# Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability

K. R. Lakshmi and S.Prem Kumar

**Abstract**— Detection of knowledge patterns in clinical data through data mining. Data mining algorithms can be trained from past examples in clinical data and model the frequent times non-linear relationships between the independent and dependent variables. The consequential model represents formal knowledge, which can often make available a good analytic judgment. Classification is the generally used technique in medical data mining. This paper presents results comparison of ten supervised data mining algorithms using five performance criteria. We evaluate the performance for C4.5, SVM, *k*-NN, PNN, BLR, MLR, PLS-DA, PLS-LDA, *k*-means and Apriori then Comparison a performance of data mining algorithms based on computing time, precision value , the data evaluated using 10 fold Cross Validation error rate, error rate focuses True Positive, True Negative, False Positive and False Negative, 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. The Best results are achieved by using Tanagra tool. Tanagra is data mining matching set. The accuracy is calculate based on addition of true positive and true negative followed by the division of all possibilities.

**Index Terms**— Accuracy, BV error rate, CV error rate, Data mining techniques, Diabetes, C4.5, SVM, *k*-NN, PNN, BLR, MLR, PLS-DA, PLS-LDA, *k*-means and Apriori algorithms.

## 1 INTRODUCTION

Basic understanding on growth and factors affecting diabetes from external sources is required before building predictive models. Our 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. Basically declared, data mining refers to extracting or "mining" knowledge from large amounts of data or databases. The development of finding useful patterns or importance in raw data has been called KDD (knowledge discovery in databases). Bulky number of data mining algorithms has been developed in modern days for mining of knowledge in databases. Of these many are supervised learning algorithms. These algorithms are generally used for categorization tasks. The importance in systems for independent decisions in medical and manufacturing applications is increasing, as data becomes available. 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.

- *Director, IERDS, Maddur Nagar, Kurnool, Andhra Pradesh, India, Phone: +918374529162. e-mail: krlakshmi_cse@yahoo.com*
- *Professor & Head, Department of CSE&IT, G.Pullaiah college of Engineering & Technology, Nandikotkur Road, Kurnool, Andhra Pradesh, India, Ph-+919866504950, e-mail: mcahod@gpcet.ac.in*

The utilization of computer tools in medical decision support is now well-known and pervasive across a wide range of medical area such as diabetes, cancer etc. 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. Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The knowledge must be new, not obvious, and one must be able to use it. 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.

Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understand 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 Heart disease, Diabetes and

Breast cancer disease predictions. 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 Review of the related literature

There are diverse kinds of studies for DM techniques in medical databases. J.W.Smith et al [16] dealing with this data base uses an adaptive learning routine that generates and executes digital analogy of perceptions-like devices, called ADAP. They used 576 training instances and obtained a classification of 76% on the remaining 192 instances. Classification is the most widely used technique in medical data mining. Later Asha Rajkumar and S. Reena [5], and A. Khemphlila and V. Boojing [11] discussed various data mining techniques for diagnosis of certain life threatening deseases. K. Srinivas et.al [12] studied the applications of Data Mining Techniques in health care and Prediction Heart Attacks making use of some data base. Utilization of different data mining techniques has been studied by Elma kolce et.al [1]. 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. Huy Nguyen A.P. et.al [17] proposed a new algorithm Homogeneity-Based Algorithm to determine over fitting and over generalization behavior of classification. The algorithms used in this paper are Support Vector Machine, Decision Tree and Artificial Neural Networks. They predict whether a new patient would test positive for diabetes. This paper studied a new approach, called the Homogeneity- Based Algorithm (or HBA) to determine optimally control the over fitting and overgeneralization behaviors of classification on this dataset (Pima Indian diabetes data set). The HBA is used in conjunction with classification approaches (such as Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), or Decision Trees (DTs)) to enhance their classification accuracy. Some expermental results seem to indicate that the proposed approach significantly outperforms current approaches. From the experiment the author concluded that it is very important both for accurately predicting diabetes and also for the data mining community, in general. M.S..Sapna [18] for predicting diabetic status the author uses data mining algorithm for testing the accuracy. The author implemented using genetic algorithm. Here Data mining algorithm is used for testing the accuracy in predicting diabetic status. Fuzzy Systems are been used for solving a wide range of problems in different application domain Genetic Algorithm for designing. Fuzzy systems allows in introducing the learning and adaptation capabilities. Neural Networks are efficiently used for learning membership functions. Diabetes occurs throughout the world, but Type 2 is more common in the most developed countries. The author implemented in Genetic Algorithm. The steps involved in this algorithm namely selection, crossover, mutation, fitness and population statistics. As a result the author concluded that the optimization of chromosome using GA is obtained and it is based on the rate of old population diabetes can be restricted in new population to get chromosomal accuracy. Recently Karthikeyini et.al [41 & 42] discussed comparison a performance of data mining algorithms for diabetes disease based on computing time, precision value , the data evaluated using 10 fold Cross Validation error rate, error rate focuses True Positive, True Negative, False Positive and False Negative, bootstrap validation and accuracy.

