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. 4, Issue 12, December 2016

Survey Paper on Diabetes Detection Using Back Propagation and C4.5 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: Electronic health records allow for complete and up-to-date medical information to be stored and made available to patients and health care providers. In addition to improving care and reducing costs and clinical errors by providing a patient's complete medical history, electronic health records and other forms of telemedicine shift 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 reports. There is a framework that 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 structured spatial-temporal shape process. in this paper, we proposed that it is possible to develop a tool for data visualization for interactive knowledge discovery by using Back propagation and c4.5 algorithm. Data Visualization is very helpful for analysts to visually discover different kinds of patterns such as clusters, relationships and associations.

KEYWORDS: Temporal signature mining, sparse coding, nonnegative matrix factorization, beta-divergence, Visual Analytics, Information visualization.

I. INTRODUCTION

In EHR information, every record comprises of numerous time arrangement of clinical factors gathered for a different patients, for example, consequences of tests in research facility and prescription requests. The record may likewise give data about patient's infections and unfriendly restorative occasions after some time. Finding inactive transient marks is critical in numerous spaces as they encode worldly ideas, for example, occasion patterns, scenes, cycles, and anomalies. Transient information mining is worried with information mining of substantial successive information sets. By successive information, we mean information that is requested concerning some list. For instance, time arrangement constitute a well-known class of consecutive information, in which records are organized by time. Successive information could be content, quality arrangements, amino corrosive groupings, and moves in a riddles or chess diversion. The requesting among the records is vital for the information portrayal/demonstrating. Time arrangement investigation has a long history. Procedures for factual demonstrating and phantom investigation of genuine or complex-esteemed time arrangement have been being used for over fifty years. Fleeting information mining strategies must be equipped for dissecting information sets that are restrictively huge for traditional time arrangement displaying methods to handle productively. Transient occasion signature digging for learning revelation is a troublesome issue. Because of immense measures of complex occasion information it is trying for people furthermore for information and data examination by machines. A proper learning representation for mining longitudinal occasion information is vital. This paper gives probability is to give intelligent and easy to understand representation of Knowledge and information with Visual Data Analytics.

Diabetes mellitus (DM) is portrayed by constant hyperglycemia and impeded starches, lipids, and proteins digestion system brought on by total or fractional inadequacy of insulin emission as well as insulin activity. There are two essential types of diabetes, insulin-subordinate diabetes mellitus (sort 1 diabetes mellitus, T1DM) and non-insulin-subordinate diabetes mellitus (sort 2 diabetes mellitus, T2DM). T2DM is the most widely recognized type of DM, which represents 90% to 95% of every diabetic patient and is relied upon to increment to 439 million by 2030. In China, the most recent measurable information demonstrate that diabetes and pre-diabetes are predominant among

DOI: 10.15680/IJIRCCE.2016. 0412127

individuals more seasoned than 20-year-old, with the rates being 9.7% and 15.5% for T1DM and T2DM, separately. T2DM for the most part results from the communication among hereditary, ecological and other hazard components. Moreover, loss of first-period of insulin discharge, strange pulsatility of basal insulin emission, and expanded glucagon discharge likewise quicken the improvement of T2DM [4, 5]. In spite of the fact that T2DM patients are by and large free of exogenous insulin, they may require it when blood glucose levels are not all around controlled with eating routine alone or with oral hyperglycemia medications. Moreover, individuals with T2DM are frequently joined by inconveniences, for example, cardiovascular ailments, diabetic neuropathy, nephropathy, and retinopathy. Diabetes and its related inconveniences bring down the nature of individuals' lives and produce huge financial and social burdens.

Electronic health records (EHRs) are transforming public health surveillance. Systems that can automatically extract, analyse, organize, and communicate EHR data to public health agencies increase the breadth, clinical detail, and timeliness of public health surveillance. EHR-based systems have rich potential to improve public health surveillance for diabetes, but little work has been done thus far to characterize the accuracy of raw electronic data for diabetes surveillance or to create custom algorithms to accurately distinguish between type 1 versus type 2 diabetes.
Accurate discrimination between type 1 and type 2 diabetes is critical given the different pathophysiology, epidemiology, prevention, management, and prognosis of these two diseases (2–7). Traditional public health surveillance systems either do not distinguish between these two conditions at all or rely on self-reports to make the distinction. In addition, most existing public health surveys are too small to meaningfully track low-prevalence conditions such as type 1 in general and type 2 in youth.

