AN ANALYSISOF DATA MINING CLASSIFICATION METHODS FOR MEDICAL DATASET

Dr. Y. Angeline Christobel
Dean, School of Computational Studies
Hindustan College of Arts and Science
Padur, Chennai
angeline_christobel@yahoo.com

ABSTRACT
In knowledge discovery process, classification is an important technique of data mining and widely used in various fields. The development of data-mining applications such as classification and clustering has shown the need for machine learning algorithms to be applied to large scale data. The aim of this paper is to investigate the performance of different classification methods for diabetic medical dataset. The performance of C4.5, Naïve Bayes, SVM and Multilayer Perceptron algorithms are evaluated based on Accuracy, Sensitivity, Specificity and Error rate.

Key Words: Data Mining, Classification, C4.5,
SVM,MLP, Naïve Bayes

I. INTRODUCTION

Data mining is the extraction of hidden predictive information from large databases [1]. It uses well established statistical and machine learning techniques to build models that predict some behavior of the data. Data mining tasks can be classified into two categories: Descriptive and predictive data mining. Descriptive data mining provides information to understand what is happening inside the data without a predetermined idea. Predictive data mining allows the user to submit records with unknown field values, and the system will guess the unknown values based on previous patterns discovered from the database.

Data mining models can be categorized according to the tasks they perform:
1. Classification and Prediction
2. Clustering
3. Association Rules

Classification and prediction is a predictive model, but clustering and association rules are descriptive models.Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Classification is the task of examining the features of a newly presented object and assigning it to one of a predefined set of classes. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have [3, 4]. The most popular classification and prediction methods are

1. Decision Trees
2. Rule based
3. Bayesian
4. Support Vector Machines
5. Artificial Neural Network
6. Ensemble methods
7. Lazy Learners

Decision tree induction is the learning of a decision tree from class-labeled training tuples.
A *rule based classifier* is a technique for classifying records using a collection of "if … then" rules.
Bayesian classifiers are statistical classifiers and are based on Bayes theorem
Support Vector Machines has its roots in statistical learning theory and has shown promising empirical results in many applications.
An *Artificial Neural Network* is a computational model based on biological neural networks.
An *Ensemble method* constructs a set of base classifiers from training data and performs classification by taking a vote on the predictions made by each base classifier.
Lazy learning is a learning method where the system tries to generalize the training data before receiving queries.

The main objective of this paper is to analyze C4.5, Naïve Bayes, SVM and Multilayer Perceptron algorithms on the data set "Diabetes" obtained from the UCI Machine Learning Repository based on Accuracy, Sensitivity, Specificity and Error rate. These algorithms are among the top 10 algorithms in data mining [14].

II. CLASSIFICATION ALGORITHMS

Classification is a process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown[2].Classification algorithms have a wide range of applications like churn pre-diction, fraud detection, artificial intelligence, and credit card rating etc.

The C4.5, Naïve Bayes,SVM and Multilayer Perceptronalgorithms are discussed below.

A) C4.5 Algorithm

C4.5 is a popular and powerful decision tree classification algorithm used to generate a decision treedeveloped by Ross Quinlan. It is a successor of ID3. It constructs the decision tree with a 'divide and conquer' strategy. It eliminates the problem of unavailable values, continuous attributes value ranges, pruning of decision trees and rule derivation. In C4.5, each node in a tree is associated with a set of cases. Also these cases are assigned weights to take into account unknown attribute values. At the beginning, only the root is present and it is associated with the whole training set, and all the weights are equal to one. At each node the divide and conquer algorithm is executed, trying to exploit the locally best choice with no backtracking allowed. In building a decision tree, it is dealt with training set that have records with unknown attributes by considering only those records where those attribute values are available. The records that have unknown attribute values are classified by estimating the probability of the various possible results. C4.5 produces tree with variable branches per node. When a discrete variable is chosen as the splitting attribute in C4.5 there will be one branch for each value of attributes [5, 9].

B) Naïve Bayes Algorithm

The Naïve Bayes classifier produce probability estimates rather than predictions. For each class value they estimate the probability that a given instance belongs to that class.The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence [2].

