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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Prediction of Diabetes Risk Factor Using Back Propagation and C 4.5 Algorithm

Aparna Phalak[1], Priti Sharma[2]

Research Scholar, Department of Computer Engineering, SSBT's COET, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, SSBT's COET, Maharashtra, India[2]

ABSTRACT: Prediction of diabetes is the most essential factor in Hospitalization. It is the proper way to predict more than 500 people diabetes at a time. Due to improper prediction of diabetes it may affect to socialization and localization, because in news or newspaper diabetes prediction rate news arrive and this prediction rate is calculated by using simple paper work so the lots of time takes to calculate the diabetes prediction rate so the focus of healthcare from reactive and hospital based to more proactive and patient-based. Electronic medical records contain patient demographics, progress notes, problems, and medications, vital signs, past medical history, immunizations, laboratory data and radiology report to our diabetes prediction system. There is a framework that predict the diabetes and enables the representation, extraction, and mining of high order latent event structure and relationships within single and multiple event sequences by mapping the heterogeneous event sequences to a geometric image by encoding events as a Bar graphs or bar charts. For creating diabetes prediction system two algorithms are used for prediction which are Back propagation and C 4.5 algorithms. These algorithms help to predict the diabetes and save time and complicated work and give the perfect solution for predicting more than 500 people diabetes.

KEYWORDS: Temporal signature mining, sparse coding, SVM, Logistic Regression, Back propagation algorithm, C 4.5 Algorithm

I. INTRODUCTION

Data mining can be defined as an activity that extracts some new nontrivial information Contained in large databases. The goal is to discover hidden patterns, unexpected trends or other subtle relationships in the data using a combination of techniques from machine Learning, statistics and database technologies. This new discipline today finds application in a wide and diverse range of business, scientific and engineering scenarios. For example, large databases of loan applications are available which record different kinds of personal and financial information about the applicants (along with their repayment histories). These databases can be mined for typical patterns leading to defaults which can help determine whether a future loan application must be accepted or rejected. Several tera bytes of remote sensing image data are gathered from satellites around the globe. Data mining can help reveal potential locations of some (as yet undetected) natural resources or assist in building early warning systems for ecological disasters like oil slicks etc. Other situations where data mining can be of use include analysis of medical records of hospitals in a town to predict, for example, potential outbreaks of infectious diseases, analysis of customer transactions for market research applications etc.

Temporal data mining is concerned with data mining of large sequential data sets. By sequential data, they mean data that is ordered with respect to some index. For example, time series constitute a popular class of sequential data, where records are indexed by time. Other examples of sequential data could be text, gene sequences, protein sequences, lists of moves in a chess game etc. Here, although there is no notion of time as such, the ordering among the records is very important and is central to the data modelling [1].

EHR are electronic versions of the paper charts in your doctors or other health care provider's office. An EHR may include your medical history, notes, and other information about your health including your symptoms, diagnoses, medications, lab results, vital signs, immunizations, and reports from diagnostic tests such as x-rays. Providers are working with other doctors, hospitals, and health plans to find ways to share that information. The information in EHR can be shared with other organizations involved in your care if the computer systems are set up to talk to each other.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Information in these records should only be shared for purposes authorized by law or by you [1]. You have privacy rights whether your information is stored as a paper record or stored in an electronic form. The same federal laws that already protect your health information also apply to information in.

In EHR data, each record (data instance) consists of multiple time series of clinical variables collected for a various patients, such as results of tests in laboratory and medication orders. The record may also provide information about patient‘s diseases and adverse medical events over time. Finding latent temporal signatures is important in many domains as they encode temporal concepts such as event trends, episodes, cycles, and abnormalities. Temporal data mining is concerned with data mining of large sequential data sets. By sequential data, we mean data that is ordered with respect to some index. For example, time series constitute a popular class of sequential data, in which records are arranged by time. Sequential data could be text, gene sequences, amino acid sequences, and moves in a puzzles or chess game. The ordering among the records is very important for the data description/modelling. Time series analysis has a long history. Techniques for statistical modelling and spectral analysis of real or complex-valued time series have been in use for more than fifty years. Temporal data mining methods must be capable of analysing data sets that are prohibitively large for conventional time series modelling techniques to handle efficiently. Temporal event signature mining for knowledge discovery is a difficult problem. Due to vast amounts of complex event data it is challenging for humans and also for data and information analysis by machines. An appropriate knowledge representation for mining longitudinal event data is important. This paper gives possibility is to provide interactive and user friendly representation of Knowledge and data with Visual Data Analytics.