We hypothesized that EHR data could substantially enrich diabetes surveillance by facilitating continuous evaluation of very large populations and leveraging clinical data to distinguish between type 1 and type 2 diabetes. We therefore sought to develop EHR-based surveillance algorithms to detect and classify type 1 versus type 2 diabetes and then apply them to a live, prospective, EHR-based surveillance system to test their performance.

II. Related work

Patient Electronic Health Records (EHRs) is one of the major carriers for conducting data driven healthcare research. There are various challenges if we work directly with EHRs, such as sparsity, noisiness, heterogeneity, bias, etc. One important aspect for mining EHR is how to explore the temporal relationships among different medical events within patient EHRs. Many approaches have been proposed for temporal mining of EHRs.

Electronic phenotyping refers to the process of identifying phenotypes from patient EHRs, which is the procedure of extracting clinically relevant features. There are quite a few existing electronic phenotyping works. For example, Ho*et al*. [3, 4] formulates the patient as tensors, wherein every mode represents a specific type of medical event. The entries in the tensor encode the interaction of those features (e.g., the frequency of a medication and a primary diagnosis). Then they proposed a tensor factorization based approach for identification of the phenotypes. Zhou *et al*. [14] formulates EHRs as temporal matrices with medical events as one dimension and time as the other dimension. They propose an optimization based technology for discovering the phenotypes within which the raw medical features have similar time evolving patterns. Lasko *et al*. proposed a deep learning method for obtaining phenotypes from continuous lab value signals, where they first adopted Gaussian process regression to impute the missing lab test values.

Kale *et al*. [7] applied deep learning to discover the physiomes from the physiological streams obtained in Pediatric Intensive Care Unit (PICU). For all these works, they either define a phenotype as some evolving pattern on the values of a specific medical feature (e.g., lab test or physiological stream), or a group of medical features (e.g., diagnosis, medication or both). They did not consider the temporal relationships across different medical events, which could be crucial as they suggest important information on the impending disease conditions.

Knowledge representation from temporal data is a hot research topic in both data analytics and biomedical informatics. For continuous time data, one popular approach is to transform them into discrete symbolic representations (string, nominal, categorical, and item sets). Popular approaches include Piecewise Linear Approximation (PLA), Adaptive Piecewise Constant Approximation (APCA), Symbolic Aggregate approximation (SAX), Piecewise Aggregate Approximation (PAA), etc. One can refer to for a survey on these approaches. For discrete time data, Mörchen *et al*. proposed the Time Series Knowledge Representation (TSKR) as a pattern language (grammar) for temporal knowledge discovery from multivariate time series and symbolic interval data, where the temporal knowledge representation is in the form of symbolic languages and grammars that have been formulated as a means to perform intelligent reasoning and inference from time dependent event sequences. More recently, Wang *et al*. proposed a convolutional framework to extract temporal signatures in discrete time data using the Temporal Event Matrix Representation (TEMR), which is shown to have wide applicability to a variety of data and application domains that involve largescale longitudinal data.

The temporal graph we propose in this paper provides an alternative way to represent the temporal knowledges contained in discrete time data. The temporal graphs capture temporal structures hidden in the sequences in a more compact way, where the nodes in the graph are events appeared in the EHR and the directed edges encode the temporal relationships between pairwise events. In this representation, the events missing in patient EHRs will not appear in the graph, and the repeated pairwise events with the same ordering will only appear once in the graph. With this representation, the temporal graph is robust and resistant to sparse, noisy, and irregular observations.
Moreover, this representation is very intuitive to interpret the temporal relationships among different medical events in patient EHRs. Another advantage is that with graph based representation, the detected phenotypes (or patterns) will also be in the form of graphs, which can be viewed as a nature aggregation of sequential patterns. In this way, we can effectively alleviate the pattern explosion problem.

