A Comparative Study of Performance Metrics of Data Mining Algorithms on Medical Data

Ashok Suragala
Department of Information Technology
Jawaharlal Nehru Technological University Kakinada
University College of Engineering Vizianagaram
Vizianagaram, Andhra Pradesh, India
ashok.cse@jntukucev.ac.in

Vamsi Shankar Simhadri
Department of Information Technology
Jawaharlal Nehru Technological University Kakinada
University College of Engineering Vizianagaram
Vizianagaram, Andhra Pradesh, India
vamsishankar.it@gmail.com

Sai Likhith Yadav Seera
Department of Information Technology
Jawaharlal Nehru Technological University Kakinada
University College of Engineering Vizianagaram
Vizianagaram, Andhra Pradesh, India
likhithsyadav18@gmail.com

***Abstract*—Computers have brought about significant improvements in technology that lead to the creation of huge volumes of data, especially in health care systems. The availability of huge amounts of data resulted in a greater need for data mining techniques in order to generate useful knowledge. With the growth in data in bio-medical and health-care communities, accurate analysis of medical data benefits early disease detection and patient care.**

Data mining is one of the major approaches for developing sophisticated algorithms for classification of data. Data mining has been criticized by some for not following all of the requirements of classical statistics [5].Classification of diseases is one of the major application of data mining and in recent years, many significant attempts are made to increase the accuracy of disease diagnosis through data mining.

We used four popular data mining algorithms Naive Bayes Classifier, K-Nearest Neighbours(KNN) Classifier, Artificial Neural Networks(ANN), and Support Vector Machine(SVM) algorithms to develop the prediction models using ILPD (Indian Liver Patient data set) obtained from the UCI Machine learning repository. We used 10-fold cross-validation method to measure the estimate of the six prediction models for performance comparison purposes. We found out that support vector machine gives the best performance in a classification with an accuracy of 74.82% and Naive Bayes performed the worst with an accuracy of 56.55%. We have evaluated the other performance metrics below in the following paper.

***Index Terms*—Classification,Naive Bayes,KNN,SVM,ANN,Data mining.**

INTRODUCTION

There is a growing need for accurate classification of disease as if the dis- ease is detected early, then it becomes easy to cure rather than in the future stages. Classification techniques are very popular in various automatic medical diagnosis tools. There is a substantial increase in mobile devices which are used for monitoring humans body conditions. With the help of automatic classification tools for liver diseases, one can detect the disease in an early stage and curing of the disease becomes easy,reported that SVM classifier produces the best predictive performance for the chemical .

Lung-Cheng reported that Nave Bayesian classifier produces high performance than SVM and C 4.5 for the CDC Chronic fatigue syndrome dataset [3]. Paul R Harper [2] reported that there is not necessary a single best classification tool but instead the best performing algorithm will depend on the features of the to be analyzed. Michael J Sorich [1] reported that SVM classifier produces best predictive performance for the chemical datasets. In this paper, we demonstrate the performance of four data mining algorithms namely Naive Bayes, KNN classifier, Support vector machines, and ANN Classifier algorithm. [6]In short, data mining is not aiming to replace medical professionals and medical researchers, but to supplement their efforts to speed up research in this field so diseases can be diagnosed and treated much better.

CLASSIFICATION ALGORITHMS

Classification algorithms are widely used in various applications for classifying the data. Data classification is a two phase process in which the first step is the training phase where the classifier algorithm builds the classifier with the training set of tuples and the second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples. Below we introduce all the algorithms which we demonstrate further.

A. Naive Bayes Classifier

The Naive Bayes Classifier is the most well-known representation of statistical learning algorithm. The Naive Bayes model is a heavily simplified Bayesian probability model[13].

Naive Bayes Algorithm uses probability theory as an approach to concept classification.

Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods [14], [15]. Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem.[9]It is not a single algorithm, but a family of algorithms in which all of them share a common principle, i.e. every pair of features being classified is independent of each other. The Naive Bayes classifier operates on a strong independence assumption [13]. It is very simple and shows high precision and speed when applied to large databases.

It works on one inference that is the effect of an attribute value of a given class is independent of the values of the other attributes. This inference is called class conditional independence. Using a few statistical tests like Chi-squared and mutual information tests we can find the conditional independence relationships among the features and use these relationships as constraints to construct a Bayesian Network.

