ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

Survey on Data Mining Techniques for Diabetics Research and Analysis

S.Sibi[1], S.Pushpalatha[2]

PG Scholar, Department of Computer Science, PSNA College of Engineering and Technology, Dindigul, India[1, 2]

ABSTRACT Diabetics mellitus is a metabolic disorder in the human body that shows high sugar level in the blood over a long period of time. It could cause many complications in future if it is not treated on time. Data mining plays an important role in the medical field especially in the analysis and prediction of diabetics. Data mining algorithms can be applied on diabetics data to identify useful information that can be used for modelling and diagnosis of diabetics. The various risk factors related to diabetics can also be predicted using the current medical data. Many such data mining approaches have been used for analysis of diabetics. In this paper, a detailed survey has been made on the use of data mining techniques in the field of diabetics. The application of data mining in the diabetics health care has also been discussed along with the various applications. Different data mining strategies that have been used for analysis of diabetics data has been discussed here. The focus of this survey is to make the reader understand about the need and use of data mining techniques to extract useful information for efficient analysis and treatment of diabetics.

KEYWORDS: Data mining; health care; diabetics mellitus; medical analysis

I. INTRODUCTION

Diabetics mellitus is a deadly disorder that refers to the high glucose level or sugar level within the blood and this causes symptoms such as frequent thirst, hunger and urination. The high sugar level in blood stays for a longer period and it should be regulated using certain means such as exercise, diet and medications. Medication involves using of insulin injections to make sure they are properly used up by the body to keep the sugar level low. If not treated for a long-term then it could lead to serious conditions such as kidney disease, ulcer, stroke or even death. The cause for diabetics is because the pancreas does not produce enough insulin and in some other cases the body will not properly accept the insulin that has been generated. In the former case it is called as Type-1 diabetics and the latter is called as the Type-2 diabetics. Another case is when pregnant women without diabetics before generates high sugar levels in blood and this is called as Gestational diabetics which will be cure once the baby has been born.

A diabetic patient should be careful with his everyday activities and should follow a repeated pattern of treatment for keeping the blood sugar level low. These activities include a healthy diet based on current sugar level, insulin injection as prescribed by the doctor, regular exercise and reducing the body weight. Based on these patterns, the sugar level of the patients is recorded over time and a suitable diet or exercise pattern is fixed. These data are recorded for many patients and for many duration of time and are called as the diabetics datasets. By using efficient data mining techniques on these data, useful information can be extracted that can be used for diagnosis of diabetics and also for prediction purposes.

Data mining is the process of extracting useful information from a given set of raw data. Data mining techniques such as classification, clustering, selection, prediction, etc. have been used for this purpose of data analysis. Based on the type of application and the type of data various algorithms have been used under each technique. Data mining has been really useful in the diagnosis of diabetics. In the recent years, many data mining techniques have been proposed for the analysis of diabetics data and for providing useful information for its diagnosis. This can be applied in many diabetics application such as predicting the sugar level based on the

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

patterns of everyday sugar levels, identifying the risk factors, etc. and this will make it easier for the doctors and patients during the treatment time.

The remainder of this paper is organized as follows: Section 2 discuss the survey of various data mining techniques used in diabetics research. Section 3 shows the overall comparison between the existing methods in diabetics research and the inferences made. Section 4talks about the need for data mining in diabetics research and the future of data mining in diabetics research and its application. Finally, Section 5 provides the conclusion and remarks.

II. DATA MINING IN DIABETICS RESEARCH

This section talks about the various applications of data mining in diabetics research and diagnosis. The scope of data mining is large and it has been used in the medical field the most. Data mining in medical and health care is an important and useful research area. One such application of data mining in health case is for analysis of diabetics. Diabetics being a deadly disorder that could even lead to death, there is need for identifying information that can useful for effective treatment for the patients. Also by using certain symptom analysis the risk of diabetics can also be identified or the type of treatment request can also be predicted.

Comparison of three data mining models for predicting diabetes or prediabetes by risk factors.