There is a large number of information visualization techniques which have been developed over the last decade to support the exploration of large data sets. In visual data exploration, the user is directly involved in the data mining process. Daniel A. Keim propose a classification of information visualization and visual data mining techniques which is based on the data type to be visualized and the technique of visualization , interaction and distortion.

Existing system uses a novel Temporal Event Matrix Representation (TEMR) and learning framework to perform temporal signature mining for large-scale longitudinal and heterogeneous event data. TEMR framework represents the event data in the form of matrix, in which where one dimension corresponds to the type of the events and the other dimension represents the time information. If event _i'happened at time _j'with value _m', then the (i, j)th element of the matrix is _m'. This is a very flexible and intuitive framework for encoding the temporal knowledge information contained in the event sequences.

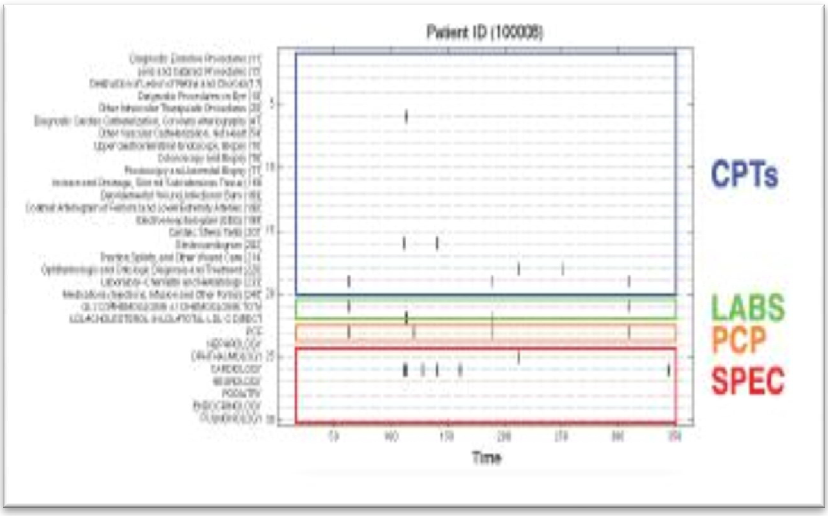

Fig 1. An example of a diabetic patient's electronic record over one year [14]

Figure.1 illustrates a simple example on representing the longitudinal medical record of a diabetic patient over one year using TEMR approach, in which the vertical axis corresponds to the different events such as primary care procedures, physician visits, lab tests, and specialist visits and the horizontal axis represents the time information associated with these events. A dot is used in the matrix for corresponding event happened at the corresponding time. There is analogy between matrix and image. TEMR gives a flexible and intuitive way of encoding comprehensive temporal knowledge. It contains event ordering, duration, and heterogeneity. Authors developed a matrix approximation-based technology to detect the hidden signatures from the event sequences and developed an online updating technology. This enables the representation, extraction, and mining of high order latent event structure and relationships within single and multiple event sequences. The knowledge representation maps the heterogeneous event sequences to a geometric image by encoding events as a structured spatial-temporal shape process.

III. PROPOSED ALGORITHM

The main objective proposed system is to provide interactive and user friendly representation of Knowledge and data with Visual Data Analytics. Visual data mining techniques are tightly integrated with the systems used to manage the vast amounts of relational and semi structured information, including database management and data warehouse systems. The final goal is to bring the power of visualization technology to every desktop to allow accurate, faster, and more intuitive exploration of very large data resources.

Visual Analytics often comprises the usage of multiple views, which requires a well-designed and intuitive user interface taking into consideration the display and arrangement of the visualization and allow the user to interactively parameterize views. Visual analytics is used for the analysis of vast amounts of data to identify and visually distill the most valuable and relevant information content. In this system, automated analysis techniques are combined with the interactive visualizations. Due to this, it becomes very easy to understand and to make decision from given very large and complex data sets. Visual analytics creates tools and techniques which make possible for people to combine information and derive meaningful results from large, changing, indeterminate, and often inconsistent data, find out the expected and unexpected things. It also provides defensible, timely, and easily understandable assessments and effectively communicates assessment for action.