The Naïve Bayesian classifier [2] works as follows:

1. Each data sample is represented by an*n*-dimensional feature vector, $X = \{x1, x2, \ldots xn\}$, depicting *n* measurements made on the sample from *n* attributes respectively $A_1, A_2, \ldots A_n$.

2. Suppose that there are m classes, $C_1, C_2, \ldots C_m$. Given an unknown data sample, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the Naïve Bayesian classifier assigns an unknown sample X to the class C_i if and only if:

$$P(C_i|\ X) > P(C_j|\ X) \text{ for } 1 \le j \le m\ ,\ j \ne i$$

Thus $P(C_i|\ X)$ is maximized. The class C_i for which $P(C_i|\ X)$ is maximized is called the maximum posteriori hypothesis.

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

3. As P(X) is constant for all classes, only $P(X|Ci)P(Ci)$ need to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely i.e. $P(C_1) = P(C_2) = \ldots = P(C_i)$ Note that the class prior probabilities may be estimated by

$$P(Ci) = \frac{si}{s}$$

Where s_i is the number of training samples of class Ci and *s* is the total number of training samples.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute$P(X|Ci)$. In order to reduce computation in evaluating $P(X|Ci)$the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample, ie., that there are no dependence relationships among the attributes. Thus

$$P(X/Ci) = \prod_{k=1}^{n} P(xk|Ci)$$

The probabilities $P(x1|Ci)$, $P(x2|Ci)$, …, $P(xn|Ci)$, can be estimated from the training samples xk refers to the value of attribute A_k for sample X which may be categorical or continuous valued.

5. In order to classify an unknown sample X, $P(X|Ci)P(Ci)$ is evaluated for each class Ci. Sample X is then assigned to the class Ci if and only if

$P(X|Ci)P(Ci) > P(X|Cj)P(Cj)$ for $1 \leq j \leq m, j \neq i$

In other words, it is assigned to the class Ci, for which $P(X|Ci)P(Ci)$ is the maximum.

C) SVM (Support Vector Machine) Algorithm

This algorithm is introduced by Vapnik et al. [11], is a very powerful method that has been applied in a wide variety of applications. The basic concept in SVM is the hyper plane classifier, or linear separability. To achieve linear separability, SVM applies two basic ideas: margin maximization and kernels, that is, mapping input space to a higher-dimension space, feature space.

SVM is an algorithm with strong regularization properties, that is, the optimization procedure maximizes predictive accuracy while automatically avoiding over-fitting of the training data. Neural networks and radial basis functions, both popular data mining techniques, have the same functional form as SVM models; however, neither of these algorithms has the well-founded theoretical approach to regularization that forms the basis of SVM.

SVM projects the input data into a kernel space. Then it builds a linear model in this kernel space. A classification SVM model attempts to separate the target classes with the widest possible margin. A regression SVM model tries to find a continuous function such that maximum number of data points lie within an epsilon-wide tube around it. Different types of kernels and different kernel parameter choices can produce a variety of decision boundaries (classification) or function approximators (regression).

D) Multi-layer Perceptron(MLP) Algorithm

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(\cdot) : R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, \ldots, x_m$ and a target y, it can learn a non-linear function approximation for either classification or regression. It is different from logistic regression, in that between the input and the output layer; there can be one or more non-linear layers, called hidden layers. Figure 1 shows a one hidden layer MLP with scalar output.

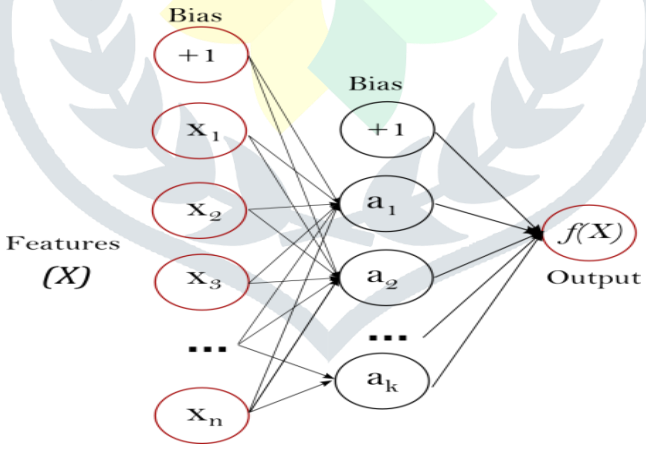


Figure 1: One hidden layer MLP.