In this paper, the performance of three data mining model were compared and diabetes, prediabetes were predicted using the 12 common risk factors. Models are Logistic Regression, Artificial neural networks, Decision tree models. These models are constructed using the training dataset and tested by testing dataset. Different risk factor which are used for the predictive analysis are, cigarette smoking, alcohol drinking, consumption of food items, work stress etc., an anthropometric measurements were taken on patients for calculating the weight, height and BMI. The three data mining models are evaluated in terms of Classification, Sensitivity and Specificity. These terms are calculated according to the specific formula. Author used SPSS Modeller Version 4.1 for data mining prediction models. The difference between proportions of the models was examined using the Chi-Square test. As a result, the C5.0 decision tree model is the best predictive model based on the classification accuracy.

Application of data mining: diabetes health care in young and old patients.

The author gives a predictive analysis of diabetic treatment by using a data mining technique called regression. Diabetes is the major health problem in Saudi Arabia and the author collects the dataset from WHO website (World Health Organization).5 age groups has been taken and divided into two groups of young and old groups. The regression technique is basically used for predicting a number; it takes a dataset and develops a formula that which fits a data. ODM (Oracle Data Miner) is software mining tool used for a predictive analysis of treating diabetes. It helps the data analysts to mine their oracle data for finding the hidden information, patterns and new insights. Data mining process of identifying the best mode of treatment for different age group of people are divided into six steps, such as data selection, preparation, analysis, result database knowledge evaluation and pattern prediction and deployment. A support vector machine algorithm is used for learning the classification and regression rule from data. It is implemented into two ways, mathematical programming and kernel functions. Predictive treatment modes like insulin, smoke cessation, exercise; drug, weight and diet are assessed and controlled. The treatment plans are given to both old and young age groups for the reduction of diabetes.

Analysis of diabetic patients through their examination history

A huge amount of medical data is mined to support the physicians and the health care organization. This paper uses a multi-level clustering technique, which helps the patients to identify the pathway that should be followed with a given disease. In this clustering technique a density based algorithm is used to discover the clusters of arbitrary shape and the outlier are filtered out. To measure the analysis of the disease a vector space model is used to represent the data, this model works with the help of TF-IDF method. That is Term-Frequency and Inverse Document Frequency, which identifies the count of the term that occurs in each document. An experimental validation method is used to collect the data of diabetic patients, this approach is used to identify

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

the group of patients with similar examination history, and the process of finding similarity is measured by the method called cosine similarity. Data mining tool called Rapid Miner Toolkit is used in this paper, it is an open source system in which there are number of data mining algorithms are presented and it can automatically analyze the huge dataset and an useful knowledge can be obtained. It uses a DBSCAN algorithm to obtain the useful data that which helps to analysis the diabetic patient through their examination history.

Design of fuzzy classifier for diabetes disease using modified artificial bee colony algorithm

FayssalBelaufa the author proposes an efficient algorithm to improve the performance of diagnosing the diabetes disease, for such purpose a novel Artificial Bee Colony (ABC) algorithm is used. ABC is a recent optimization algorithm that which affect the foraging behavior of a bee colony. The honey bees in the algorithm are classified as employed bees, onlooker bees and scout bees. By this classification a fitness value is used. To improve the performance effectively a modification is made to the ABC algorithm by adding a mutation operator called fuzzy classifier. It consists of linguistic rules which are easy to interpret by the user. A blended crossover operator (BLX-) of a genetic algorithm is used, when the updating of current best solution is failed. By this modified ABC algorithm the diversity of ABC is not enhanced without affecting the quality of the solution. It is a modified version of ABC and it is a tool that automatically create and optimize the membership functions and rules. A diabetes dataset is taken from the UCI machine learning repository. A 10- fold cross validation method is used to evaluate the performance of the proposed system; it is evaluated based on the classification, sensitivity and specificity values. The proposed modified ABC is used as an evolutionary algorithm to create an optimal fuzzy classifier which is used for obtaining the classification values. A fuzzy method with modified ABC algorithm is a powerful tool for diagnosis of diabetes disease.

Realization of a service for the long-term risk assessment of diabetes-related complications

This paper proposes a clinical decision support system for assessing the risk of developing diabetes-related complications in diabetes patients, for such purpose a novel computerized tool is used named as LTRA, which is a long-term risk assessment system. This system core is a set of mathematical models, by this the clinical information of the patient is processed and an estimation of probability over the time of experiencing the adverse event is provided. The major issue occurs when dealing with clinical records is the presence of missing information. Missing data are treated as an intelligent approach; it is based on the average of survival function with the probability distribution of possible realization of the missing information. This missing information module is devised in order to allow the LTRA system to provide risk evaluation even when the information required by the predictive model is missing. The system has been deployed as a web service; it is fully integrated within the diabetic-management platform. A comparison is made between the LTRA system along with UKPDS risk engine (non-diabetes patients), which results in the effective performance of LTRA. But these two methods are performed poorly on the type I diabetes validation cohorts. It occurs due to few events available for this group of patients.