B. K-Nearest Neighbors(KNN) Algorithm

KNN, also known as K-Nearest Neighbors, is one of the simplest supervised Machine Learning algorithms which is mainly based on the feature similarity, it is mainly focused on the classification problems in the industry.That is, it classifies a data point based on how its neighbours are classified.KNN stores all available cases and classifies new cases based on a similarity measure from the existing.

KNN algorithm is a commonly used algorithm as it is known for its easy interpretation, effectiveness in predicting and low calculation time. The value of "k" in the KNN algorithm is a factor that refers to the number of nearest neighbours in order to include in a majority voting process. Choosing the correct value of "k" is a process called "Parameter Tuning" and it is important for better accuracy. KNN algorithm is well used when the data are labelled, noise-free and the data set is small. Because KNN is a "lazy learner"(It does not learn a discriminative function from the training set), it cannot be used for complex data sets.

[10]This algorithm also requires calculating Euclidean Distance to find the nearest neighbours of the unknown data point from all the points in the data set. The most common classification is found out from the samples in the data set so that the new sample is assigned to this classification.

C. Support Vector Machine(SVM)

Support Vector Machine is specified in supervised machine learning method mainly used in classification. SVM deals with the data sets in order to sort the data into one of the categories. In this method, each element in the given data set is represented as a point which is plotted in n-dimensional space. The value of each feature represents the value of a particular coordinate in the plane.Here, n is the number of features of the data set.

Then, [8]classification is performed in order to find the suitable hyper-plane that differentiates the two classes of support vectors. Support Vectors in the plane are simply the points of coordinates that represent the individual observation of data items in a given data set. For the given support vectors, identifying the right hyper-plane for segregating the two classes better is the primary objective in the classification.

Generally, if the distances between the nearest point or class and hyper-plane are maximized, it is well considered in deciding the right hyper-plane. This distance is known as margin. So, SVM is a frontier which best segregates the two classes of support vectors. Advantages in using SVM are that it supports High dimensional input space, Sparse document vectors as well as Regularization parameter.

D. Artificial Neural Networks(ANN)

[7]Artificial neural networks (ANN's) are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modelling extremely complex nonlinear functions. Formally defined, are analytic techniques modelled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations after executing a process of so-called learning from the existing data.

We used a popular ANN architecture called multi-layer (MLP) with back-propagation in the following paper. The multilayer is sometimes colloquially referred to as "neural networks", especially when they have a single hidden layer.The Back Propagation Algorithm is a multi-layered Neural Networks for learning the rules, credited to Rumelhart and McClelland.

METHODOLOGY

Data Source

[4]In order to perform the research experiments reported in the following paper, we used the data contained in the UCI Machine Learning repository with the name ILPD. The ILPD data set consists of 416 liver patient records with the disease and 167 patient records with no disease. This data set contains 11 attributes and is regarded as one of the most comprehensive sources of data from Andhra Pradesh.

Data Understanding and Preparation

The data understanding and the data preparation also known as data stages are among the most important steps in the data mining applications. A vast majority of time spent on developing data mining applications is accounted for the stage. Almost most of the time and effort in this research project were spent on cleaning and preparing the data for predictive modelling.

The ILPD data consisted of 583 records/cases and 11 attributes. These 11 attributes are Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline, Sgpt Alanine Aminotransferase,

Aspartate, Total Protein, Alb Albumin, A/G ratio Albumin and Globulin ratio and outcome variable i.e. affected by disease or not. The outcome variable was identified as the predictable attribute with value 0 for patients with liver disease and value 1 for patients with no liver disease. The data consisted of several missing records, so we decided to remove all the missing records.

Since these attributes are considered to be important in predicting, rather than deleting the attributes, the records containing the missing data were removed from the. Further analysis was performed to check the effect other variables of deleting these records. This analysis showed that there was no significant change in the distribution of the other variables.

For instance, the histograms presented in below figure1 show the distribution plot before deletion and figure 2 shows the distribution plot after deletion of the records. Comparing the two plots we find that there is no significant change in the distribution of age attribute and Total Bilirubin attribute before and after the deletion of these missing valued records.