II. RELATED WORK

The prediction of the Diabetes Detection is an interesting task for researchers. In the literature, number of methods are applied to accomplish this task. The prediction methods use various approaches, from highly informal ways to more formal ways (e.g. linear or non-linear regressions).

The prediction methods on the basis of type of data and the type of tool that each method is using to predict the Diabetes risk factor which are categorized as: Technical Analysis Methods, Fundamental Analysis Methods, Traditional, Time Series Prediction Methods, Machine Learning Methods

The common thing between these techniques is that they are used to predict and thus can have benefit from the patients healthbehaviour.

There are mainly 2 models involved in the operational research as The Operational Research as Classification and Prediction models such that Fuzzy model and classifier model. In classifier model two classifier models are covered which are probabilistic model and statistical Model.

1. Fuzzy Models: In this type of model the uncertain data is considered. It explicitly represents uncertain data via random variables or stochastic processes. It characterizes the system performance and also estimates the performance of the system. And Fuzzy models operate on information granules that are fuzzy sets and fuzzy relations. in fuzzy model Information granules are abstract realizations of concepts used in modelling As modelling is realized at higher, more abstract level, fuzzy model models give rise to a general architecture in which they highlight three main functional modules, that is input interface, processing module, output interface The goal of this notebook is to demonstrate how Fuzzy Logic can be used for modelling. They plan to show how fuzzy sets can be used to represent a real system or process. In this demonstration, they use modelling data from a paper in which the authors investigate the effects of using various fuzzy operators for constructing models [Stachowicz and Kochanska, 1987]. To demonstrate fuzzy modelling, they use many functions from Fuzzy Logic, along with standard Mathematics functions. Fuzzy Logic contains numerous functions for working with fuzzy sets and fuzzy logic.

2. Classifier Models: classifier Models consist of a number of dimensions you can look at to give you a sense of what will be a reasonable algorithm to start with, namely: Number of training examples, Dimensionality of the feature space, Do I expect the Problem to be linearly separable?, Are features independent?, Are features expected to Linearly dependent with the target variable?, Is over fitting expected to be a problem?, What are the system's requirement in terms of speed/performance/memory usage? All those questions have one solution that is classifier algorithm. Speed memory, input target variable, output target variable, over fitting problem all these problems are solve by using

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

classifier algorithm. It gives the optimal solution among the different types of solutions which is called as best possible solution.

III. PROPOSED ALGORITHM

This section gives the detailed description of the Back propagation algorithm and C4.5 algorithm. Back propagation algorithm takes the inputs of the training and testing dataset these dataset contains the value of patient Age, Sex, HBA1C, Resting BP, Plasma Glucose, Cholesterol, Pulse Rate, Hypertension, Date, Foot Ulcers, Severity Index, Treatment. The objective of the training is to minimize the divergence between real data and the output of the network. This principle is referred to as Supervised Learning

The algorithm of the Back propagation algorithm is as below:

Algorithm 1 Back Propagation Algorithm

1. Accept input sample
2. Perform its weighted summation.
3. Apply it to input layer neurons
4. Process all inputs at each neuron by transfer function to get individual.
5. Hidden layer and repeat 1,2,3,4 steps pass it as an input to all neurons of for hidden layer neurons.
6. Pass output of hidden layer neurons to all output layers and repeat 1,2,3,4 steps to get final output.
7. Display the final output

C 4.5 Algorithm

This section gives the detailed description of the C4.5 algorithm to solving problem of visualization so many information visualization techniques have been developed to support the exploration of large data sets. There are various interactive visual data mining tools available for visual data analysis. It is possible to perform clinical assessment for visual interactive knowledge discovery in large electronic health record databases. but C 4.5 algorithm have some good feature and solve speed, memory, rule set, missing values, and over fitting problem that's the reason use c 4.5 algorithm. The algorithm is as shown in Algorithm 2. The flowchart of the Back propagation algorithm is as shown