## 3 Data Analysis

The most important methodology use for this paper throughout the analysis of journals and publications in the field of medicine. The explore focused on more recent publications. The data study consists of diabetes dataset. It includes name of the attribute as well as the explanation of the attributes. UCI and Indian Council of medical Research–Indian Diabetes (ICMR-INDIAB) study has provides data from three states and one Union Territory, representing nearly 18.1 percent of the nation's population. When extrapolated from these four units, the conclusion is 62.4 million people live with diabetes in India, and 77.2 million people are on the threshold, with pre-diabetes. It factored in anthropometric parameters like body weight,BMI (body Mass Index),height and weight limits and also tested fasting blood sugar after glucose load (known diabetes exempted),and cholesterol for all participant. The occurrences of pre-diabetes (impair fasting glucose and/or impair glucose tolerance) was 8.3 percent, 12.8 percent, 8.1 percent, 14.6 percent correspondingly. Nineteen years to the lead of that deadline, India has 62.4 million, and further 77.2 million (potential diabetes) in the pre-diabetes period. According to the diabetes atlas of 2009, there were 50.8 million people with diabetes in India. All data collected were stored electronically. The following fields linked all records: name, date of birth, and individual study identification number. All statistical analyses were performed using SAS for Windows version 9.0 software (SAS Institute, Inc., Cary, NC) on an IBM-compatible computer. Preliminary descriptive analysis was conducted to check for the distribution of the variables of interest, and log transformation was carried out where data were not normally distributed.

The ICMR–INDIAB study is the first effort to provide accurate and comprehensive state and national level data on prevalence of diabetes in India. It addresses limitations of previous non representative studies and, when completed, should provide robust and reliable estimates of diabetes prevalence in India, removing the need for modeling projections from one or two studies. This study is also unique in that it is designed to cover both rural and urban areas and provide estimates for prediabetes, dyslipidemia, hypertension, obesity, and the level

of glycemic control among the confirmed cases of diabetes. Thus the ICMR–INDIAB study will provide an accurate snapshot of the burden associated with diabetes in India. The Indian Council of Medical Research and the Madras Diabetes Research Foundation deserve praise for this massive undertaking, which will highlight areas for policy action and establish a national framework for non communicable disease (NCD) surveillance. The ICMR–INDIAB survey lays the foundation for effective NCD prevention and control and for applied public health research. New figures for diabetes prevalence in India indicate that the epidemic is progressing rapidly across the nation, reaching a total of 62.4 million persons with diabetes in 2011. Phase one results of the Indian Council of Medical Research – India Diabetes (ICMR-INDIAB) Study have provided data from three States and one Union Territory, representing nearly 18.1 per cent of the nation's population. When extrapolated from these four units, the conclusion is 62.4 million people live with diabetes in India, and 77.2 million people are on the threshold, with pre-diabetes. These results have been published in an article authored by R.M. Anjana et al [20 and 22], and also estimated that, in 2011, Maharashtra will have 6 million individuals with diabetes and 9.2 million with pre-diabetes, Tamilnadu will have 4.8 million with diabetes and 3.9 million with pre-diabetes, Jharkhand will have 0.96 million with diabetes and 1.5 million with pre-diabetes, and Chandigarh will have 0.12 million with diabetes and 0.13 million with pre-diabetes. Projections for the whole of India would be 62.4 million people with diabetes and 77.2 million people with pre-diabetes.

