ISSN 22773061

A Review on associative classification for Diabetic Datasets A Simulation Approach

Deepti Jain [1] Department of Computer Science & Engineering BUIT, Bhopal, India
deeptij85@gmail.com
Divakar singh[2] Department of Computer Science & Engineering BUIT, Bhopal, India
divakar_singh@rediffmail.com

ABSTRACT

Association rules are used to discover all the interesting relationship in a potentially large database. Association rule mining is used to discover a small set of rules over the database to form more accurate evaluation. They capture all possible rules that explain the presence of some attributes in relation to the presence of other attributes. This review paper aims to study and observe a flexible way, of, mining frequent patterns by extending the idea of the Associative Classification method. For better performance, the Neural Network Association Classification system is also analyzed here to be one of the approaches for building accurate and efficient classifiers. In this review paper, the Neural Network Association Classification system is studied and compared in order to find best possible accurate results. Association rule mining and classification rule mining can be integrated to form a framework called as Associative Classification and these rules are referred as Class Association Rules. This review research paper also analyzes how data mining techniques are used for predicting different types of diseases. This paper reviewed the research papers which mainly concentrated on predicting Diabetes.

KEYWORDS

Data Mining, Association Rule, Back propagation neural network, Class association rules.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 7, No 1
editor@cirworld.com
www.cirworld.com, member.cirworld.com

INTRODUCTION

1.1 Introduction

Data mining refers to the process of extracting knowledge from large amounts of data. The mining process is an iterative sequence of steps**.** Normally there exists two categories of data mining.

(i) Descriptive data mining.

(ii) Predictive data mining.

For carrying out summarizations or generalizations descriptive data mining is used whereas for finding out the inference or predictions, Predictive data mining is used. Association rule mining falls under the descriptive category**.**

Association rules aims in extracting important correlation among the data items in the databases. A predictive approach tries to assign a value to a future or unknown value of other variables or database fields, whereas description tries to summarize the information in the database and to extract patterns. Association rule mining is a descriptive data mining task.

Association rules find interesting associations and/or relationships among large set of data items. They capture all possible rules that explain the presence of some attributes in relation to the presence of other attributes. An association rule can be expressed in the form of $X \rightarrow Y$, whereby X and Y are disjointed item sets. An association rule must satisfy two important measures, namely support and confidence. Support indicates how frequently the items in the rule occur together whereas confidence determines how frequent items in Y appear in transactions containing X. It is important to analyze the relationships among the risk factors and Apriori algorithm is used for this purpose. It is an powerful algorithm for mining frequent item sets for association rules.

The association rule mining and classification rule mining can be integrated to form a framework called as Associative Classification and these rules are referred as Class Association Rules. A class association rule is obviously a predictive task. By using the discriminative power of the Class Association Rules we can also build a classifier.

1.2 Diabetes

Insulin is one of the most important hormones in the body. It aids the body in converting sugar, starches and other food items into the energy needed for daily life. However, if the body does not produce or properly use insulin, the redundant amount of sugar will be driven out by urination. This disease is referred to diabetes. The cause of diabetes is a mystery, although obesity and lack of exercise appear to possibly play significant roles**.**

RESEARCH SURVEY

2.1 Survey Report on Diabetic's

Diabetes is a substantial medical issue in the United States. According to the National Center for Health Statistics (NCHS), the number of non-institutionalized adults with diagnosed diabetes was 13.4 million in 2002. Diabetes caused numerous deaths, 71,372 deaths in 2001, ranking it the 6th highest cause of death in the United States. The number of deaths can be minimized by motivating patients to self-manage their diabetes. In fact "Diabetes Self-Management Education (DSME) is the cornerstone of care for all individuals with diabetes who want to achieve successful health-related outcomes. The project research team is made up of professors and graduate students with backgrounds in information technology, nursing, and sociology who have an interest in using emerging technologies in healthcare applications.

