IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 6, August 2014.
www.ijiset.com
ISSN 2348 – 7968

Predicting Diabetes by cosequencing the various Data Mining Classification Techniques

P. Radha[1], Dr. B. Srinivasan[2]

[1] Ph.D Research Scholar, Computer Science Department, Karpagam University, Coimbatore, Tamil Nadu, India

[1] Assistant Professor, Vellalar College for Women, Erode, Tamil Nadu, India

[2] Computer Science Department, Gobi Arts and Science College, Gobichettipalayam, Tamil Nadu, India

Abstract

Diabetes affects between 2% to 4% of the global population and its avoidance and effective treatment are undoubtedly crucial public issues in the 21st century. Although human decision making is often optimal, it is poor when there are huge amounts of data to be classified. Medical data mining has been a great potential for exploring hidden patterns in the data sets of medical domain. Data mining algorithms can be trained in clinical data to predict the disease. Classification is the generally used technique in medical data mining. This paper presents results comparison of five supervised data mining algorithms using five performance criteria. The performance is evaluated by the five algorithms C4.5, SVM, *k*-NN, PNN, and BLR. Comparison of performance of data mining algorithms based on computing time, precision value, the data evaluated using 10 fold Cross Validation error rate, bootstrap validation and accuracy. A typical confusion matrix is furthermore displayed for quick check. The study describes algorithmic discussion of the dataset for the disease acquired from UCI and ICMR-INDIAB, on line repository of large datasets. Tanagra tool is used to achieve the best results. Tanagra is data mining matching set.

Keywords: ***Diabetes, Data Mining, Classification, C4.5, SVM, k-NN, PNN, BLR***

1. Introduction

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. It is "the science of extracting useful information from large databases". It is one of the tasks in the process of knowledge discovery from the database. [1] Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. It is a process to examine large amounts of data routinely collected. Data mining is most useful in an exploratory analysis because of nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. There are two primary goals of data mining tend to be *prediction* and *description. Prediction* involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand *Description* focuses on finding patterns describing the data that can be interpreted by humans. The Disease Prediction plays an important role in data mining. There are different types of diseases predicted in data mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes etc… This paper analyzes the Diabetes predictions. Basic understanding on growth and factors affecting diabetes from external sources is required before building predictive models. The idea is to predict the diabetic cases and to find the factors responsible for diabetes using data mining methods. Some of the interesting facts affected by diabetes and observed from the statistics given by various researchers. The development of finding useful patterns or importance in raw data has been called KDD (knowledge discovery in databases). In the previous century, an exponential enhancement has been seen in the accuracy and sensitivity of diagnostic tests, from observe outside symptom and use refined laboratory tests and difficult imaging methods increasingly that allow detailed non-invasive inner examinations. This improved accuracy has predictably resulted in an exponential increase in the patient data available to the physician. The process of finding confirmation to decide a probable reason of patient's key symptoms from all other possible reason of the symptom are known as establishing a medical diagnosis.

Data mining is a remarkable opportunity to support physician deal with this large amount of data. Its methods can help physicians in various ways such as interpret multifaceted diagnostic tests, combining information from several sources (sample movies, images, clinical data, proteomics and scientific knowledge), given that support for differential diagnosis and providing patient-specific prediction.

ISSN 2348 – 7968

The respite of the paper is organized as follows it first gives details of classification on different methods. Then medical data mining is described. The article ends by concluding with a summary of investigated methods and future research.

2 DATA ANALYSIS

Diabetes is the most common disease nowadays in all populations and in all age groups. It is a disease in which the body does not produce or properly use insulin. The cells in our body require glucose for growth for which insulin is quite essential. When someone has diabetes, little or no insulin is secreted. In this situation, plenty of glucose is available in the blood stream but the body is unable to use it (Mohamed et al., 2002). The types of diabetes are Type-1 Diabetes, Type-2 Diabetes, Gestational Diabetes.