A. VISUAL DATA EXPLORATION

Data Exploration means finding the valuable information hidden in like a drop in the ocean when dealing with data sets containing millions of data items. The aim of Visual data exploration is to integrate human in finding information from large, applying perceptual abilities to the large data sets available in computer systems. Visual data exploration presents the data in some visual form, and allows the human to get insight into the data, draw the conclusions, and allow interacting directly with the data. Visual data mining techniques have proven to be of high value in exploratory data analysis and they also have a high potential for exploring large databases. Visual data exploration can easily deal with highly inhomogeneous as well as noisy data. Visual data exploration is very intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.

B. BACK PROPAGATION ALGORITHM

Most people would consider the Back Propagation network to be the quintessential Neural Net. Actually, Back Propagation 1,2,3 is the training or learning algorithm rather than the network itself. The network used is generally of the simple type. These are called Feed-Forward Networks or occasionally Multi-Layer Perceptrons (MLPs).

- First apply the inputs to the network and work out the output remembers this initial output could be anything, as the initial weights were random numbers.
- Next work out the error for neuron B. The error is What you want What you actually get
- Change the weight. Let W AB be the new (trained) weight and WAB be the initial weight.

Calculate the Errors for the hidden layer neurons. Unlike the output layer we can’t calculate these directly (because we dont have a Target), so we Back Propagate them from the output layer (hence the name of the algorithm). This is done by taking the Errors from the output neurons and running them back through the weights to get the hidden layer errors. For example if neuron A is connected as shown to B and C then we take the errors from B and C to generate an error for A. Having obtained the Error for the hidden layer neurons now proceed as in stage 3 to change the hidden layer weights. By repeating this method we can train a network of any number of layers.

The back propagation algorithm falls into the general category of gradient descent algorithms, which intend to find the minima/maxima of a function by iteratively moving in the direction of the negative of the slope of the function to be minimized/maximized. The main objective is to minimize the error function. The average error function to be minimized (error density).

The main steps using the Back propagation algorithm as follows:

Step 1: Feed the normalized input data sample, compute the corresponding output;
Step 2: Compute the error between the output(s) and the actual target(s);
Step 3: The connection weights and membership functions are adjusted;
Step 4: IF Error > Tolerance THEN go to Step 1 ELSE stop.

C. C4.5 ALGORITHM

The algorithm constructs a decision tree starting from a training set T S, which is a set of cases, or tuples in the database terminology. Each case species values for a collection of attributes and for a class. Each attribute may have either discrete or continuous values. Moreover, the special value unknown is allowed, to denote unspecified values. The class may have only discrete values. We denote with C1 To CN Class the values of the class.

The C4.5 algorithm constructs the decision tree with a divide and conquers strategy. In C4.5, each node in a tree is associated with a set of cases. Also, cases are assigned weights to take into account unknown attribute values. At the beginning, only the root is present, with associated the whole training set T S and with all case weights equal to 1:0. At each node the following di- vide and conquer algorithm is executed, trying to exploit the locally best choice, with no backtracking allowed.

In doing classification with C4.5, the concepts of entropy and correlation coefficient need to be explained in brief. Entropy is a measure of uncertainty among random variables in a collection of data or in other words entropy provides information about the behaviour of random processes used in data analysis. Correlation coefficient has its uses as a chief statistical tool in data analysis finding the relationship between variable sets. Different ways of calculations have been introduced to boost the efficiency of the correlation coefficient among which are Kendall, Pearson’s and Spearman’s correlation coefficients. There are several test options with WEKA providing data classification such as training set, supplied test set, percentage split and cross validation.