The researchers theorized that allowing diabetes patients to easily enter information about their daily activities (e.g. diet, exercise, blood sugar levels, and medications) into a handheld computer, and then immediately upload that information to a central server for review by a medical professional would prove more useful than simply logging this information for the patient's own use. Several PDA applications for diabetes patients exist (Accu-Check, Diabetes Pilot, SiDiary, Gluco Control, et.al.) but none had all the features the researchers felt such a system needed. There was no standard mobile self-management diabetic software system that was all encompassing in the areas of insulin medications, noninsulin medications, diet, exercise, blood sugar and weight entries. Furthermore, none had the ability to quickly send patient data to a central server so that a healthcare professional could monitor patients'

2.2 Methodologies used in Diabetes Detection

A various data mining technique is used in diabetes detection and prediction such as cluster analysis, association rules, Bayesian network and classifier support vector machine, Regression analysis, rough set, Text mining etc.

Yuji Akematsu et al[9] used regression analysis to estimate the effect of e-health to user who have these disease and then calculate the monetary effect of eHealth in reduction of medical expenditure. The objective of this research is to evaluate empirically the effectiveness of eHealth in Nishi-aizu Town, in Fukushima Prefecture, based on a mail survey to residents and their receipt data of National Health Insurance from November 2006 to February 2007. This paper mainly analyzes for what kind of diseases eHealth in this town is effective to reduce medical expenditures. Our main interests are focused on four life style related diseases, namely, heart diseases, high blood pressure, diabetes, and strokes..

As a result, eHealth is verified to provide the positive effect to heart diseases, high blood pressure, and diabetes among four diseases. These results are expected to strongly valid for establishment of evidence-based policy such as reimbursement from the medical insurance which we do not have yet in Japan.

Huang and colleagues et al[20] applied Naïve Bayes, IB1classifier, and CART C4.5 on information collected from 2,064 patients (1,148 males and 916 females), and identified the five most important factors that influence blood glucose control: (1) age; (2) diagnosis duration; (3) need for insulin treatment; (4) random blood glucose measurements; and (5) diet treatment. Using these five factors, 95% predictive accuracy and 98% sensitivity was achieved.

Supriya charoensiriwath et al[10] presented 3D body scanning technology for personalized health monitoring and diagnosis system. By having such a system which lets people view their body shape and health online would encourage them to greater care of themselves and their health. In addition this system provides an easy way for people to keep in touch with doctors and nutritionists and for doctor to monitor their patients and screen for diabetes at earlier stage

Parisut jitapakdee et al[3] proposed image processing technique for detecting the symptoms of Diabetic retinopathy (DR). Diabetic Retinopathy is a medical condition where the retina is damaged because fluid leaks from blood vessels into the retina. The presence of hemorrhages in the retina is the earliest symptom of diabetic retinopathy. The number and shape of hemorrhages is used to indicate the severity of the disease. Early automated hemorrhage detection can help reduce the incidence of blindness.

Amit Kumar Mishra et al [14] developed Information Geometry Based Scheme for Hard Exudate Detection in Fundus Images. They focuses on the detection of hard exudates (HEs) in fundus Images for identification of the condition of Diabetic Retinopathy (DR). HEs have been found to be the most specific markers for the presence of retinal oedema, and are also one of the most prevalent lesions during early stages of DR. In order to detect HEs, we have introduced a novel way of using the Distance Learning Metric using a Non-linear Kernels function. We have also introduced a new method to remove the OD in the post-processing stage based on variance calculation.

Tammy Toscos et al[7] Kay Connelly et al presents survey of how parents and teens cope with diabetes in the context of technology support. Teenagers make many transitions during adolescence toward adult lifestyles and responsibilities. Teens with Type 1 Diabetes (T1D) have the additional burden of assuming responsibility for disease management. The findings reported represent the perspective of parents and adolescents who are coping with T1D - uncovering various tensions that interfere with the effective use of technology to manage the disease. Predominant themes from a set of semi-structured interviews are used to construct implications for the design of new technology intended to support families coping with T1D. By drawing attention to tensions (emotional strain or pressure) that prevents the effective use of diabetes-related technology, we have uncovered ways in which technology might to facilitate the transition to adulthood. Emphasizing the needs of adolescents does not have to be in conflict with the needs of their parents. Our findings suggest that addressing parents' desire to improve diabetes-related communication will help in diffusing the technology-enabled conflict.

