning Based Prediction of Depression among Type 2 Diabetic Patients",ISKE,vol.18,pp:456-466,2017.

[25]. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang,"Type 2 diabetes mellitus prediction model based on data mining",Elsevier,vol.4,pp:190-210,2017.

[26]. Ahmed R. Abas " Adaptive competitive learning neural networks", Egyptian Informatics Journal (2013) 14, 183-194 l1110-8665 _ 2013 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information,

[27]. Enas M.F. El Houby , "A survey on applying machine learning techniques for management of diseases". Elsevier, (2017)

[28]. Takuo Emoto,Tomoya Yamashita,Toshio Kobayashi,Naoto Sasaki,Yushi Hirota,Tomohiro Hayashi, Anna So, Kazuyuki Kasahara,Keiko Yodoi,Takuya Matsumoto,Taiji Mizoguchi,Wataru Ogawa,Ken - ichi Hirata,"Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease",Springer,vol.3,pp:89-100,2016.

[29]. B.M. Patil, Hybrid prediction model for Type-2 diabetic patients. Expert Systems with Applications 37 (2010) 8102-8108.

[30]. Aliza Ahmad and Aida MustaphaH, Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus. ICDIPC 2011, Part I, CCIS 188, pp. 537-545, 2011.

[31]. Alexis Marcano-Cedeño, Joaquín Torres, and Diego Andina, A Prediction Model to Diabetes Using Artificial Metaplasticity. IWINAC 2011, Part II, LNCS 6687, pp. 418-425, 2011.

[32]. Humar, K. and Novruz, A., Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications, 35, 82-89, 2008.

[33]. Rebecca Schnall and Marlene Rojas, A user-centered model for designing consumer mobile health (mHealth) applications (apps). Journal of Biomedical Informatics 60 (2016) 243-2 33Yogita Gupta, Rana Khudhair Abbas Ahmed2 and Dr. Sandeep Kumar Kautish, "APPLICATION OF DATA MINING AND KNOWLEDGE MANAGEMENT IN SPECIAL REFERENCE TO MEDICAL INFORMATICS: A REVIEW" , vol 2, issue 2 , Int J Med Lab Res 2017, 2(2): 60-76