The first phase of the ICMR-INDIAB study covered Tamil Nadu, Maharashtra, Jharkhand and Chandigarh, with a sample size of 16,000 persons. "The results are amazing and 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. The prevalence of diabetes in Tamil Nadu was 10.4 per cent, in Maharashtra it was 8.4 percent, in Jharkhand, 5.3 percent, and in terms of percentage, highest in Chandigarh at 13.6. The prevalence of pre-diabetes (impaired fasting glucose and/or impaired glucose tolerance) was 8.3 percent, 12.8 percent, 8.1 percent, and 14.6 per cent, respectively. 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. "According to the Diabetes Atlas of 2009, there were 50.8 million people with diabetes in India. In just two years, this figure has gone up by 12 million. Obviously, diabetes in India is progressing exponentially. Also, we see that 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. The numbers of pre-diabetics will drop. We also expect that by then, the epidemic will shift to the economically disadvantaged groups, going by the experience of nations in the West," Dr. Mohan [19] added. Also, he explained that there was a huge window of opportunity for prevention, considering the number of modifiable risk factors among the pre-diabetes group. The three-phased study, when concluded, hopes to have done similar analyses for all the States and union territories in India.

Finally we obtained 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%, $\rho < 0.0001$). Diabetes showed a positive and independent association with age, BMI, WHR, family history of diabetes, monthly income 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. **Prediabetes** is a condition when patient **blood sugar level** triggers higher than normal, but not so high that we can validate it as type 2 diabetes. Gestational diabetes is a form of diabetes which affects pregnant women. It is thought that the hormones created during pregnancy reduce a woman's receptivity to insulin, leading to high **blood sugar** levels. **Gestational diabetes** affects on 4% of all pregnant women.

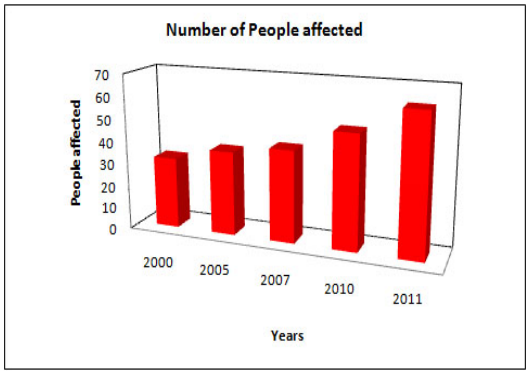


Fig. 1. Incresing Occurrence of Diabetes in India.

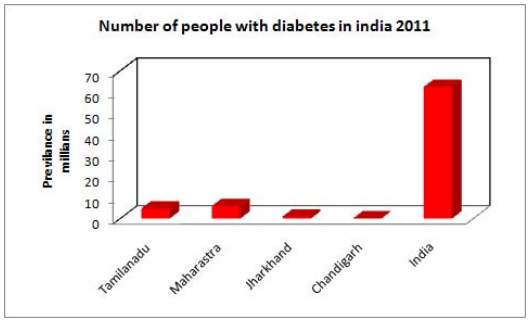


Fig. 2. Prevalance of Diabetes in India: ICMR-INDIAB Study

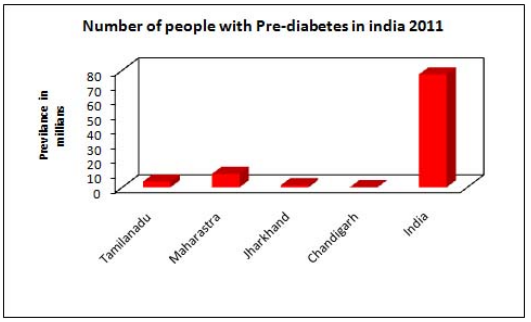


Fig. 3. Prevalance of Pre-diabetes in India: ICMR-INDIAB Study

## 4 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 modelling** 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 database 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 provided, 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 overfit the data.

### 4.1 Machine Learnimg Approches

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, BLR, MLR, PLS-DA, PLS-LDA, k*-means and Apriori).

#### 4.1.1 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. We will outline the algorithms employed in C4.5, highlight some changes in its successor See5/C5.0, and conclude with a couple of open research issues.

#### 4.1.2 SVM

Support vector machines (SVM). Support vector machines are a moderately new-fangled type of learning algorithm, origi-

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

#### 4.1.3 *k*-NN

It is the nearest neighbour algorithm. The *k*-nearest neighbour'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 [30]. 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 theinitial *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.[38] 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 membership functions for each clusters rather than a hard one.

#### 4.1.4 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.

#### 4.1.5 BLR

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.