Integrated Approach

This section gives the detailed description of the combination of Back propagation and C 4.5 algorithm. to solving problem of Diabetes Prediction and visualization of diabetes risk factor within 3 4o 4 minutes. By creating the combination of the back propagation and c 4.5 algorithm the accuracy of the prediction is improved and this is the better way to create framework to predict the diabetes and enables the representation, extraction, and mining of high order latent event structure and relationships within single and multiple event sequences Algorithm 2 C 4.5 Algorithm

Require: Id (Target Attribute, Attribute)

1. Create a root node for the tree
2. Check for the base case
3. Apply Feature Selection using Genetic Search
4. best Tree = Construct a DT using training data
5. Perform Cross validation

    Divide all examples into N disjoint subsets, E = E1, E2, to EN

    For each i = 1 TO N do

(a) Test set = Ei

(b) Training set = E - Ei

(c) Compute decision tree using Training set

(d) Determine performance accuracy Pi using Test set

    Compute N-fold cross-validation estimate of performance = (P1 + P2 + ... + PN)/N

6. Perform Reduced Error Pruning technique
7. Perform Model complexity

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

8. Find the attribute with the highest info gain (A Best)
9. Partition S into S1, S2,S3 To Sn according to the value of A Best
10. Repeat the steps for S1, S2, S3
11. Classification

IV. SIMULATION RESULTS

Experimental result shows the effectiveness of proposed system, in which the training and testing dataset is provided and single patient entry as well as multiple patient entry is uploaded in database. Figure1 shows Classification of diabetes detection using Back propagation and c 4.5 algorithm. This figure shows the 50% people have infected from HBA1c, 15% people have infected from Insulin Glargin,7% people have infected from NPH Insulin, 10% people have infected from fasting plasma glucose, 18% people have infected from Random plasma glucose. All these infection shows those people have detect diabetes by various infection.

Results of experiment state that performance of back propagation and c 4.5 algorithm is better than other algorithm. Performance get reduce when there are improper attributes are added. Accuracy of diabetes detection is better than the previous TEMR system, accuracy is improving to 99% as compare to existing approach. Figure 3 shows accuracy of diabetes detection using Back propagation and c 4.5 algorithm