Rule extraction using recursive-rule extraction algorithm with J48 graft combined with sampling selection techniques for the diagnosis of type2 diabetes mellitus in the Pima Indian dataset

Yoichi Hayashi proposed a new algorithm, which extracts the high accurate, concise and interpretable rules for Pima Indian dataset (PID) and the algorithm is sampling Re-Rx with J48 graft algorithm. The type2 diabetes mellitus (T2DM) consist of 90-95% of diabetes cases that are diagnosed for adults. Mostly used diagnostic method for T2DM is a black-box model, but it does not provide a clear reason for underlying diagnosis for the physician. To avoid such problem the rule extraction method is used, it gives more accurate explanation to the physician; it is easy and simple to understand. A model for high accuracy classification is introduced, which a white-box model is named as Recursive-rule extraction (Re-Rx algorithm). As the name implies the recursive nature provides more rules compared to other algorithms. The use of rule extraction algorithm is proposed by combining the Re-Rx with J48 graft along with sampling selection technique; this process helps to obtain the classification rules for PID dataset. As a result of the novel algorithm 83.83% of an average accuracy is obtained after the 10runs of 10-fold cross validation technique. The main drawback of this paper is that the algorithm is not tested on more recent and complete diabetes datasets, if it is done using such data then the accuracy may increase and produces the best algorithm for extracting the accurate rules.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

III. CONCLUSION

Many data mining strategies have been used in different application of diabetics research and analysis. This paper provides a survey on the recent and most useful research in diabetics using data mining techniques. Even still, there is a large gap in this area that will lead to further research in future. Predicting diabetics using the risk factor also have certain drawbacks as the symptoms related to diabetics also is common with other diseases and so there is a need for a better data mining model that can effectively identify the occurrence of diabetics. This can be done by building an effective model that takes into consideration all the factors related to diabetics disorder. This survey provides a path for future research in application of data mining in diagnosis of diabetics.

REFERENCES

[1] Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients",2012, Journal of King Saud University – Computer & Information Sciences., 1319-1578. http://dx.doi.org/10.1016/j.jksuci.2012.10.003

[2] Dario Antonelli , Elena Baralis, Giulia Bruno, Tania Cerquitelli , Silvia Chiusano, NaeemMahoto.," Analysis of diabetic patients through their examination history".,2013.,ELSEVIER-Expert Systems with Applications., 0957-4174., http://dx.doi.org/10.1016/j.eswa.2013.02.006

[3] Kamadi V.S.R.P. Varmaa, AllamAppaRao, ThummalaSitaMahalakshmi , P. V. NageswaraRao.," A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach".,2016.,ELSEVIER- Applied Soft Computing., 1568-4946., http://dx.doi.org/10.1016/j.asoc.2016.05.010

[4] Fayssal Beloufa, M.A. Chikh.," Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm".,2013., ELSEVIER- 0169-2607.,http://dx.doi.org/10.1016/j.cmpb.2013.07.009

[5] Iván Contreras, Carmen Quirós, MargaGiménez, Ignacio Conget, JosepVehi .," Profiling intra-patient typeI diabetes behaviors".,2016.,ELSEVIER-01692607.,http://dx.doi.org/10.1016/j.cmpb.2016.08.022

[6] Yoichi Hayashi , ShonosukeYukita.," Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset".,2016.,ELSEVIER- Informatics in Medicine Unlocked., 2352-9148.,http://dx.doi.org/10.1016/j.imu.2016.02.001

[7] Rahul Isola, RebeckCarvalho, Amiya Kumar Tripathy.," Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and k-NN", 2012, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 16, NO. 6, 1089-7771