#### 4.1.6 MLR

A multinomial logit (MNL) model, also known as multinomial logistic regression, is a regression model which generalizes logistic regression by allowing more than two discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.). An extension of the binary logistic model cases where the dependent variable has more than two categories is the multinomial logistic Regression. In such cases collapsing the data into two categories not make good sense or lead to loss in the richness of the data. The multinomial legit model is the appropriate technique in these cases, especially when the dependent variable categories are not ordered. Multinomial regression to include feature selection/importance methods.

#### 4.1.7 PLS-DA & PLS-LDA

PLS Regression for Classification Task PLS (Partial Least Squares Regression) Regression can be viewed as a multivariate regression framework where to predict the values of several PLS-LDA (Partial Least squares-Linear Discriminant Analysis target variables ($Y_1$, $Y_2$ …) from the values of several input variables ($X_1$,$X_2$,…)[27 & 28]. The algorithm use three axis for the diabetes disease is the following: The components of X are used to predict the scores on the Y components, and the predicted Y component scores are used to predict the actual values of the Y variables. In constructing the principal components of X, the PLS algorithm iteratively maximizes the strength of the relation of successive pairs of X and Y component scores by maximizing the covariance of each X-score with the Y variables. The PLS Regression is initially defined for the prediction of continuous target variable. But it seems it can be useful in the supervised learning problem where we want to predict the values of discrete attributes. In this tutorial we propose a few variants of PLS Regression adapted to the prediction of discrete variable. The generic name "PLS-DA" (Partial Least Square Discriminant Analysis) is often used in the literature. To predict the values of the dependent variable for unseen instances (or unlabeled instances) from the observed values on the independent variables. The process is rather basic if handle a linear regression model. Apply the computed parameters on the unseen instances.

#### 4.1.8 The *k*-means algorithm

The *k*-means algorithm is a simple iterative method to partition a given dataset into a serspecified number of clusters, *k*. This algorithm has been discovered by several researchers across different disciplines. A detailed history of *k*-means alongwith descriptions of several variations are given in [40]. Gray and Neuhoff [39] provide a nice historical background for *k*-means placed in the larger context of hill-climbing algorithms. The algorithm operates on a set of *d*-dimensional vectors, $D = \{\mathbf{x}_i \mid i = 1, \ldots, N\}$, where $\mathbf{x}_i \in R^d$ denotes the *i*th data point. The algorithm is initialized by picking *k* points in $R^d$ as the initial *k* cluster representatives or "centroids". Techniques for selecting these initial 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. Then the algorithm iterates between two steps till convergence:

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

Step 2: Relocation of "means". Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

#### 4.1.9 The Apriori algorithm

One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. Let the set of frequent itemsets of size k be $F_k$ and their candidates be $C_k$. Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.

1. Generate $C_{k+1}$, candidates of frequent itemsets of size k +1, from the frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to $F_{k+1}$.

Function apriori generates $C_{k+1}$ from $F_k$ in the following two step process:

1. Join step: Generate $R_{K+1}$, the initial candidates of frequent itemsets of size k + 1 by

taking the union of the two frequent itemsets of size k, Pk and Qk that have the first k−1

elements in common.

$R_{K+1} = P_k \cup Q_k = \{item_1, item_2, \ldots, item_{k-1}, item_k, item_{k'}\}$

$P_k = \{item_1, item_2, \ldots, item_{k-1}, item_k\}$

$Q_k = \{item_1, item_2, \ldots, item_{k-1}, item_{k'}\}$

where, $item_1 < item_2 < \cdots < itemk < itemk'\_$ .

2. Prune step: Check if all the itemsets of size k in $R_{k+1}$ are frequent and generate $C_{k+1}$ by removing those that do not pass this requirement from $R_{k+1}$. This is because any subset of size k of $C_{k+1}$ that is not frequent cannot be a subset of a frequent itemset of size k + 1.

Function subset finds all the candidates of the frequent itemsets included in transaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Apriori scans the database at most $k_{max}$+1 times when the maximum size of frequent itemsets is set at $k_{max}$.

### 4.2 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. In each iteration, 9 subsets of data are used as training data and the remaining set is used as testing data. In rotation, each subset of data serves as the testing set in exactly one 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.

