COMPARATIVE ANALYSIS OF CLASSIFICATION TECHNIQUES FOR ACCURACY ON A MULTIVARIATE DATA SET

[1]KalpanaRangra,[2]Dr. K.L.Bansal
[1]Research Scholar, Department of Computer Science , Himachal Pradesh University Shimla,India
[2]Professor,Department of Computer Science, Himachal Pradesh University Shimla, India

*Abstract:-***Current scenario offers huge amount of data and information that are available for everyone either offline or online. Besides being available on the web or even offline, data can be stored in many different kinds of databases and information repositories. With such huge amount of data, there is a need for powerful techniques for better interpretation of these data that exceeds the human's ability for comprehension and making decision in a better way. Though there are a number of techniques and many variations of the methods described, one of the techniques from the mentioned group is almost always used in real world deployments of data mining systems.There are many different methods, which may be used to predict the accuracy for different classes of objects and majority of data mining techniques can deal with different data types. This paper provides an explanation and comparative study on some of the most common data mining classificationtechniques used in day to day life applications for predictions and decision making. In order to reveal the best technique for dealing with the classification task that helps in decision making, this paper has conducted a comparative study between a number of some of the available data mining classification techniques that can aid in differentiating the accuracy and prediction efficiency of one algorithm from another and help in finding the best for respective dataset used.**

Keywords:-data mining, classification accuracy, classification error, class recall, class precision

I. Introduction

Today's databases and data repositories contain so much data and information that it becomes almost impossible to manually analyze them for valuable decision-making. Therefore, humans need assistance in their analysis capacity,humans need data mining and its applications [1].

A)Data mining:

Data mining and manipulation actually draws attentionon the fields of data visualization, computer science, psychology, and information science/information systems. The term data science as a whole intertwines with data- and knowledge-intensive domains such as prediction of health,business,public interests and much more.In coming era it will be impossible to functionally separate data science from the base sciences that it supports.Data mining revolves around certain techniques and approaches to collect ,process and analyze data. Though these methods are not purely science, and often involve more synthesis, deduction, and induction. Working with data requires a solid logical model, an understanding of mathematics, and technical ability. The best data scientists have a background with both information technology and social, biological, or medical science. As the data manipulation and data mining field is so fresh, the fundamental skills are often developed on the job, in practice.

Data mining is an interdisciplinary subfield of computer science involving computational process of large data sets. It is an advanced analysis process that aims to extract information from a data set and transform it into an understandable structure for further use. The techniques use joint methodology of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about solving problems by analyzing data already present in databases [2].

Data mining is the process of discovering interesting knowledge, such as associations, patterns,changes, significant structures and anomalies, from large amounts of data stored in databases or data warehouses or other information repositories [1]. It has been widely used in recent years due to the availability of huge amounts of data in electronic form, and there is a need for turning such data into useful information and knowledge for large applications. These applications are found in fields such as Artificial Intelligence, Machine Learning, Market Analysis, Statistics and Database Systems, Business Management and Decision Support [3].

There are numerous key data mining techniques that have been developing and used in data mining projects lately. These include statistics, association, classification, clustering, prediction, sequential patterns and decision tree. Basically Data Mining techniques have been broken up into two categories based on their evolution. Although these techniques are mentioned in the same breath, they have different analytical approaches. The final component of any data mining algorithms is data management strategy: The ways in which data are stored, indexed and accessed [4].

Data mining offers promising ways to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be explored to predict future behavior. The availability of new data mining algorithms, however, should be met with caution because these techniques are only as good as the data that has been collected. Good data is the first requirement for good data exploration. Assuming good data is available, the next step is to choose the most appropriate technique to mine the data. However, there are tradeoffs to consider when choosing the appropriate data mining technique to be used in a certain application.

There are definite differences in the types of problems that are conductive to each technique. The "best" model is often found by trial and error i.e. trying different technologies and algorithms. Often times, the data analyst should compare or even combine available techniques in order to obtain the best possible results.[12]

B)Classification

Classification is a data mining technique that maps data into predefined groups or classes. It is a supervised learning method which requires labeled training data to generate rules for classifying test data into predetermined groups or classes [2]. It is a two-phase process. The first phase is the learning phase, where the training data is analyzed and classification rules are generated. The next phase is the classification, where test data is classified into classes according to the generated rules. Classification usually comes with a degree of certainty. It might be the probability of the object belonging to the class or it might be some other measure of how closely the object resembles other examples from that class. Classification may indicate a propensity to act in a certain way (predictive) or it may indicate similarity to objects that are definitely members of a given class (definitive).