[8] Vincenzo Lagani, Franco Chiarugi, DimitrisManousos, VivekVerma, Joanna Fursse , Kostas Marias, IoannisTsamardinos.," Realization of a service for the long-term risk assessment of diabetes-related complications".,2015., ELSEVIER- Journal of Diabetes and Its Complications.,1056-8727.,http://dx.doi.org/10.1016/j.jdiacomp.2015.03.011

[9] Simone Marini, EmanueleTrifoglio, Nicola Barbarini, FrancesoSambo, Barbara Di Camillo, Alberto Malovini, Marco Manfrini, Claudio Cobelli , Riccardo Bellazzi.," A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes".,2015., ELSEVIER- Journal of Biomedical Informatics., 1532-0464. http://dx.doi.org/10.1016/j.jbi.2015.08.021

[10] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu.," Comparison of three data mining models for predicting diabetes or prediabetes by risk factors".,2012., ELSEVIER-1607-551X., http://dx.doi.org/10.1016/j.kjms.2012.08.016

[11] C. Novara, N. Mohammad Pour, T. Vincent, G. Grassi.," A Nonlinear Blind Identification Approach to Modeling of Diabetic Patients", 2015, IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY., 1063-6536.

[12] Gyorgy J. Simon, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro and Peter W. Li.," Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus", 2013, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 1041-4347.

[13] S. Upadhyaya , K. Farahmand, T. Baker-Demaray.," Comparison of NN and LR classifiers in the context of screening native American elders with diabetes".,2013.,ELSEVIER- Expert Systems with Applications., 0957-4174.,http://dx.doi.org/10.1016/j.eswa.2013.05.012

[14] Kung-Jeng Wang, Angelia Melani Adrian, Kun-Huang Chen, Kung-Min Wang.," An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus".,2015., ELSEVIER- Journal of Biomedical Informatics., 1532-0464 http://dx.doi.org/10.1016/j.jbi.2015.02.001

[15] Akay, Dragomir, Björn-Erik Erlandsson.," A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin".,2013., IEEE Journal of Biomedical and Health Informatics., 2168-2194., DOI -10.1109/JBHI.2013.2295834

[16] Nahla H. Barakat, Andrew P. Bradley, Mohamed Nabil H. Barakat," Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", 2010., IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 14, NO. 4, 1089-7771.

[17] Ivan Contreras, JosepVehi," Using Normalised Compression Distance to Identify Different Profiling Days in type 1 Diabetic Patients", 2015.,ELSEVIER 2405-8963.

[18] Simon Fong , Sabah Mohammed , Jinan Fiaidhi , CheeKeongKwoh.," Using causality modeling and Fuzzy Lattice Reasoning algorithm for predicting blood glucose".,2013., ELSEVIER- Expert Systems with Applications., 0957-4174.,http://dx.doi.org/10.1016/j.eswa.2013.07.035

[19] AdemKarahoca , M. AlperTunga.," Dosage planning for type 2 diabetes mellitus patients using Indexing HDMR".,2012., ELSEVIER- Expert Systems with Applications., 0957-4174.,doi:10.1016/j.eswa.2012.01.056

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 1, March 2018

6th National Conference on Frontiers in Communication and Signal Processing Systems (NCFCSPS '18)

13th-14th March 2018

Organized by

Department of ECE, Adhiyamaan College of Engineering, Hosur, Tamilnadu, India

[20] Kamadi V.S.R.P. Varma, AllamAppaRao, T. SitaMaha Lakshmi, P.V. NageswaraRao," A computational intelligence approach for a better diagnosis of diabetic patients",2013, ELSEVIER- Computers and Electrical Engineering., 0045-7906., http://dx.doi.org/10.1016/j.compeleceng.2013.07.003.

[21] AzraRamezankhani, OmidPournik, Jamal Shahrabi, DavoodKhalili, FereidounAzizi, FarzadHadaegh," Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study", 2014, ELSEVIER- Diabetes Research and Clinical Practice., 0168-8227., http://dx.doi.org/10.1016/j.diabres.2014.07.003.

[22] Hamid R. Marateb, MarjanMansourian , ElhamFaghihimani, MasoudAmini, Dario Farina.," A hybrid intelligent system for diagnosing microalbuminuria in type 2 diabetes patients without having to measure urinary albumin".,2013., ELSEVIER- Computers in Biology and Medicine., 0010-4825.,http://dx.doi.org/10.1016/j.compbiomed.2013.11.006