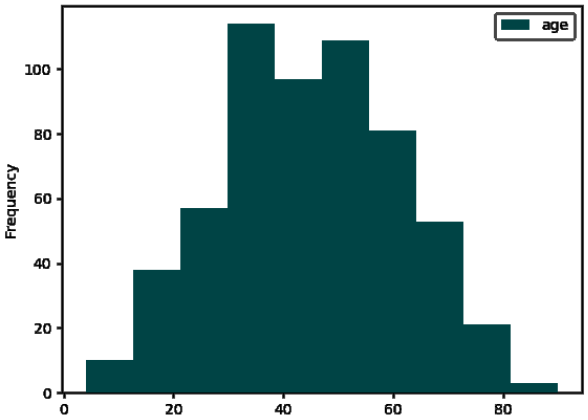

Fig. 1. This figure shows the histogram of the distribution of the age variable before deleting the missing data from the data set

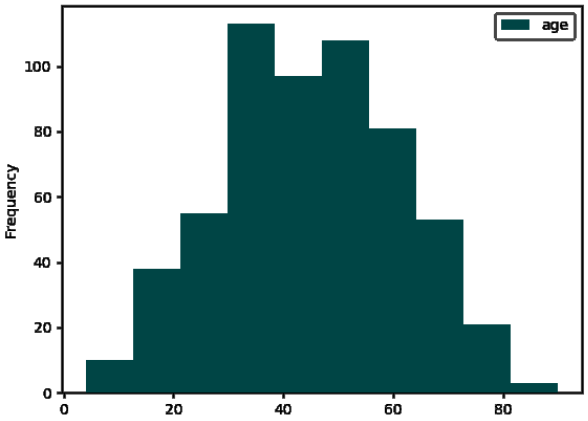

Fig. 2. This figure shows the histogram of the distribution of the age variable after deleting the missing data from the data set

Sensitivity, Specificity and Accuracy

In this study we used three types of performance evaluation measures namely Sensitivity, Specificity and Accuracy.

1) Sensitivity: Sensitivity is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly called as Recall, and corresponds to the true positive rate.

Sensitivity is the probability that a test result will be positive when the disease is present in the body.

$$\frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \tag{1}$$

where

$TruePositive$ = numbers of true positive predictions for the considered class

$FalseNegative$ = numbers of false negative predictions for the considered class

$(TruePositive + FalseNegative)$= total number of test examples of the considered class

2) Specificity: Recall/sensitivity is related to specificity, which is a measure that is commonly used in two class problems where one is more interested in a particular class. Specificity corresponds to the true-negative rate.

Specificity is the probability that a test result will be negative when the disease is not present in the body.

$$\frac{True\ Negative}{True\ Negative\ +\ False\ Positive} \tag{2}$$

where

$TrueNegative$ = numbers of true negative predictions for the considered class

$FalsePositive$ = numbers of false positive predictions for the considered class

$(TrueNegative + FalsePositive)$= total number of test examples of the considered class

3) Accuracy: Accuracy is the overall probability that a patient will be accurately classified.

$$\frac{T_P\ +\ T_N}{(T_P\ +\ T_N\ +\ F_P\ +\ F_N)} \tag{3}$$

where

$T_P = TruePositive$=numbers of true positive predictions for the considered class;

$T_N = TrueNegative$=numbers of true negative predictions for the considered class;
$F_P = FalsePositive$=numbers of false positive predictions for the considered class;
$F_N = FalseNegative$=numbers of true negative predictions for the considered class;
$(T_P + T_N + F_P + F_N)$= total number of test examples of the considered class

K-fold cross validation

In order to reduce the bias related to the random sampling of the training and testing data samples in comparing the classification accuracy of two or more methods, researchers use the k-fold cross-validation method. In this method, the complete is split into k subsets of approximately equal size. The classification model is then trained and tested "k" times. In this study, to estimate the performance of classifiers a 10-fold cross-validation approach is used. In 10-fold cross validation method, the entire is split into 10 mutually exclusive subsets. We then average the results obtained from the data.

TABLE I
CLASSIFICATION OF VARIOUS ALGORITHMS ON THE DATA SETS.