The leftmost layer, known as the input layer, consists of a set of neurons $\{x_i | x_1, x_2, \ldots, x_m\}$ representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $w_1x_1 + w_2x_2 + \ldots + w_mx_m$, followed by a non-linear activation function $g(\cdot) : R \rightarrow R$ - like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values.The Multi-layer Perceptron has the capability to learn non-linear models and models in real-time (on-line learning) using partial_fit.

III. PERFORMANCE EVALUATION

Classifier performance depends on the characteristics of the data to be classified. In this paper, k-fold cross validation is used for evaluating the classifiers.

In k-fold cross validation, the initial data are randomly partitioned into k mutually exclusive subset or folds $d_1,d_2,\ldots,d_k$, each approximately equal in size. The training and testing is performed k times. In the first iteration, subsets d_2, …, d_k collectively serve as the training set inorder to obtain a first model, which is tested on d_1; the second iteration is trained in subsets d_1, $d_3,\ldots$, d_k and tested on d_2; and so on[2].

Performance of the selected algorithms is measured for Accuracy, Sensitivity, Specificity and Error rate from the confusion matrix obtained.

The Accuracy, Sensitivity, Specificity and Error rate can be defined as follows:

Accuracy = (TP+TN) / (TP + FP + TN + FN)
Sensitivity = TP/(TP + FN)
Specificity = TN /(TN + FP)
Error rate = (FP+FN) / (TP + FP + TN + FN)

Where TP is the number of True Positives
TN is the number of True Negatives
FP is the number of False Positives
FN is the number of False Negatives

IV. EXPERIMENTAL RESULTS

In this paper, 10-fold cross validation is applied for evaluating the performance of the classifiers. The "Diabetes" dataset, which is obtained from the UCI machine learning library [13] is used.Algorithm for attribute selection was applied on dataset to preprocess the data. The dataset contains 768 instances, 8 attributes and one class label.

Table 1 shows the Accuracy, Sensitivity, Specificityand Error rate of C4.5, Naïve Bayes, SVMand Multi-layer Perceptronalgorithms.

Figure2 shows the graphical representation of difference in Accuracy.

Figure3 shows the graphical representation of difference in Sensitivity.

Figure4 shows the graphical representation of difference in Specificity.

Figure5 shows the graphical representation of difference in Error rate.

Table 1: Comparison of Data Mining Models

Algorithms	Accuracy	Sensitivity	Specificity	Error rate
C4.5	74.21%	79.02%	63.67%	26.17%
Naïve Bayes	76.3%	80.22%	67.76%	23.69%
SVM	77.34%	78.49%	73.97%	22.65%
Multilayer Perceptron	62.36%	79.84%	42.85%	24.6%