Type-1 diabetes occurs when the body's immune system is attacked and the beta cells (these cells produce insulin) of pancreas are destroyed. This results in insulin deficiency. The only treatment to Type-1 diabetes is insulin. **Type-2 diabetes** is caused by relative insulin deficiency. Pancreas in Type-2 diabetes still produces insulin but it may not be effective or may not produce sufficient amount of insulin to control blood glucose Type-2 diabetes is the most common type of diabetes which usually develops at age 40 and older. **Gestational Diabetes** occurs to a pregnant women without a previous diagnosis of diabetes. **Prediabetes** is a condition when patient blood sugar level triggers higher than normal, but not so high that it can validate it as type 2 diabetes.

The most important methodology used for this paper throughout is by the analysis of journals and publications in the field of medicine. The data study consists of diabetes dataset. It includes name of the attribute as well as the explanation of the attributes. Pima Indian Diabetes Dataset and Indian Council of Medical Research–Indian Diabetes (ICMR-INDIAB) study provides data about the Diabetes. World Health Organization (WHO) report had shown a marked increase in the number of diabetics and this trend is expected to grow in the next couple of decades. In the International Diabetes Federation Conference 2003 held in Paris, India was labeled as " Diabetes Capital of the World". Of about 190 million diabetics worldwide, more than 33 million are Indians. According to the diabetes Atlas of 2009, there were 50.8 million people with diabetes in India. The worldwide figure is expected to rise to 330 million, 52 million of them Indians by 2025.

The first effort was made by ICMR–INDIAB to provide accurate and comprehensive state and national level data on prevalence of diabetes in India. The ICMR-INDIAB conducted a study and the results are amazing. The results provide evidence for increase in prevalence of diabetes not only in urban areas but also in rural areas. The study also provides authentic new data on the total number of people with diabetes in India," Dr. Mohan [19] added. The study began in late 2008 and was completed by 2010. It factored in anthropometric parameters like body weight, BMI (body mass index), height and waist circumference, and also tested fasting blood sugar, followed by blood sugar after a glucose load (known diabetics exempted), and cholesterol for all participants. Questions were also asked about food habits, physical activity, and smoking, alcohol usage, among others. Projections made in the past about the total number of diabetics in the country for the future may need to be revised. For instance, in May 2004, in Diabetes Care, volume 27, Sarah Wild et al [21] proposed that India would have 79.4 million people with diabetes in 2030. Nineteen years ahead of that deadline, India has 62.4 million, and a further 77.2 million (potential diabetics) in the pre-diabetes stage. Also, it has shifted to the 25-34 years age group," Dr. Mohan [19] explained. "The epidemic is likely to stabilize in the population at about 20-25 per cent or so".

Finally it was observed that age standardized prevalence's of diabetes and impaired glucose tolerance was 12.1% and 14.0% respectively, with no gender difference. Diabetes and impaired glucose tolerance showed increasing trend with age. Subjects under 40 years of age had a higher prevalence of impaired glucose tolerance than diabetes (12.8% vs 4.6%, < 0.0001).

Diabetes showed a positive and independent association with age, BMI, WHR (Waist-Hip Ratio), family history of diabetes, and sedentary physical activity. Age, BMI and family history of diabetes showed associations with impaired glucose tolerance. This national study shows that the prevalence of diabetes is high in urban India. There is a large pool of subjects with impaired glucose tolerance at a high risk of conversion to diabetes.

Diabetes mellitus is characterized by recurrent or persistent hyperglycemia, and is diagnosed by demonstrating any one of the following as shown in the table (1) which is WHO diabetes diagnostic criteria [5]

Fasting plasma glucose level $\geq$ 7.0 mmol/l (126 mg/dl)

Plasma glucose $\geq$ 11.1 mmol/l (200 mg/dl) two hours after a 75 g oral glucose load as in a glucose tolerance test

Symptoms of hyperglycemia and casual plasma glucose $\geq$ 11.1 mmol/l (200 mg/dl)

Glycated hemoglobin (Hb A1C) $\geq$ 6.5%.