Chunquan Huang et al[11] gave a detailed research on Evaluation of Diabetes Metabolic Function Based on Support Vector Machine aims to provide a survey using Support Vector Machine (SVM) to predict and assess metabolic functions of diabetes based on bio-heat transfer theory and infrared thermal imaging technology. Two metabolic characteristic values, metabolic function parameter and blood perfusion rate, are extracted from thermography data of cold water stimulation experiment as inputs of SVM to set up models by different kernel functions. For more than 2000 clinical data used in the paper, the prediction accuracy averaged 90%. The research provides a new attempt to evaluate diabetes metabolic function, hoping for contribution to early detection of diabetes.In addition, control models can be selected to illuminate the significance of SVM applied to chronic metabolic diseases, such as Artificial Neural Networks (ANN).

Yung-Hsiu Lin et al, [19] presents Developing Diabetes Self-Care Supporting Service: A Systemic Approach A method of systemic approach is used in this study as the map for building a diabetes self-care support (DSCS) service. Based on the self-care theory, DSCS service is designed as an integrated care system that functions at both at patient's home and remote service center. The service was evaluated by a group of diabetic patients for one month and a service acceptance survey was administrated. In this study, the method of systemic approach is useful in analyzing the complex service system and the results from the survey support the needed self-care service The results from the evaluation of DSCS service indicate that the Preference and Ease of Use of the self-care functions are significantly correlated with the Intention to Use of this specific service.

Yu Ting Yeh et al[18] presents Assessment of User Satisfaction with an Internet based Integrated Patient Education System for Diabetes Management. The internet based integrated patient education system supplies diabetic patients with thorough education and treatment needs. On one hand it fulfills the needs of diabetic patients or their family member by assessing online for medical records, prescriptions, laboratory results and amount of pharmaceutical education received.it reinforces patient to accept nursing education and pharmaceutical education online. Through the assistance of online diabetes patients, self-management and education materials, not only the quality of treatment will be improved but care as well. Patients understanding towards drug use safety and self-care knowledge and skills will also be greatly advanced with the help of online system.

Ping Zuo et al [21]described Analysis of Noninvasive Measurement of Human Blood Glucose with ANN-NIR Spectroscopy. Based on the task of noninvasive blood glucose measure, An outside body and inside body measure amphibious human blood glucose measuring system is designed by us in china which is used to measure and model analyze glucose liquid of different concentration, and in this paper we put forward a new technology in analyzing the absorption spectrum of infrared ray by blood. After the measurement of the absorption spectrum of infrared by the total blood and normal human serum and higher glucose blood, this paper makes artificial neural network training through the Levenberg- Marquardt BP neural network which depends on the character parameter in the value of the 16 special wavelengths. The research result has valuable promise to blood analysis of the infrared spectroscopy and diagnosing the

disease. The experiment proves that this method, compared with the current method, is characterized by rapid training speed and accurate predication result.

Breault et al[23] applied a classification and regression tree (CART) using the CART data-mining software (Salford Systems, San Diego, CA) on data of 15,902 diabetes patients and observed that the most important variable associated with bad glycemic control (HbA1c >9.5) is age. Patients below the threshold of 65.6 years old have worse glycemic control than older people, which was very surprising to the clinicians. Using this knowledge, they can target the specific age groups that are more likely to have poor glycemic control. In order to predict next-morning fasting blood glucose (FBG), Yamaguchi and colleagues20 used dataForest software (Yamatake Co., Japan)21 to apply classification on a data set collected from four type 1 diabetes mellitus (T1DM) patients over a period of 150 days. The authors constructed a model for predicting next-morning FBG based on FBG, metabolic rate, food intake, and physical condition and concluded that the physical condition is highly correlated with FBG and its best variable to predict FBG.

Bellazzi et al[17] used a combination of structural time series (STS) analysis, based on Bayesian network, and temporal abstraction (TA) to interpret past BGL data in order to extract and visualize the trends and daily cycles of BGL. First, data was analyzed with STS, with the results in the form of time varying series over a specific time period. Then, the second step was to apply TA on the results from the first step for further interpretation. At the end of the process, the final results were a trend diagram and a daily cycle diagram that visually represent the BGL.