## 5 Research Findings

### 5.1 Data mining in Diabetes Disease Prediction

Ten different supervised classification algorithms i.e. C4.5, SVM, K-NN, PNN, BLR, MLR, PLS-DA, PLS-LDA, *k*-means and Apriori have been used analyze dataset in. 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.

### 5.2 Tanagra

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.

### 5.3 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 1).

| S.No. | Name | Description |
|---|---|---|
| 1. | Pregnancy | Number of times pregnant |
| 2. | Plasma | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| 3. | Pres | 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. | Pedi | Diabetes pedigree function |
| 8. | Age | Age (in years) |
| 9. | Class | Class variable (0 or 1) |

**Table 1**. Attributes of diabetes dataset

### 5.4 Performance shown by Algorithms

SVM, PNN, BLR, MLR, PLS-DA, *k*-means and Apriori 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 2).

| | Classified as Healthy | Classified as not Healthy |
|---|---|---|
| **Actual Healthy** | TP | FN |
| **Actual not Healthy** | FP | TN |

**Table 2:** 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 labelled instances that were predicted as positive.

The fitness criteria are calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

$$\text{Accuracy} = \frac{TP + FN}{TP + FP + TN + FN}$$

$$\text{Positive Precision} = \frac{FP}{TP + FP}$$

$$\text{Negative Precision} = \frac{FN}{TN + FN}$$

$$\text{Error Rate} = \frac{FP + FN}{TP + FP + TN + FN}$$

**Step 1:** The ten algorithms can be filtered by using lowest computing time (<550ms). The ten can be reduced seven algorithms namely (SVM, PNN, BLR, MLR, PLS-DA, *k*-means and Apriori).

**Step 2**: The above algorithms can filtered by using positive precision values. If the precision value is greater than 0.1.we get the six algorithms namely (SVM, PNN, BLR, MLR, PLS-DA, *k*-means and Apriori).

**Step 3**: The above algorithms can filter by using Cross Validation Error rate (< 0.3) i.e. lowest error rate. The above six algorithms can be reduced. We get four algorithms namely (SVM, BLR, MLR, PLS-DA, *k*-means and Apriori)

**Step 4**: The above algorithms can filter by using Bootstrap Validation Error rate (< 0.29) i.e. lowest error rate. The above four algorithms can be reduced. We get three algorithms namely

(BLR, MLR, PLS-DA, *k*-means and Apriori)

**Step 5**: The above algorithms can filter by using highest accuracy and lowest computing time. The above three algorithms can be reduced to one. We get best one for PLS-DA.

**Step 6**: Stop the process. We get the best one.

The step5 consists of values of different classification. According to these values the accuracy was calculated. The figures (4-6) represents the resultant values of above classified dataset using data mining supervised classification algorithms and it shows the highest accuracy and lowest computing among the three. It is logical from chart that compared on basis of performance and computing time, precision value, Error rate (10 fold Cross Validation, Bootstrap Validation) and finally the highest accuracy and again lowest computing time. PLS-DA algorithm shows the superior performance compared to other algorithms.