Figure 2: Comparison based on Accuracy

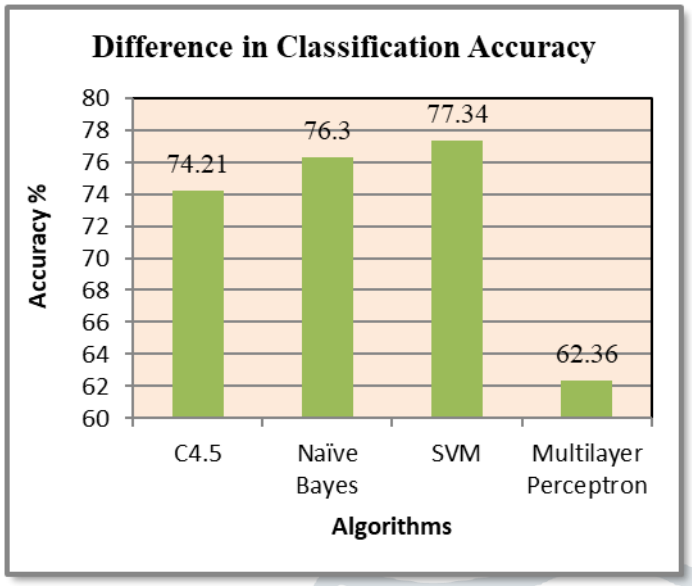


Figure 3: Comparison graph based on Sensitivity

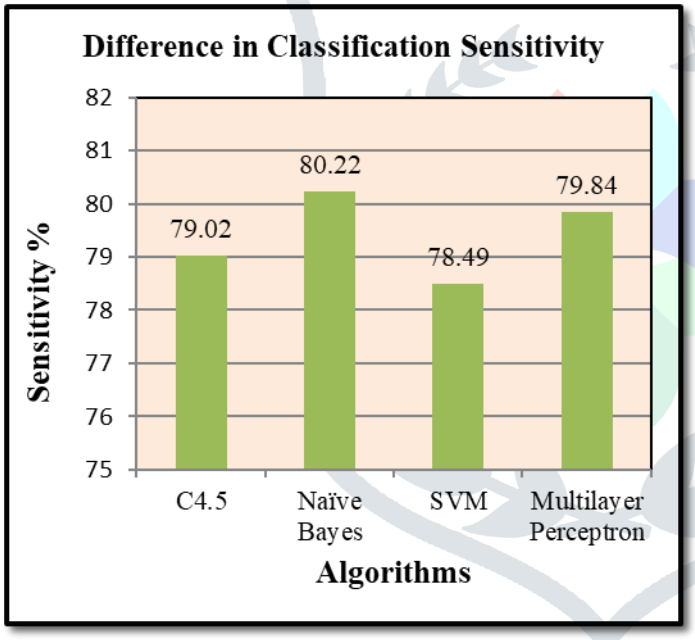


Figure 4: Comparison graph based on Specificity

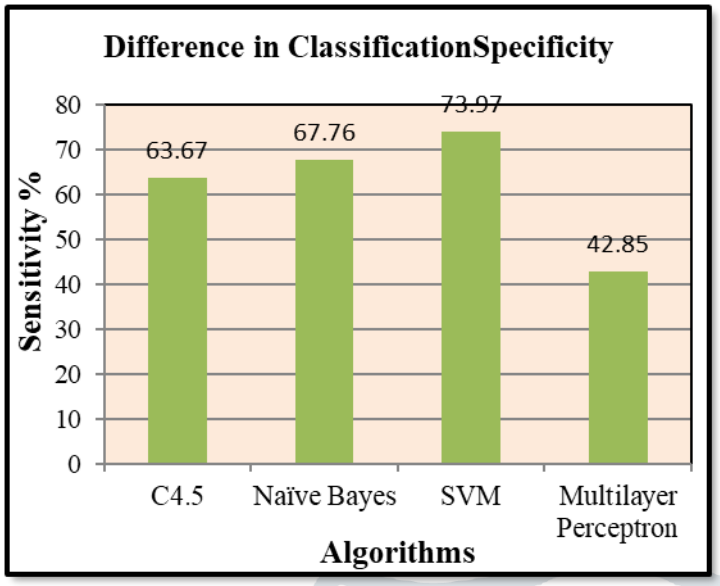
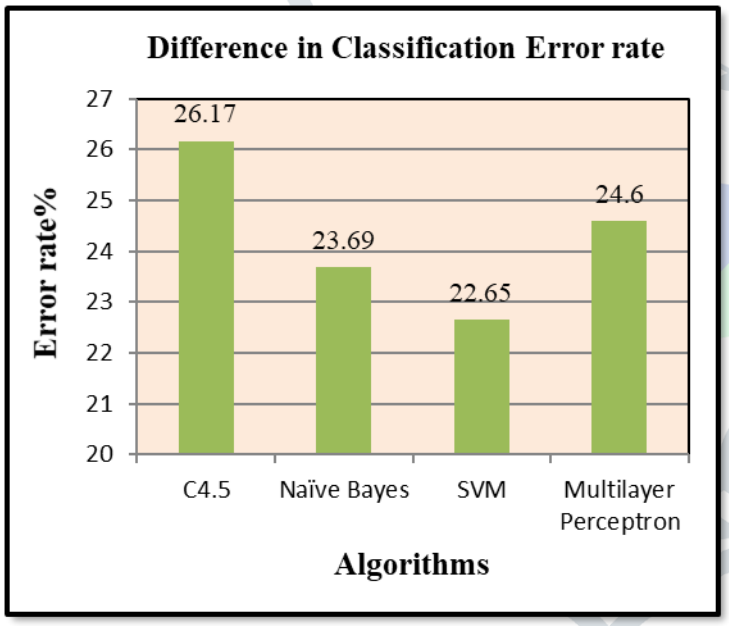


Figure 5: Comparison graph based on Error rate.