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is.[8]

Issues Regarding Classification

1) Missing data values

Missing data values cause problems during both the training phase and to the classificationprocess itself. For example, the reason for non-availability of data may be due to [5]:

- Equipment malfunction
- Deletion due to inconsistency with other recorded data
- Non-entry of data due to misunderstanding
- Certain data considered unimportant at the time of entry
- No registration of data or its change

This missing data can be handled using following approaches [6]:

- Data miners can ignore the missing data
- Data miners can replace all missing values with a single global constant
- Data miners can replace a missing value with its feature mean for the given class
- Data miners and domain experts, together, can manually examine samples with missing
- values and enter a reasonable, probable or expected value

2) Measuring Accuracy

Determining which data mining technique is best depends on the interpretation of the problem by users. Usually, the performance of algorithms is examined by evaluating the accuracy of the result. Classification accuracy is calculated by determining the percentage of tuples placed in the correct class. At the same time there may be a cost associated with an incorrect assignment to the wrong class which can be ignored.

Classification algorithms

Following categories of classification techniques has been used in the work.

1)Tree induction

It's a greedy strategy that split the records based on an attribute test that optimizes certain criterion. The basic issue in this technique is to determine how to split the records i.e. to specify the attribute test condition and how to determine the best split. The second important consideration is to determine when to stop splitting.

Decision stump

Decision stump is a machine learning model consisting of a one-level decision tree.It can also be interpreted as a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature.Depending on the type of the input feature, several variations are possible. For nominal features, one may build a stump which contains a leaf for each possible feature value or a stump with the two leaves, one of which corresponds to some chosen category, and the other leaf to all the other categories. For binary features these two schemes are identical. A missing value may be treated as a yet another category. They are oftenused as weak learners or base learner in techniques like bagging and boosting .Sometimes they are also called as one rule.

Decision Tree

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision tree learning is a method commonly used in data mining.The goal is to create a model that predicts the value of a target variable based on several input variables. In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

2) Lazy modeling

Lazy classifiers store all of the training samples and do not build a classifier until a new sample needs to be classified. It differs from eager classifiers, such as decision tree induction, which build a general model (such as a decision tree) before receiving new samples. In lazy classifiers, no general model is built until a new sample needs to be classified. Lazy classifiers are simple and effective. However, it's slow at predicating time since all computation is delayed to that time.

K NN

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.*k*-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

3) Bayesian classification

Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.The Bayesian Classification represents a supervised learning method as well as a statisticalmethod for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem.

Naive Bayes

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesianstatistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class isunrelated to the presence (or absence) of any other feature.Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods.An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

4)Rule Induction

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data.The extractionof useful if-then rules from data based on statistical significance. Rules are the most popular symbolic representation of knowledge derived from data.Rules are natural and easy form of representation possible inspection by human and their interpretation.Rule induction is more comprehensive than any other knowledge representation.Standard form of rules include " IF Conditions THEN Class", "Class IF Conditions; Conditions → Class".

C) Rapid Miner

RapidMiner is an open source data mining tool that provides data mining and machine learningprocedures including data loading and transformation, data preprocessing and visualization,modeling, evaluation, and deployment [7][9]. Rapid Miner uses nested graphs to describe the knowledge flow process.The process can contain loading data, preprocessing, modeling using different types of algorithms, performance measuring, report generating and so on. Knowledge flows consists of Operators where each has a given number of inputs and outputs with type checking. Each Operator also have its attributes which can be set while selecting a given operator

II. Implementation

A)System requirement

System requirement is the basic and foremost platform for any practical approach to be followed for researchand study. The algorithms for classification were implemented on Rapid Miner 5.3, which was configured on Intelcore i3 processor, 2GB RAM,64 bit operating system having Window 7 Home Premium installed in machine.

B) Selection of dataset

The data set was downloaded from KDNuggets[9].It is the latest dataset which is collected from IPEDS survey .This file was created on July 2014, contains directory information for every institution in the 2013 IPEDS universe.It includes name, address, city, state and zip code and Identifies institutions as currently active.The dataset consisted of 250 attributes out of which only 66 were selected for the study purpose and the tool automatically made selection accordingly for classification.The file format is CSV which is readily accepted by Rapid Miner. [10,11]

C) Selection of tool

Along with selection of dataset choice of tool for implementation is yet another important issue for consideration.The current study is implemented on Rapid miner version 5.3, which is an open source tool.It supports more than 40 file formats for datasets and includes more than 100 algorithms for data manipulation.Not only this it incorporates all the features of most handy data mining tool WEKA along with other tools as R,KEEL,Statisticaetc.Though addition of the features makes Rapid Miner a bit complex but since it has a lot many operators, it makes the task quite easier.

D) Selection of technique

While choosing a technique, there is no specific rule that would let you select particular technique over another one. At times those choices are made reasonably arbitrarily based on the accessibility of data mining analysts who are most skilled in one technique over another, but for resolving the purpose of the current work classification was considered as the baseline data manipulation technique.Under broad spectrum of classification few algorithms were selected for comparison and analytical study.

III Results and analysis

Accuracy

Accuracy of classifier refers to ability of classifier predict the class label correctly and the accuracy of predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

Classification error

In terms of machine learning and pattern classification, the data set can be discretely divided into 2 or more classes. Each element of the dataset is called an instance and the class it belongs to is called the label. The Bayes error rate of the dataset classifier is the probability of the classifier to incorrectly classify an instance.

Root mean square error

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. Basically, the RMSD represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of accuracy.