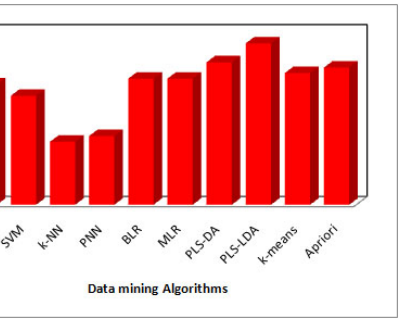


Fig. 4. Predicted Accuracy

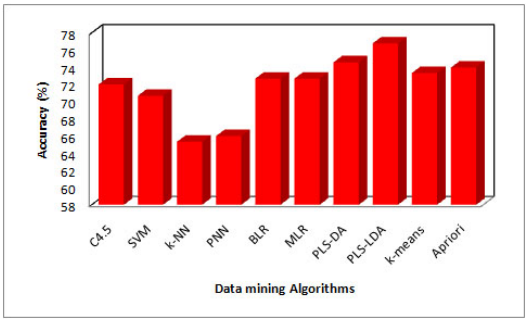


Fig. 5. Performance of cross validiation error rate

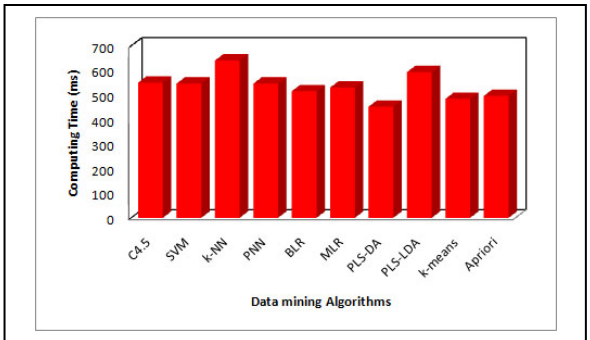


Fig. 6. Performance of computing time

## 6 Conclusion

The main goal medical data mining algorithm is to get best algorithms that describe given data from multiple aspects. The algorithms are very necessary for intend an automatic classification tools. With help of automatic design tools to reduce a wait in line at the experts. The PLS-DA was the best one among ten (five criteria are satisfied). Three axis are used the redundancy cut value is 0.0250, positive and negative values are predicted based on the recall and 1-precision values. It can be classified as function as positive and negative and finally constant value of positive and negative. The first one is computing time in 452 milliseconds it is the lowest, second one is Cross Validation error rate is 0.2667 , third positive precision values are greater than 0.1 , fourth one Bootstrap Validation error rate is 0.2726 lowest (i.e. repetition is 1, test error rate 0.2747,Bootstrap ,Bootstrap+) compare to others and finally three values(Accuracy, Specificity and Sensitivity) are calculated by using formula and the prediction one is accuracy. Then the Accuracy of PLS-DA is 74% from the above results PLS-DA algorithm plays a vital role in data mining techniques.

## Acknowledgments

The authors are thankful to Prof. C.Uma Shankar, Dept. of OR&SQC and Dr. M. Veera Krishna, Department of Mathematics, Rayalaseema University, Kurnool, Andhra pradesh, India, for their valuable guidance and suggestions with thought provoking discussions throughout the period of my research and in the preparation of this paper, and IJSER Journal for the support to develop this document.

## 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.

[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]. F.Hosseinkhah, H.Ashktorab, R.Veen, M. M. Owrang O (2009), "Challenges in Data Mining on Medical Databases ", *IGI Global*, pp. 502-511.

[5]. Asha Rajkumar,Sophia Reena.G., "Diagnosis of Heart Disease Using Data Mining Algorithm", *Global Journal of Computer Science and Technology*, Vol-10, 2010, pp. 38-46.

[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 "DataMining 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 Regression 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) "DataMining 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 accurate 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 fasting glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Research-INdia DIABetes (ICMR-INDIAB) study". *Diabetologia*, Vol. 54 (12) , Dec 2011, pp. 3022-3027.

[23]. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", second edition, Morgan Kaufmann Publishers an imprint of Elsevier.

[24]. Cover, T., Hart P., "Nearest Neighbour Pattern Classification", *IEEE Trans Inform Theory*, Vol. 13(1), 1967, pp. 21–27.

[25]. Breiman, L., Friedman, J., Olsen,R., Stone, C., 1984. "Classification and Regression Trees", Chapman & Hall, 1984, pp. 185-189.

[26]. Dayle, L., Sampson, Tony J., Parker, Zee Upton, Cameron, P., Hurst, "Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine and SIMCA classification, *Journal of Chemo metrics*, Vol. 20(8–10), September 2011, pp. 341–351.

[27]. Barker, M., & Rayens, W., "Partial least squares for discrimination", *Journal of Chemo metrics*, Vol. 17(3), 2003, pp. 166–173.

[28]. Bylesjo, M., Rantalainen, M., Cloarec, O., "OPLS discriminant analysis: Combining the strengths of PLS-DA", 2006.

[29]. Breiman,L.,Friedman,J.,Olsen,R., Stone.C, "Classification and Regression Trees", Chapman & Hall, 1984.

[30]. Cover.,T.M., Hart, P.E.,"Nearest neighbor pattern classification", *IEEE Trans. Inform Theory*, Vol. IT-13, Jan 1967, pp. 21-27.

[31]. Barker, M., & Rayens, W, "Partial least squares for discrimination", *Journal of Chemo metrics*, Vol. 17(3), 2003, pp. 166–173.

[32]. Ramakrishna, Gehrkev, "Database Management Systems", *International Edition, TMH*, p. 929.

[33]. David,A.,Aoyama, Jen-Ting,T., "TimeLine and visualization of multiple-data sets and the visualization querying challenge", *Journal of visual languages and Computing*, Vol. 18, 2007, pp. 1-21.