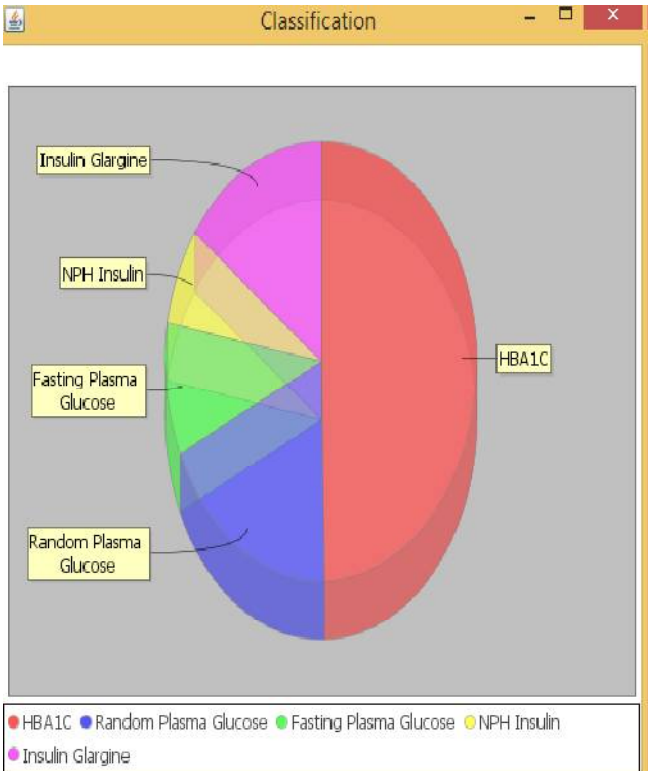

Figure 1: classification of diabetes detection

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

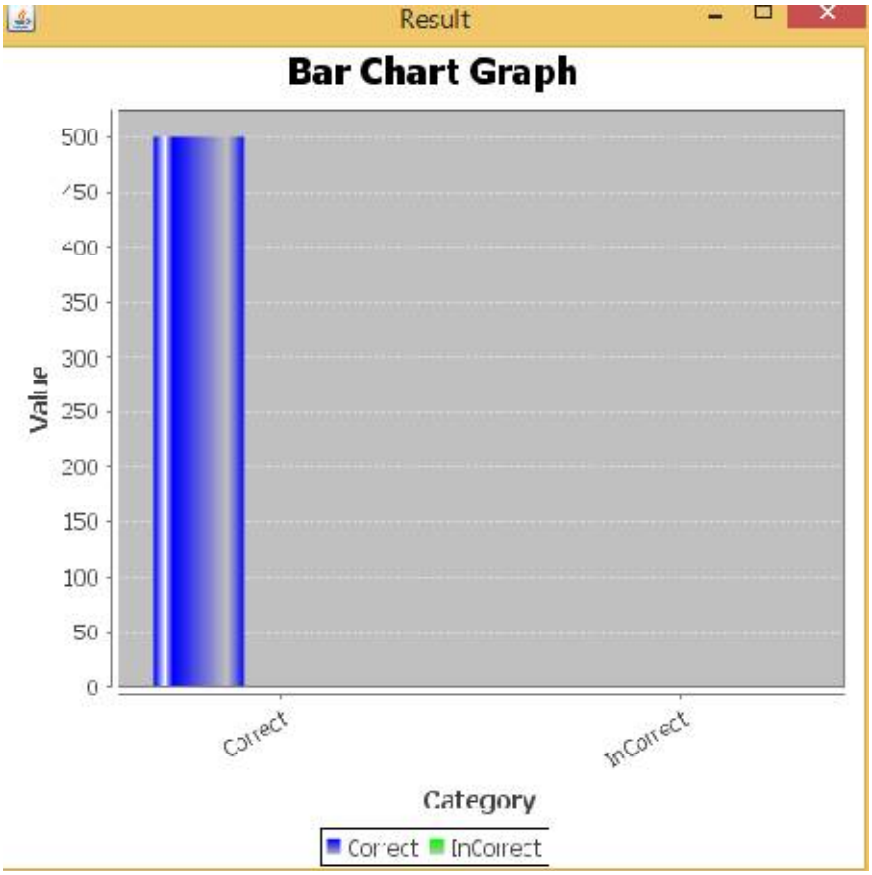

Figure 2: accuracy of diabetes detection

V. CONCLUSION AND FUTURE WORK

The diabetes prediction for more than 500 people within 3 to 4 minutes is very difficulttask Back propagation and C 4.5 algorithm combines for fast automatic data mining algorithm the intuitive power of the human mind, which improve the quality and speed of the data mining process. The algorithms considers the Age,Sex,HBA1C,Resting BP, Plasma Glucose, Cholesterol, Pulse Rate, Hypertension, Date, Foot Ulcers, Severity Index, Treatment attributes for predicting the diabetes. The goal of predicting the diabetes is to bring the power of prediction is to every desktop to allow accurate, better, faster and intuitive result. Classification techniques are used to manage the vast amounts of relational and semi structured information, including database management and data warehouse systems.

In future different dataset will be used for comparing values and find out exact solutionand multiple algorithms will be used for improving the accuracy of algorithm.

REFERENCES

[1] K. M, M. J. and Lazarus R, and et al, Integrating clinical practice and public health surveillance using electronic medical record systems," PubMed, p. S154S162, 2012.
[2] V. K, H. RF, and L. D, Increasing incidence of type 1 diabetes in 0 to 17 year old colorado youth," PubMed, no. 30, p. 503509, 2007.
[3] Z. Xu, X. Qi, A. K. Dahl, and W. Xu, Waist to height ratio is the best indicator for undiagnosed type 2 diabetes," Diabetic Med, June 2013.
[4] R. N. Feng, C. Zhao, C. Wang, Y. C. Niu, K. Li, F. C. Guo, S. T. Li, C. H. Sun, and Y. Li, Bmi is strongly associated with hypertension and waist circumference is strongly associated with type 2 diabetes and dyslipidemia, in northern chinese adults,"SIGCOMM Comput. Commun. Rev., May.
[5] [Online]. Available: http://care.diabetesjournals.org