ISSN 2348 – 7968

Condition	2 hour glucose	Fasting glucose	HbA_{1c}
Unit	mmol/l(mg/dl)	mmol/l(mg/dl)	%
Normal	<7.8 (<140)	<6.1 (<110)	<6.0
Impaired fasting glycaemia	<7.8 (<140)	≥ 6.1(≥110) & <7.0(<126)	6.0–6.4
Impaired glucose tolerance	≥7.8 (≥140)	<7.0 (<126)	6.0–6.4
Diabetes mellitus	≥11.1 (≥200)	≥7.0 (≥126)	≥6.5

Table 1 WHO diabetes diagnostic criteria[1]

3 METHODOLOGY

There are various numbers of data mining methods. One approach to categorize different data mining methods is based on their function ability as below [3]:

Regression is a statistical methodology that is often used for numeric prediction.

Association returns affinities of a set of records.

Sequential pattern function searches for frequent sub sequences in a sequence dataset, where a sequence records an ordering of events.

Summarization is to make compact description for a subset of data.

Classification maps a data item into one of the predefined classes.

Clustering identifies a finite set of categories to describe the data.

Dependency modeling describes significant dependencies between variables.

Change and deviation detection is to discover the most significant changes in the data by using previously measured values.

Classification algorithms require that the classes be defined based on data attribute values. Pattern recognition is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes. Data classification is a two-step process.

Step 1: A classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of data-base tuples and their associated class labels. Each tuple is assumed to belong to a predefined class called the class label attribute. Because the class label of each training tuple is pro-vided, this step is also known as **supervised learning.** The first step can also be viewed as the learning of a mapping or function, y = f (X), that can predict the associated class label y of a given tuple X. Typically, this mapping is represented in the form of classification rules, decision trees, or mathematical formulae.

Step 2: The model is used for classification. First, the predictive accuracy of the classifier is estimated. If we were to use the training set to measure the accuracy of the classifier, this estimate would likely be optimistic, because the classifier tends to over fit the data.

3.1 Machine Learning Approaches

Machine learning algorithms can be classified as supervised learning or unsupervised learning. In supervised learning, training examples consist of input/output pair patterns. Learning algorithms aim to predict output values of new examples based on their input values. In unsupervised learning, training examples contain only the input patterns and no explicit target output is associated with each input [13]. The unsupervised learning algorithms need to use the input values to discover meaningful associations or patterns. In supervised machine learning algorithms (*C4.5, SVM, k-NN, PNN,* and BLR).

3.1.1 C4.5

Decision trees are controlling categorization algorithms. Accepted decision tree algorithms consist of C4.5. At the equivalent time as the name imply, this performance recursively separate inspection in branches to build tree for the purpose of improving the calculation accuracy. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. C4.5 generates classifiers expressed as decision trees, but it can also construct classifiers in more comprehensible rule set form.

3.1.2 SVM

Support vector machines (SVM). Support vector machines are a moderately new-fangled type of learning algorithm, originally introduced. Naturally, SVM aim at pointed for the hyper plane that most excellent separates the classes of data. SVMs have confirmed the capability not only to accurately separate entities into correct classes, but also to identify instance whose establish classification is not supported by data. Although SVM are comparatively insensitive define distribution of training examples of each class. SVM can be simply extended to perform numerical calculations. Two such extension, the first is to extend SVM to execute regression analysis, where the goal is to produce a linear function that can fairly accurate that target function. An extra extension is to learn to rank elements rather than producing a classification for individual elements. Ranking can be reduced to comparing pairs of instance and producing a +1

ISSN 2348 – 7968

estimate if the pair is in the correct ranking order in addition to −1 otherwise.