Class Precision

The precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives,(which are items incorrectly labeled as belonging to the class).

Class Recall

Recall is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

Table 1 compares the performances of six classification techniques based on five parameters namely accuracy, classification error, root mean square error(RMSE), class precision, class recall for a multivariate dataset.From the table it can be visualized that among six selected classification algorithms (i.e. decision tree, decision stump. Naïve Bayes, Naïve Bayes kernel,Rule Induction, K-NN),K-NN which is a lazy modeling technique showed maximum accuracy of 100%with 0 error rate followed by Naïve Bayes Kernel which had the prediction accuracy of 99.37%. and second least error rate.Similarly the values of class recall and precision can also be studied and compared from the table .It can also be concluded from the table that the value of RMSE for all algorithms is less than zero still large variation of results can be seen for each technique.The value of class precision is again highest for K-NN.

Table I. Comparative analysis for values of different performance parameters for six selected algorithms.

CATEGORY	Algorithm	Accuracy	Classification Error	Root Mean Square Error	Class Precision	Class recall
Tree Induction	**DECISION STUMP**	**95.92%**	**4.08**	**0.201+/-0.00**	**44.94%**	**49.49%**
	DECISION TREE	**99.14%**	**0.86%**	**0.081+/-0.00**	**98.08%**	**95.70%**
Bayes Classification	**NAÏVE BAYES**	**83.91%**	**16.09%**	**0.40+/-0.00**	**65.85%**	**95.63%**
	NAÏVEBAYES(KERNEL)	**99.37%**	**0.63**	**0.074+/-0.00**	**96.36%**	**99.83%**
Rule Induction	**RULE INDUCTION**	**92.15%**	**7.85%**	**0.280+/-0.00**	**23.03%**	**25%**
Lazy Modeling	**K-NN**	**100%**	**0**	**0.00+/-0.00**	**100%**	**100%**

Figure 1 shows the graphical representation of the results interpreted from table. The graph clearly shows K-NN classification having highest accuracy, class precision, class recall and least classification error and RMSE value. So it can be concluded that lazy modeling technique for classification is best suitable for the selected multivariate data set of universal post-secondary school analysis. There after Bayesian classification ranks second highest in accuracy and precision for classification.

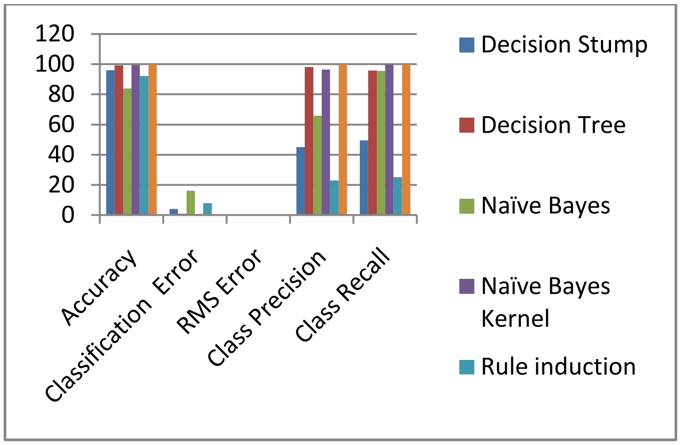

Fig.1

IV. Conclusion

Undoubtedly one of the toughest things to do when deciding to devise a data mining technique is the determination of which technique to use and when. Most of the time the technique to be used are determined by trial and error. Currently, many data mining and knowledge discovery techniques and are available worldwide for different usage This research has conducted a comparison study between a number of available data mining classification techniques depending on their ability for classifying data correctly and accurately. The accuracy measure; which represents the percentage of correctly classified instances, is used for judging the performance of the selected techniques and algorithms.The classification task itself is affected by the type of dataset and the way the classifier was implemented within the toolkit.Finally, K-NN(a Lazy modeling classification technique)algorithm has achieved the highest performance improvements. As a future research, it can be extended to test some other data mining techniques such as clustering, to for more than one dataset of different types.

References

[1] Goebel, M., Gruenwald, L., A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations Newsletter, v.1 n.1, p.20-33, June 1999

[2]. Ian H. Witten and Eibe Frank, Data Mining- Practical Machine Learning Tools and Techniques- second Edition.

[3] Data mining definitions available at http://cplus.about.com/od/glossar1/g/Datamining.htm-

[4] Ming, H., Wenying, N. and Xu, L., (2009) "An improved decision tree classification algorithmbased on ID3 and the application in score analysis", Chinese Control and Decision Conference(CCDC), pp1876-1879

[5] Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.

[6]Kantardzic, M., (2011) Data Mining: Concepts, Models, Methods and Algorithms, Wiley-IEEE Press

[7] Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevier.

[8] FabricioVoznika,LeonardoViana, "Data mining Classification".

[9] RapidMiner, http://rapid-i.com/content/view/181/190/

[10]www.KDnuggets .com

[11] http://nces.ed.gov/ipeds/

[12] https://inventory.data.gov/dataset/

[13]http://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/naive-bayes-classifier.pdf