Mean of Ten-Fold Cross Validation				
Serial No.	***Name of Classifier***	***Accuracy***	***Sensitivity***	***Specificity***
1	Artificial Neural Networks(MLP)	0.5559	0.8621	0.3493
2	Naive Bayes Classifier	0.6878	0.2943	0.8756
3	Support Vector Machines(SVM)	0.7077	0.0000	0.9975
4	K-Nearest Neighbors(KNN)	0.6766	0.3132	0.8197

RESULTS

Classification results

In this study, we have calculated the mean of the previous data and we have replaced the missing values with the mean of the previous records. We have applied various data mining algorithms to the data which we have replaced with mean value and found out that there is a very slight decrease in the mean accuracy of the result when the data mining algorithms are applied to it. There isn't any significant change in the accuracy even though the missing values are replaced by the mean value. Naive Bayes has a slight increase in the accuracy compared to the other classifier models.

In the following paper, the models were evaluated based on the measures which are discussed above (classification sensitivity, specificity and accuracy). The results were achieved using 10 fold cross-validation for each model, and are based on the mean results obtained from the test data set. In this study, we can find out that SVM has the least mean sensitivity of 0, and SVM also produced the highest mean specificity of 0.99755. SVM also produced the highest accuracy compared to other data mining algorithms with a mean accuracy of 0.7077.

While Naive Bayes performed well with the highest mean sensitivity of 0.9594 and some folds reached a high sensitivity of 1. Naive Bayes fared with a mean specificity of 0.3962 which has the least specificity and also the accuracy of naive Bayes is the least among all with a mean accuracy of 0.5559. KNN and ANN fared well with mean accuracies of 0.6878 and 0.6766 respectively. All other results can be seen in the below table For each fold of each model type, the detailed prediction results of the validation data sets are presented in the form of confusion matrices.

The SVM approach provided with better results as the number of records in the experiment was less in number and the data records are noiseless. Support Vector Machine algorithm works better in the case of continuous data instead of discrete data. The data set we choose determines the speed and accuracy in the functionality of the SVM algorithm.The SVM algorithm fails to increase the accuracy of the data sets when there is a higher number of noisy data records. While it works well with noiseless data which unstructured or semi-structured.

Naive Bayes failed to provide better results as the number of records in the experiment is minimal. Naive Bayes provide better results in the huge data sets of a large number of folds. Although it requires short computational time for training, it is less accurate as compared to the other classifiers on some data sets. It has good performance and it is improved by removing irrelevant features. It can deal with noisy data and is a non-deterministic algorithm which works based on the probability theorem.

A confusion matrix is a matrix which represents the classification results. In a two-class prediction problem (such as the one in this paper) the upper left cell denotes the number of samples which classifies as true while they were true (true positives), and lower right cell denotes the number of samples classified as false while they were actually false (true-false). The other two cells (upper right cell and lower left cell) denote the number of samples which are misclassified.

Specifically, the lower left cell denoting the number of samples classified as false while they were actually true (false negatives), and the upper right cell denoting the number of samples classified as true while they actually were false (false positives). Once the confusion matrices were constructed, the performance measures(accuracy, sensitivity and specificity) of each fold were calculated using the respective formulas presented in the previous section.

DISCUSSION

Data collection

One of the key components of predictive accuracy results is the quantity and quality of the collected data [12]. However, the data gathered in medical fields are generally collected as a result of patient-care activity to benefit the patient,

and collection for research purposes is only a secondary consideration. As a result, medical databases contain many features and biased data that create problems for data mining techniques. Medical databases may consist of a large volume of diverse data, including diverse data fields.

The diversity of the data gathered complicates the use of data mining techniques. Additionally, as with any large databases, medical databases contain missing valued data that must be dealt with prior to the use of the data mining techniques. Further, as a result of the method of collection, several databases including medical databases may contain data that is redundant, incomplete, imprecise or inconsistent, which can affect the use and results of the data mining techniques.

Also, the collection method can introduce noisiness into the data and can affect the results of the data mining techniques which are applied to the dataset. All of the above creates several problems for data mining, and as a result, may require more data reduction and data preparation than the data derived from sources [8,11].

Even with these many problems associated with medical databases, the use of medical data for research has many benefits. The research can provide useful information for diagnosis, treatment and early detection of disease. As mentioned above, the results of data mining are directly affected by the quantity and quality of the data.

Medical data governed under the rules of the Common Rule (45 CFR 46) and HIPPA and is subject to the penalties thereunder if the proper procedures are not followed [11].

By improving data collection, data mining can yield even better results and benefits. By making data collection process a main focus, the methods of collecting medical data can be formalized and standardized. Thus, reducing the problem of missing, redundant or inconsistent valued data.