3.1.3 *k*-NN

It is the nearest neighbor algorithm. The *k*-nearest neighbor's algorithm is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms. The algorithm operates on a set of d-dimensional vectors, $D = \{\mathbf{x}_i \mid i = 1. . . N\}$, where $\mathbf{x}_i \in k^d$ denotes the *I* th data point. The algorithm is initialized by selection *k* points in k^d as the initial *k* cluster representatives or "centroids". Techniques for select these primary seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data *k* times [26]. Then the algorithm iterates between two steps till junction:

Step 1: Data Assignment each data point is assign to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Each group representative is relocating to the center (mean) of all data points assign to it. If the data points come with a possibility measure (Weights), then the relocation is to the expectations (weighted mean) of the data partitions.

"Kernelize" *k*-means though margins between clusters are still linear in the embedded high dimensional space, they can become non-linear when projected back to the original space, thus allowing kernel k-means to deal with more complex clusters. Dhillon et al.[24] have shown a close connection between kernel *k*-means and spectral clustering. The *k*-medoid algorithm is similar to *k*-means except that the centroids have to belong to the data set being clustered. Fuzzy c-means is also similar, except that it computes fuzzy member-ship functions for each clusters rather than a hard one.

3.1.4 PNN

Prototype NN classification is an easy to understand and easy to implement classification techniques. Despite its simplicity, it can perform well in many situations. The new prototype p is simply the average vector of p^1 and p^2, or the average vector of weighted p^1 and p^2. The-class of the new prototype is the same as the one of p^1 and p^2. Continue the merging process until the number of incorrect classifications of patterns in T starts to increase.

3.1.5 BLR

BLR (*Binary Logistic Regression)* Predictive analysis in health care primarily to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease and other lifetime illnesses. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care. Logistic regression is a generalization of linear regression. It is used primarily for predicting binary or multi-class dependent variables.

3.2 Evaluation of Computational Results

The accuracy of a learning system needs to be evaluated before it can become useful. Limited availability of data often makes estimating accuracy a difficult task. Choosing a good evaluation methodology is very important for machine learning systems development. There are several popular methods used for such evaluation, including holdout sampling, cross validation, leave-one out, and bootstrap sampling. In the holdout method, data are divided into a training set and a testing set. Usually 2/3 of the data are assigned to the training set and 1/3 to the testing set. After the system is trained by the training set data, the system predicts the output value of each instance in the testing set. These values are then compared with the real output values to determine accuracy. In cross validation, a data set is randomly divided into a number of subsets of roughly equal size. Ten-fold cross validation, in which the data set is divided into 10 subsets, is most commonly used. The system is trained and tested for 10 iterations. During each iteration 9 subsets of data are used as training data and the remaining set is used as testing data. In rotation, each sub-set of data serves as the testing set in exactly each iteration. The accuracy of the system is the average accuracy over the 10 iterations. In the bootstrap method, *n* independent random samples are taken from the original data set of size *n*. Because the samples are taken with replacement, the number of unique instances will be less than *n*. These samples are then used as the training set for the learning system, and the remaining data that have not been sampled are used to test the system.

4 RESEARCH FINDINGS

4.1 Data mining in Diabetes Disease Prediction

Five different supervised classification algorithms i.e. C4.5, SVM, K-NN, PNN, and BLR have been used to analyze dataset. Tanagra tool is powerful system that contains clustering, supervised learning, Meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms.

4.2 Tanagra

ISSN 2348 – 7968

Tanagra is a data mining suite build around graphical user interface. Tanagra is particularly strong in statistics, offering a wide range of uni and multivariate parametric and nonparametric tests. Equally impressive is its list of feature selection techniques. Together with a compilation of standard machine learning techniques, it also includes correspondence analysis, principal component analysis, and the partial least squares methods. Tanagra is more powerful, it contains some supervised learning but also other paradigms such as clustering, supervised learning, meta supervised learning, feature selection, data visualization supervised learning assessment, statistics, feature selection and construction algorithms. The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in this domain , and allowing to analyze either real or synthetic data. Tanagra can be considered as a pedagogical tool for learning programming techniques. Tanagra is a wide set of data sources, direct access to data warehouses and databases, data cleansing, interactive utilization.