Theresult shows that out of C4.5, Naïve Bayes, SVM and Multllayer Perceptron algorithms, SVMperforms better classification. The error rate of SVM is low and the accuracy, sensitivity is very high compared to the other three models.

V. CONCLUSION

In this paper, the performance of C4.5, Naïve Bayes, SVM and Multilayer Perceptron were analyzed. The experiments were conducted on the medical dataset "Diabetes" from UCI Machine Learning Repository. The Classification Accuracy, Sensitivity, Specificity and Error rate is validated by 10-fold cross validation method. The Study shows that Support Vector Machine (SVM) turned out to be a best classifier.

VI. REFERENCES

1. KietikulJearanaitanakij,"Classifying Continous Data Set by ID3 Algorithm", Proceedings of fifth International Conference on Information Communication and Signal Processing, 2005.
2. J. Han and M. Kamber,"Data Mining Concepts and Techniques", Morgan Kauffman Publishers, USA, 2006.
3. Agrawal, R., Imielinski, T., Swami, A., "Database Mining:A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, pp. 914-925, December 1993.
4. Chen, M., Han, J., Yu P.S., "Data Mining: An Overview from Database Perspective", IEEETransactions on Knowledge and Data Engineering, Vol. 8 No.6, December 1996.
5. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
6. S.B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007
7. J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996.
8. J.R. Quinlan, "Induction of decision trees," In Jude W.Shavlik, Thomas G. Dietterich, (Eds.), Readings in Machine Learning. Morgan Kaufmann, 1990. Originally published in Machine Learning, vol. 1, 1986, pp 81–106.
9. Salvatore Ruggieri,"Efficient C4.5 Proceedings of IEEE transactions on knowledge and data Engineering",Vo1. 14,2,No.2, PP.438-444,20025
10. P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under Zero-one loss, Machine learning 29(2-3)(1997) 103-130.11
11. Vapnik, V.N., The Nature of Statistical Learning Theory, 1st ed., Springer-Verlag,New York, 1995.
12. Michael J. Sorich, John O. Miners,Ross A. McKinnon,David A. Winkler, Frank R. Burden, and Paul A. Smith, "Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP- Glucuronosyltransferase Isoforms"
13. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29]
14. XindongWu · Vipin Kumar · J. Ross Quinlan Joydeep Ghosh · Qiang Yang ·Hiroshi MotodaGeoffrey J. McLachlan · Angus Ng · Bing Liu ·Philip S. Yu ·Zhi-Hua Zhou · Michael Steinbach ·David J. Hand · Dan Steinberg, "Top 10 algorithms in data mining "Springer 2007
15. Niculescu-Mizil, A., &Caruana, R. (2005). Predicting good probabilities with supervised learning. Proc.22nd International Conference on Machine Learning (ICML'05).
16. Thair Nu Phyu, "Survey of Classification Techniques in Data Mining MultiConference of Engineers and Computer Scientists" 2009 Vol I IMECS 2009, Hong Kong
17. Arbach, L.; Reinhardt, J.M.; Bennett, D.L.; Fallouh, G.; Iowa Univ., Iowa City, IA, USA "Mammographic masses classification: comparison between backpropagation neural network (BNN), K nearest neighbors (KNN), and human readers", 2003 IEEE CCECE
18. D. Xhemali, C.J. Hinde, and R.G. Stone. Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. International Journal of Computer Science, 4(1):16-23,2009.
19. Rishi Gupta, Raghav Agarwal, Machine Learning Comparison Classification Algorithms, International Journal of Advanced Engineering and Global Technology I Vol-03, Issue-11, December 2015
20. Himani Raina, OmaisShafi Analysis Of Supervised Classification Algorithms, International Journal Of Scientific & Technology Research Volume 4, Issue 09, September 2015