2.3 Table 1. Comparing various Technologies used for Diabetes Dataset

S.no	Author	Methodology Used	Accuracy	Prediction level
1	Yuji Akematsu	Regression analysis	73% approx.	Moderate
2	Huang and colleagues	Naïve Bayes, IB1classifier	95% approx.	Good
3	Supriya charoensiriwath	3Dbody scanning technology	45% approx.	Good
4	Parisut jitapakdee	Image processing	NA	NA
5	Amit Kumar Mishra	Information Geometry	96% avg.	High
6	Chunquan Huang	SVM	90% avg.	High
7	Yung-Hsiu	DSCS	Less than 35%	Poor
8	Ping Zuo	ANN-NIR	83.5% approx.	High
9	Breault and colleagues	classification and regression tree	76% avg.	High
10	Bellazzi	STS Analysis	80% avg.	High

NEURAL NETWORK ASSOCIATIVE ASSOCIATION SYSTEM

The proposed system, Neural[8] network Associative Association, consists of three parts: Quantization[10] Phase, Generating Class Association Rule Phase and the Association phase. Before Quantization phase, we have to convert the character data into numeric data. This is important as neural network can only tackle the numeric data. Preprocessing phase. Following diagram represent proposed model

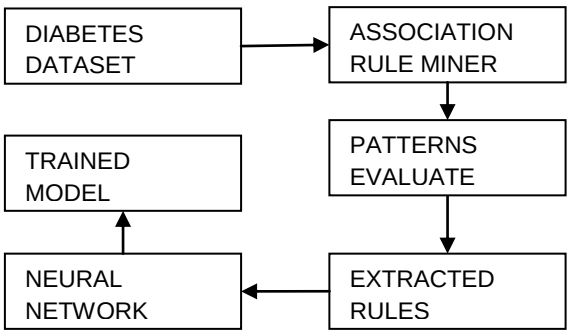

FIGURE1: Associative Classification for Diabetic Dataset

In figure we are giving the description of associative classification of diabetic dataset using Neural network.Back propagation Neural Network is a feed forward neural network which is comprised of a paradigm of processing units, oriented in a series of two or more mutually exclusive sets of neurons or layers. The first layer is an input layer which provides the holding site to the inputs of the neural network. The generic role of this layer is to hold the input values and to supply these values to the units in the next layer. The last, output layer is the point at which complete mapping of the network input is provided. In these two layers lies zero or more layer of hidden units. In this internal layer, additional remapping takes place.

3.1. Quantization Phase

Association datasets[11] may contain many continuous properties. Association rules mining with continuous properties is a research content. So our system involves quantizing the continuous attributes based on the pre-defined class target. For quantization, we have used the algorithm form.

3.2 Generating Class Association Rules Phase

This move presents a way for generating the Class Association Rules (CARs) from the pre-defined dataset.

Association Rules

Consider:

D= {d1, d2... dn} is a database that consist of set of n data and dI

I= {i1, i2… im} is a set of all items that appear in D

The association rule has a format, A→ B, by support=s%, confidence=c% and A ∩ B = ø and where A, B I.

Support value is frequency of number of data that consist of A and B or P (A U B).

CONCLUSION

Data mining has played a important role in diabetic research. Data mining would be a valuable asset for a diabetes researcher because it can unearthen hidden knowledge from a huge amount diabetes related data. We believe that data mining can significantly help diabetes research and ultimately improve the quality of health care of diabetes patients. The application of data mining techniques in the selected articles were useful for extracting valuable knowledge and generating new hypothesis for further scientific research/experimentation and improving health care for diabetes patient. The result could be used for scientific research and real life practice to improve the quality of health care for diabetes patient.

FUTURE WORK

In future we are using Feed forward Neural network based approach for diabetic risk estimation.

REFERENCES

[1]. P. N. Tan, M. Steinbach and V. Kumar, "Association analysis: Basic concepts and algorithms", in "Introduction to Data Mining", Addison Wesley,2006,Ch.6,www.users.cs.umn.edu/ ~kumar/dmbook/ch6.pdf (Accessed: August 2012).

[2]. Xiaoxin Yin,Jiawei Han, “CPAR: Classification based on Predictive Association Rules”, In Proc. Of SDM, pp.no. 331-335,2012.

[3] Parisut Jitpakdee, A Survey on Hemorrhage Detection in Diabetic Retinopathy Retinal Images, 978-1-4673-2025-2/12/$31.00 ©2012 IEEE.