4.3 Data Source

To evaluate these data mining classification Pima Indian Diabetes Dataset was used. The dataset has 9 attributes and 768 instances. Attributes are exacting, all patients now are females at least 21 years old of Pima Indian heritage. If the 2 hour post load Plasma glucose was as a minimum 200 mg/dl (Table 2).

Sno	Name	Description
1.	Pregnancy	Number of times pregnant
2	Plasma	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3.	Pressure	Diastolic blood pressure (mm Hg)
4.	Skin	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	Mass	Body mass index (weight in kg/(height in m)^2)
7	Pedigree	Diabetes pedigree function
8	Age	Age (in years)
9	Class	Class variable (0 or 1)

Table 2. Attributes of diabetes dataset

C4.5, SVM, K-NN, PNN, and BLR in a lowest computing time that we have experimented with a dataset. A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy. Confusion matrix is a matrix representation of the classification results (table 3).

	Classified as Healthy	Classified as not Healthy
Actual Healthy	TP	FN
Actual not Healthy	FP	TN

Table 3: confusion matrix

From the confusion matrix to analyze the performance criterion for the classifiers in disease detection accuracy, precision, recall have been computed for all datasets (Table 3). Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. Recall is the percentage of positive labeled instances that were predicted as positive.
The fitness criteria are calculated as follows:

$$\text{Sensitivity} = TP / (TP + FN)$$

$$\text{Specificity} = TN / (FP+TN)$$

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

The following table 4 consists of values of different classification algorithms. According of these values the lowest computing time (< 550 ms) can be determined.

Algorithm used	Time Taken (ms)	Accuracy %	Positive Recall	Error Rate
C 4.5	550	86	0.38	0.28
SVM	546	74.8	0.368	0.29
K-NN	640	78	0.473	0.34
PNN	546	67	0.481	0.34
BLR	515	75	0.372	0.273

Table 4. Performance study of Algorithm

SVM, PNN, BLR in a lowest computing time that was experimented with the dataset.

5 CONCLUSION

The main goal of medical data mining algorithm is to get best algorithms that describe given data from multiple aspects. There are different data mining classification algorithm that can be used for the identification of diabetes disease among patients. In this paper five classification techniques (C 4.5, SVM, K-NN, PLR, and BLR) are applied to predict the diabetes disease in patients. The algorithms are very necessary for intend an automatic classification tools. In our study first the five techniques were first filtered by using the computing time in which BLR has the lowest computing time with 75% accuracy and error rate of 0.27. The second one with accuracy rate is SVM compared with other techniques. The other techniques accrue high computing time with improved accuracy and error rate is high compared with BLR. Then the accuracy of BLR is 75% from the above results BLR algorithm plays a vital role in data mining techniques.

References

[1]. Elma kolce (cela), Neki Frasheri, “A Literature Review of Data Mining Techniques used in Healthcare Databases”, *ICT Innovations 2012 Web Proceedings* -Poster Session.

ISSN 2348 – 7968

[2]. D.S.Kumar, G.Sathyadevi, S.Sivanesh Decision, "Support System for Medical Diagnosis Using Data Mining ", *International journal of computer applications,* Vol. 4, No. 5, 2011.

[3]. N.Satyanandam, Dr.Ch.Satyanarayana, Md.Riyazuddin, A.Shaik."Data Mining Machine Learning Approaches and Medical Diagnose Systems" A Survey. *International journal of computer applications*, Vol. 2, No. 2, 2009.

[4]. Karthikeyini.V., Pervin begum.I., Tajuddin.K., Shahina Begum, "Comparative of data mining classification algorithm (CDMCA) in Diabetes Disease Prediction", *International journal of Computer Applications*, Vol.60, No. 12, Dec. 2012, pp. 26-31.

[5]. en.wikipedia.org/wiki/ Diabetes_mellitus Wikipedia

[6]. E.Knorr.E and R.Ng, "Algorithms forming distance -based outliers in large datasets", in *proceedings of 1998 International Conference on Very Large Data Bases* (VldB'98), pp. 392-403 New York, 1998.