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

[6] B. J. Lee and J. Y. Kim, Identification of type 2 diabetes risk factors using phenotypes Consisting of anthropometry and triglycerides based on machine learning," IEEE, vol. 20, no. 1, 2016.
[7] P. T. Katzmarzyk, C. L. Craig, and L. Gauvin, Adiposity, physical fitness and incident diabetes: The physical activity longitudinal study," Computer Communications, March 2007.
[8] G. F. A. Berber, R. Gomez-Santos and L. Sanchez-Reyes, Anthropometric indexes in the prediction of type 2 diabetes mellitus, hypertension and dyslipidaemia in a mexican population," 2001), MONTH = December volume = 25, number = 12, pages =17941799.
[9] B. Balkau, D. Sapinho, A. Petrella, L. Mhamdi, M. Cailleau, D. Arondel, and M. A.Charles, Prescreening tools for diabetes and obesity associated dyslipidaemia comparing bmi, waist and waist hip ratio," vol. 60, no. 3, pp. 295304),, March 2006.
[10] |Prescreening tools for diabetes and obesity associated dyslipidaemia comparing bmi, waist and waist hip ratio," vol. 60, no. 3, p. 295304, March 2006.
[11] I. S. Okosun, K. M. Chandra, S. Choi, J. Christman, G. E. Dever, and T. E. Prewitt,Hypertension and type 2 diabetes comorbidity in adults in the united states: risk of overall and regional adiposity," vol. 9, no. 1, p. 19, January 2001.
[12] L. A. Sargeant, F. I. Bennett, T. E. Forrester, R. S. Cooper, and R. J.Wilks, Predicting incident diabetes in jamaica: the role of anthropometry," August.
[13] N. T. D. S. le, T. T. Hanh, K. Kusama, D. Kunii, T. Sakai, N. T. Hung, and S. Yamamoto, Anthropometric characteristics, dietary patterns and risk of type 2 diabetesmellitus in vietnam," Auguest.
[14]S.A.Sarwade and R. Makhijani, A review on mining signatures from event sequencesand visual interactive knowledge discovery in large electronic health record database,"ISSN, vol. 3, December 2013. [Online]. Available: www.ijarcsse.com
[15] F.Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi, and A. F. Laine, A framework for mining signatures from event sequences and its applications in healthcare data," vol. 35, no. 2, February 2013.
[16] B. Cao, D. Shen, J. Sun, X. Wang, Q. Yang, and Z. Chen, Detect and track latentfactors with online nonnegative matrix factorization," pp. 2689{2694, 2007.
[17] F. Chung, Spectral graph theory," 1997.
[18] C. Ding, T. Li, and M. Jordan, Pattern analysis and machine intelligence," IEEE,vol. 32, no. 1, pp. 45{55, January 2010.
[19] N. Robinson, The disadvantages of logistic regression."
[20] S. Singh and P. Gupta, Comparative study id3, cart and c4.5 decision tree algorithm:A survey," ISSN, vol. 27, no. 27, July 2014.
[21] L. Auria and R. A. Moro, Support vector machines as a technique for solvency analysis 811," ISSN, Auguest 2008.
[22] F. Wang, C. Liu, Y. Wang, J. Hu, and G. Yu0, A graph based methodology fortemporal signature identification from ehr," Online, November 2015.
[23] A. Garg and D. Roth, Understanding probabilistic classifiers," ECML, no. 1, pp.
[24] L. Auria and R. A. Moro, Support vector machines as a technique for solvency analysis 811," ISSN, pp. 1,16, August 2008.
[25] K. M. Leung, Naive Bayesian classifier," no. 1, pp. ,16, November 2007.
[26] T. M. Mitchell, \Generative and discriminative classifiers: Naive Bayes and logistic Regression," pp. 1,17, February 2016. [Online]. Available: www.cs.cmu.edu
[27] J. Eggert and E. Korner, Sparse coding and nmf," vol. 2, no. 1, pp. 2529{2533, 2004.
[28] W. Fei, L. Ping, and K. Christian, Online nonnegative matrix factorization for document clustering," 2011.
[29] B. Hoang, Ashley caudill originally published on the ieee emerging technology portal,"pp. 2006,2012. [Online]. Available: http://www.ieee.org/go/emergingtechE
[30] J. N. E. Sharon SilowCarroll and D. Rodin, Health management associates using electronic health records to improve quality and efficiency: The experiences of leading hospitals," July 2012.