[34]. Chau, M., Shin,D., "A Comparative study of Medical Data classification Methods Based on Decision Tree and Bagging algorithms", *Proceedings of IEEE International Conference on Dependable, Autonomic and Secure Computing*, 2009, pp.183-187.

[35]. Palaniappan, S., Awang, R., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp.108-115.

[36]. Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules", Seminar Presentation at University of Tokyo, 2004.

[37]. Liang Yanhong, Tan Runhua, "Text Mining-based Patent Analysis in Product Innovative Process", Hebei University of Technology.

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

[39]. Gray RM, Neuhoff DL., "Quantization", *IEEE Trans Inform Theory*, Vol. 44(6), 1988, pp. 2325–2384.

[40]. Jain AK, Dubes RC., "Algorithms for clustering data", *Prentice-Hall*, Englewood Cliffs, 1988.

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

[42]. 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.

## Appendix:

| S. No. | Alg. | CT (ms) | TP | FN | FP | TN | Acc. (%) | Spe. | Sen. | CVE rate | P (Prec) | N (Prec) | BVE rate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | C4.5 | 550 | 31 | 23 | 19 | 77 | 72.00 | 0.8021 | 0.5741 | 0.2800 | 0.3800 | 0.2300 | 0.3196 |
| 2 | SVM | 546 | 24 | 30 | 14 | 82 | 70.67 | 0.8541 | 0.4444 | 0.2933 | 0.3684 | 0.2678 | 0.2929 |
| 3 | *k*-NN | 640 | 20 | 34 | 18 | 78 | 65.33 | 0.8125 | 0.3703 | 0.3466 | 0.4736 | 0.3035 | 0.3532 |
| 4 | PNN | 546 | 42 | 12 | 39 | 57 | 66.00 | 0.5937 | 0.7778 | 0.3400 | 0.4814 | 0.1739 | 0.3406 |
| 5 | BLR | 515 | 32 | 22 | 19 | 77 | 72.67 | 0.8021 | 0.5925 | 0.2733 | 0.3725 | 0.2223 | 0.2754 |
| 6 | MLR | 530 | 32 | 22 | 19 | 77 | 72.67 | 0.8021 | 0.5925 | 0.2733 | 0.3725 | 0.2223 | 0.2754 |
| 7 | PLS-DA | 452 | 25 | 21 | 16 | 83 | 74.48 | 0.8384 | 0.5435 | 0.2552 | 0.3902 | 0.2019 | 0.2782 |
| 8 | PLS-LDA | 593 | 36 | 20 | 16 | 83 | 76.78 | 0.8384 | 0.6429 | 0.2323 | 0.3077 | 0.1941 | 0.2726 |
| 9. | *k*-mean | 484 | 28 | 23 | 18 | 81 | 72.66 | 0.8182 | 0.5491 | 0.2733 | 0.3913 | 0.2212 | 0.2734 |
| 10. | Apriori | 496 | 27 | 20 | 18 | 85 | 74.67 | 0.8252 | 0.5745 | 0.2533 | 0.4000 | 0.1905 | 0.2733 |

**Table 3**: Comparison of supervised Algorithms based on performance

Alg.-Algorithm names, CT- Computing Time, TP-True Positive,FN-False Negative, FP-False Positive, TN True Negative, Acc-Accuracy, Spec-Specificty, Sen-Senitivity, CVE rate-CrossValidation Error rate, P(Prec)-Positive Precision, N(Prec)-Negative Precision, BVE rate-Bootstrap Validation Error rate.

## BIOGRAPHY OF AUTHORS

**K.R.Lakshmi:** She has completed Master degree in Computer Applications in 2010 from Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India. She is a Director, IERDS, Maddur Nagar, Kurnool, Andhra Pradesh, India, Her teaching and research areas Data mining techniques. She has published 2 articles in international well reputed journals.

**S.Prem Kumar:** He received Ph.D. degree in Computer Science and Technology from Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, in 2010. He is Professor of computer science and engineering, Department of CSE&IT, G.Pullaiah college of Engineering & Technology, Nandikotkur Road, Kurnool, Andhra Pradesh, India. His teaching and research areas include Data mining techniques, mobile computing and Internet frame works. He has published 10 articles in national and international well reputed journals.