[7]. E.Jiawei Hen and Micheline Kamber "Data Mining Concepts and Techniques", *CA:Elsevier Inc,SanFranciso*, 2006

[8]. U.M.Piatetsky-Shapiro and G.Smyth "From Data Mining to Knowledge Discovery : An Overview", 1996, pp.1 -36.

[9]. S.C.Liao & M.Embrenchts, "Data Mining techniques applied to medical information", *Med.Inform*, 2000, pp.81-102.

[10]. L.Breiman, J.Friedman, J.Olsen C.Stone, "Classification and Re-gression Trees", *Chapman & Hal*, 1984, 122-134.

[11]. A.Khemphila,V.Boojing, "Comparing Performance of logistic regression, decision tree and neural network for classifying heart disease patients", *Proceeding of International conference on Computer Information System and Industrial Management Application*, 2010, pp.193-198.

[12]. K.Srinivas, B.Kavitha Rani,A.Govrdhan, Applications of Data Mining Techniques in health care and Prediction Heart Attacks, *International Journal on Computer Science and Engineering* (IJCSE), vol. II, 2010, pp.250 -255.

[13]. D.Rubben, Jr.Canals (2009) "Data Mining in Health care :Current Applications and Issues".

[14]. Tanagra Data Mining tutorials http:// data-mining -tutorials-blogspot.com.

[15]. UCI Machine Learning Repository pima Indian diabetes dataset

[16]. Smith, J.,W., Everhart, J.,E., Dickson, W.,C., Knowler, W.,C. and Johannes, R.,S., Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in Proceedings of the Symposium on Computer Applications and Medical Care, *IEEE Computer Society Press*, 1988, pp. 261- 265.

[17]. Huy Nguyen Anh Pham and Evangelos Triantaphyllou "Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization" Department of Computer Science, 298 Coates Hall, Louisiana State University, Baton Rouge, LA 70803.

[18]. Ms.S.Sapna, Dr.A.Tamilarasi "Data mining – Fuzzy Neural Genetic Algorithm in predicting diabetes" Department Of Computer Applications (MCA), K.S.R College of Engineering "BOOM 2K8", *Research Journal on Computer Engineering*, March 2008.

[19]. Mohan V, Shanthirani S, Deepa R, Premalatha G, Sastry NG, Saroja R. Chennai Urban Population Study (CUPS No. 4) Intra urban differences in the prevalence of the metabolic syndrome in southern India, the Chennai Urban Population Study (CUPS No. 4) *Diabet Med.*, Vol. 18(4), 2001, pp. 280–287.

[20]. Anjana RM, Ali MK, Pradeepa R, Deepa M, Datta M, Unnikrishnan R, Rema M, Mohan V. The need for obtaining accu-rate nationwide estimates of diabetes prevalence in India - rationale for a national study on diabetes. *Indian Journal of Medical Research*, Vol. 133(4), 2011, pp. 369–380.

[21]. Sarah Wild et al , Global prevalence of diabetes estimates for the year 2000 and projections for 2030, *Diabetes Care*, Vol. 27, No. 10, Oct. 2004, p. 25-60.

[22]. Anjana R.M., et.al and ICMR–INDIAB Collaborative Study Group. "Prevalence of diabetes and prediabetes (impaired fast-ing glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Re-search-INdia DIABetes (ICMR-INDIAB) study". *Diabetologia*, Vol. 54 (12) , Dec 2011, pp. 3022-3027.

[23]. Karthikeyini.V., Pervin begum.I., "Comparison a performance of data mining algorithms (CPDMA) in prediction of Diabetes Disease", *International journal of Computer Science and Engineer-ing*, Vol.5, No. 03, March 2013, pp. 205-210.

[24]. Dhillon IS, Guan Y, Kulis B., "Kernel k-means: spectral cluster-ing and normalized cuts", *KDD*, 2004, pp. 551–556.