[4]. Mary DeRosa, “Data Mining and data analysis for counterterrorism”, Center for Strategic and International Studies, March 2011.

[5]. Prachitee B. Shekhawat, Sheetal S. Dhande, “A classification technique using associative classification”, International journal of computer application(0975-8887) vol. 20-No.5,pp.no. 20-28,April 2011.

[6]. Xindong wu,Vipin Kumar,et.al. “Top 10 algorithms in data mining”, Knowledge Information System(2010) 14:1-37 DOI 10.1007/s10115-007-0114-2,Springer-Verlag London Limited 2011.

[7]Tammy Toscos, Kay Connelly, A survey of how parents and teens cope with diabetes in the context of technology support, 978-1-936968-15-2 © 2011 ICST.

[8]. Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, 2e Elsevier Publication, 2011.

[9] Yuji Akematsu, Empirical Analysis of the Effect of eHealth to Medical Expendirures of Lifestyle-related Diseases, 978-1- 4244-6376-31101$26.00 ©2010 IEEE.

[**10**]Supiya Charoensiriwath, SizeThailand e-Health: A Personalised Health Monitoring and Diagnosis System Using 3D Body Scanning Technology, 978-1-890843-21-0/10/$26.00 ©2010 IEEE.

[11] Chunquan Huang, The Research on Evaluation of Diabetes Metabolic Function Based on Support Vector Machine, 978-1-4244-6498-2/10/$26.00 ©2010 IEEE.

[12]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, “Advances in Knowledge Discovery and Data Mining”, MIT Press, Cambridge, Massachusetts, USA,2010.

[13].B.Liu, W. Hsu and Y.Ma, “Integrating Classification and Association rule mining”, KDD, pp. no.80-86,2010.

[14] Amit Kumar Mishra, An Information Geometry Based Scheme for Hard Exudate Detection in Fundus Images(AISTATS), 2009.

[15]. Chung, Kusiak, “Grouping parts with a neural network”, Journal of Manufacturing System, volume 13, Issue 4, 02786125, pp.no. 262, April 2009.

[16].Cheng Jung Tsai, Chein-I Lee, Wei-Pang Yang, “A discretization algorithm based on Class Attribute Contingency Coefficient”, Inf. Sci., 178(3),pp. no. 714-731,2009.

[17].Bellazzi R, Abu-Hanna A. Data mining technologies for blood glucose and diabetes management. J Diabetes Sci Technol. 2009;3(3):603–12.

[18]Yu Ting Yeh, Assessment of User Satisfaction with an Internet based Integrated Patient Education System for Diabetes Management, 978-1-4244-2281-4/08/$25.00_c 2008 IEEE

[19]. Yung-Hsiu Lin, Developing Diabetes Self-Care Supporting Service: A Systemic Approach, pp. 71-77, 2007.

[20].Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients’ data. Artif Intell Med. 2007;41(3):251–62.

[21].Ping Zuo, Yingchun Analysis of Noninvasive Measurement of Human Blood Glucose with ANN-NIR Spectroscopy, 0-7803-9422-4/05/$20.00 ©2005 IEEE.

[22].M. J. Zaki, "Mining non-redundant association rules", Data Mining Knowl. Disc., 9, 223-248,2004

[23].Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artif Intell Med. 2002;26(1–2):37–54.

[24].R.Agrawal, T.Imieilinski, and A.Swami, “Mining Association Rules between sets of items in large databases”,SIGMOD, pp. 207-216,2000.

[25].Robert J. Schalkoff, “Artificial Neural networks”, McGraw Hill publication, ISE.

[26].http://archive.ics.uci.edu/ml/datasets.html.

Ms. Deepti Jain[1], Assistant Professor in Computer Science & Engineering Department at Trinity Institute of Technology & Research Bhopal, MP, India, has done her bachelors in Computer Science from Truba IE&IT Bhopal, in 2008. Her area of interest includes Data Mining, Image Processing, and Neural Networks etc.

Dr. Divakar singh[2], Professor & HOD of Computer Science & Engineering Deptt., BUIT Barkatullah University, Bhopal received his degree B.E., M.TECH & PHD in computer science engineering. He has research interests in image analysis & image mining,softcomputing & machine learning & carries a total of 12 years teaching experience in the same subjects