Limitations of data mining

While data mining can provide some insightful information and support to the medical staff by identifying patterns that may not be readily apparent, there are limitations to what data mining can do. Not all patterns found through data mining are fascinating. For a pattern to be fascinating, it should be logical, practical and actionable.

Therefore, data mining requires human involvement to exploit the extracted knowledge. Data mining can only provide assistance in making the diagnosis or prescribing the suitable treatment, but it still cannot replace the physicians ability to understand and analyze the disease.

Most data mining techniques and tools use training and testing datasets drawn from the same sample data set. It can be argued that the testing dataset used is not independent, and therefore, the results have a greater probability of being biased. Also, all data mining results are not accurate and useful, in data mining we can find a lot of patterns in data which are useless.

CONCLUSION

In this paper, we report on an effort where we developed several models for ILPD dataset. Specifically, we used four popular data mining methods (Naive Bayes Classifier, KNN Classifier, Artificial Neural Networks and support vector machines). We acquired a dataset (ILPD dataset) from the UCI machine learning repository and after going through a process of data cleansing and transformation, we used the data set to develop models in order to measure the unbiased prediction accuracy of the three methods, we used a 10-fold cross-validation procedure. That is, we divided the dataset into 10 partitions and we have calculated the average of results obtained (accuracy, sensitivity and specificity).

The results indicated that the support vector machine performed the best with a mean classification accuracy of 70.77%, the ANN model came out to be second best with a mean classification accuracy of 68.96%, the KNN model came out to be third best with a mean classification accuracy of 68.53% and the Naive Bayes model came out to be the worst with a mean classification accuracy of 55.59%. Interestingly, naive Bayes has got the highest sensitivity among all other models with a mean sensitivity of 95.94%.

Although data mining techniques are capable of extracting patterns hidden deep in medical datasets, without the consistent feedback from the medical professional, the results obtained from data mining are useless. The patterns found via data mining methods should be evaluated by medical professionals who have years of experience in the problem domain to decide whether they are logical, actionable, accurate enough and novel to fuel new clinical research directions.

In short, data mining is not aiming to replace medical professionals and medical researchers, but to supplement their efforts to speed up research in this field so diseases can be diagnosed and treated much better.

REFERENCES

[1] 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”.“Journal of Chemical Information and Computer Sciences”,2003,vol. 43,no. 6,pp. 2019–2024.

[2] Paul R. Harper,“A review and comparison of classification algorithms for medical decision making”,“Health Policy,Elsevier”,2005,vol.71,no. 3,pp.315–331.

[3] Lung-Cheng Huang, Sen- Yen Hsu and Eugene Lin,“A comparison of classification methods for predicting Chronic Fatigue Syndrome based on genetic data”,“Journal of Translational Medicine”,2009,vol. 7,no.81.

[4] Bendi Venkata Ramana, Prof. M. Surendra Prasad Babu,Prof. N. B. Venkateswarlu;ILPD (Indian Liver Patient Dataset) Data Set,UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/machine-learning-databases/00225/].

[5] Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:124.

[6] Richards G, Rayward-Smith VJ, Sonksen PH, Carey S., Weng C.;Data mining for indicators of early mortality in a database of clinical records; Artif Intell Med 2001;22:21531.

[7] Haykin S. Neural networks: a comprehensive foundation.New Jersey: Prentice Hall; 1998.

[8] SVM Tutorial: Classification, Regression, and Ranking.Hwanjo Yu and Sungchul Kim.Handbook of Natural computing, 2012.
[9] Rish, Irina (2001). An empirical study of the naive Bayes classifier. IJCAI Workshop on Empirical Methods in AI.
[10] Altman, N. S. (1992).“An introduction to kernel and nearest-neighbor nonparametric regression”. The American Statistician. 46 (3): 175185
[11] Berman JJ. Confidentiality issues for medical data miners.Artif Intell Med 2002;26:2536.
[12] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79:85762.
[13] Wafa’ S.Al-Sharafat, and Reyadh Naoum Development of Genetic-based Machine Learning for Network Intrusion Detection World Academy of Science, Engineering and Technology 55, 2009.
[14] I. Rish. An empirical, study of the naive bayes classifier. Proceedings of IJCAI-01, 2001.
[15] S. Eyheramendy, D. Lewis, and D. Madigan. On the naive bayes model for text categorization. Proceedings Artificial Intelligence Statistics, 2003.