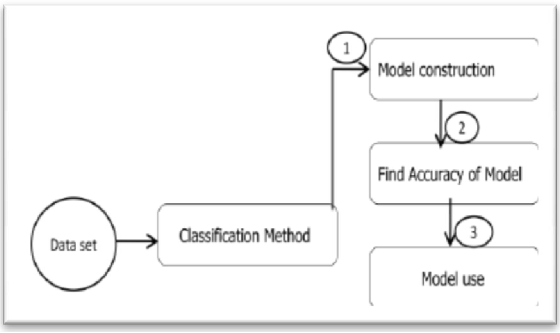

Fig: Classification Steps

IV. CONCLUSION

Integration of visualization techniques and more established methods combines fast automatic data mining algorithms with the intuitive power of the human mind, which improve the quality and speed of the data mining process. Visual data mining techniques are used to manage the vast amounts of relational and semi structured information, including database management and data warehouse systems. From the above discussion and paper reviewed, it can be possible to present a novel temporal event matrix representation and learning framework. It can be possible to extend the existing system with the development of an interactive tool for visualization. The ultimate goal is to bring the power of visualization technology to every desktop to allow accurate, better, faster and intuitive exploration of very large data resources.

REFERENCES

[1] David Gotz, Fei Wang, Adam Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. Journal of biomedical informatics. 2014;48:148–159. [PubMed] [2] David Gotz, Fei Wang, Adam Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. Journal of biomedical informatics. 2014;48:148–159. [PubMed]

[2] 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.*, vol. 30, no. 6, pp. e201–e207, Jun. 2013.

[3] 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," *J. Epidemiol.*, vol. 22, no. 4,pp. 317–323, May 2012.

[4] Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, ShahramEbadollahi, And Andrew F. Laine,‖A Framework For Mining Signatures From Event Sequences And Its Applications In Healthcare Data‖, *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 35, No. 2, February 2013.

[5]Healthcare Information and Management Systems Society.

[6]B. Cao, D. Shen, J.T. Sun, X. Wang, Q. Yang, and Z. Chen, —Detect and Track Latent Factors with Online Nonnegative Matrix Factorization,‖ *Proc. 20th Int'l Joint Conf. Artificial Intelligence*, pp. 2689-2694, 2007.

[7] F.R.K. Chung, *Spectral Graph Theory*. *Am. Math. Soc*., 1997.

[8] C. Ding, T. Li, and M.I. Jordan, —Convex and Semi-Nonnegative Matrix Factorizations,‖ *IEEE Trans. PatternAnalysis and Machine Intelligence*, vol. 32, no. 1, pp. 45-55, Jan. 2010.

[9] M. Dong, —A Tutorial on Nonlinear Time-Series Data Mining in Engineering Asset Health and Reliability rediction: Concepts, Models, and Algorithms,‖*Math. Problems in Eng*., vol. 2010, pp. 1- 23, 2010.

[10] J. Eggert and E. Korner, —Sparse Coding and NMF,‖ *Proc. IEEE Int'l Joint Conf. Neural Networks*, vol. 2, pp. 2529-2533, 2004.

[11]W. Fei, L. Ping, and K. Christian, —Online Nonnegative Matrix Factorization for Document Clustering,‖ *Proc. 11thSIAM Int'l Conf. Data Mining*, 2011.

[12] B. Cao, D. Shen, J.T. Sun, X. Wang, Q. Yang, and Z. Chen, —Detect and Track Latent Factors with Online Nonnegative Matrix Factorization,‖ *Proc. 20th Int'l Joint Conf. Artificial Intelligence*, pp. 2689-2694, 2007.

[13] F.R.K. Chung, *Spectral Graph Theory*. *Am. Math. Soc*., 1997.

[14] C. Ding, T. Li, and M.I. Jordan, —Convex and Semi-Nonnegative Matrix Factorizations,‖ *IEEE Trans. PatternAnalysis and Machine Intelligence*, vol. 32, no. 1, pp. 45-55, Jan. 2010.

[15] M. Dong, —A Tutorial on Nonlinear Time-Series Data Mining in Engineering Asset Health and Reliability rediction: Concepts, Models, and Algorithms,‖*Math. Problems in Eng*., vol. 2010, pp. 1- 23, 2010.

[14] S.A.Sarwade and R.K.Makhijani----"A Review on Mining Signatures from Event Sequences and Visual Interactive Knowledge Discovery in Large Electronic Health Record Database" Volume 3, Issue 12, December 2013