Journal of Data Mining and Knowledge Discovery
ISSN: 2229-6662 & ISSN: 2229-6670, Volume 3, Issue 3, 2012, pp.-91-95.
Available online at http://www.bioinfo.in/contents.php?id=42

PERFORMANCE OF DIFFERENT CLASSIFIERS ON SCHOOL STUDENTS' DATA IN KARNATAKA

SRIMANI P.K.[1] AND BALAJI K.[2]*
[1]R&D Division, B.U., DSI, Bangalore-560078, Karnataka, India.
[2]Department of MCA, Surana College PG Centre, #17, Kengeri ST, Bangalore-560060, Karnataka, India.
*Corresponding Author: Email- manobalaji@gmail.com

Received: October 25, 2012; Accepted: November 06, 2012

Abstract- This paper deals with the study of the performance of different classifiers on edu. data in particular the school students' data belonging to different zones of Karnataka. Mining educational data is an emerging interdisciplinary research area that mainly facilitates to take effective decisions with regard to the development of the cities under consideration. The data comprises 610 instances and 16 attributes and the results predict that Decision Stump is the best classifier with accuracy 97.918.

Keywords- Classifiers, Students' data, Data mining, Accuracy prediction, Confusion matrix

Citation: Srimani P.K. and Balaji K. (2012) Performance of Different Classifiers on School Students' Data in Karnataka. Journal of Data Mining and Knowledge Discovery, ISSN: 2229-6662 & ISSN: 2229-6670, Volume 3, Issue 3, pp.-91-95.

Introduction

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information – information which can be utilized for increasing the revenue, reducing costs, or both. Many analytical tools for analysing data are available and data mining software is one such tool which facilitates the users to effectively analyse the data from different dimensions or angles, categorize it, identify and summarize the relationships therein, Technically data mining can be referred to as the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining is the data analysis component of Knowledge Discovery in Databases (KDD). In fact KDD encompasses all steps from the collection and management of data through to data analysis. Frequent themes are analysis (both exploratory and formal), methods capable of handling the computations and automation for large data sets.

Data mining basically involves huge volumes of data being sifted through by data mining software for searching patterns in the data and no doubt Data mining, is a type of artificial intelligence primarily used for analyzing scientific and business data. In recent years Data Mining's application is found in voting trends and patterns in political campaigns.

In decision making, these patterns play a vital role as they identify the areas which can be improved in the process. For the improvement of profitableness and effectiveness associated with (i) the interactions with the customers (ii) the management of risk and (iii) detection and fraud, business organizations employ Data Mining techniques. In other words, the discovery of patterns through data mining assist business organizations makes timelier and better decisions.

Knowledge Discovery Database:

Knowledge Discovery in Databases (KDD) is the automated discovery of patterns and relationships in large databases.

Knowledge discovery database involves different fields such as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems, and data visualization. The KDD Process is user involved, highly iterative multistep process, which can be seen in [Fig-1].

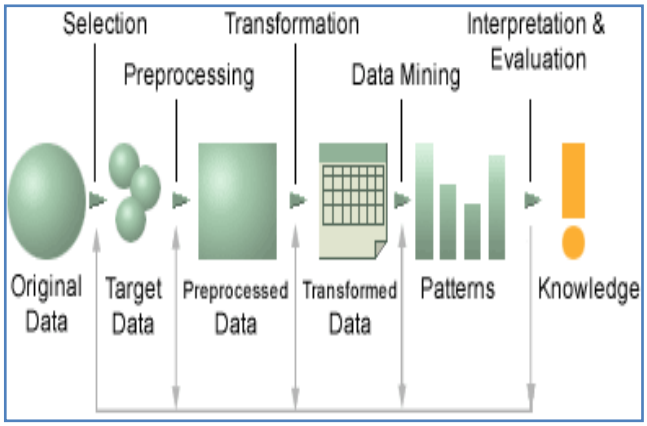

Fig. 1-The KDD Process

[Fig-1] reveals that (i) the Knowledge Discovery Database is a process that is user involved and highly iterative and (ii) the initial data is the organizational data which is collected from several locations and put in some central locations which are normally referred to as Data Warehouses or Data Marts.

In order to resolve the inconsistencies and the discrepancies associated with the data of different locations, data transformation has to be performed on the raw data prior to its placement in the data warehouse. Here the discrepancies or the inconstancies are due to the different data types and different field names for the same information and data fields. Data transformation resolves inconsistencies between the data of one location from that of another. For example, inconsistencies may be differences in data type for the same information and different field names for the same data field.

Data warehouses hold both detailed and summary data. Detailed data is used for pattern analysis, where summarized data may hold the results of previous analyses. Data warehouses also contain historical data whereas operational data is usually current. Efficient organization of the data warehouse is essential for efficient data mining.

Once the data is organized, a selection process occurs where some subset of this data becomes the *target data* upon which further analysis is performed. While creating the target data analyst should keep it in his mind the following:

- The domain of interest
- The end user's needs and the underlying data mining tasks.

Data entry mistakes can occur and/or the data may have missing or unknown entries. During the data cleaning and preprocessing stage noise is removed from the data. The Outliers and anomalies which are present in the data pose special problems for the data analyst during the data cleaning process.

Our goal is to find the outliers and anomalies which are the representations of these rare patterns in the data and hence they should not be removed. This step in the process can be the most time consuming. Data Reduction and Coding step employs transformation techniques that are used to reduce the number of variables in the data by finding useful features with which to represent the data.

The KDD process constitutes data mining as a step in which the transformed data is used and a search for this actual patterns is performed. The search for patterns the analysis are performed under the data mining task and representation model. At this stage, it is absolutely necessary the suitable data mining algorithm (for example linear/logistic regression, neural networks, association rules, etc.) for the data mining task viz., classification, database segmentation, rule generation, etc.

As discussed earlier, Data mining tasks comprised: classification, linear regression analysis, rule formation, or cluster analysis. Patterns generated in the data mining step may not be new or interesting. It is therefore necessary to remove redundant and irrelevant patterns from the set of useful patterns. The end user has to be informed wherever 'good patterns' are discovered and this can be done either through textual reports or visualizations in the form of graphs, spreadsheets, diagrams and so on.

The interpretation step takes the reported results and interprets this into *knowledge.* Further it may require restoring possible conflicts with previously discovered knowledge since new knowledge may even be in conflict with knowledge that was believed before the process began. When this is done to user's satisfaction, the knowledge is documented knowledge will be reported to the interesting parties which may involve visualization.

It is important to stress that the KDD process is not linear. In the sense that the results obtained in one phase can be fed into the same or different phase. Current KDD systems involve a highly interactive human component at several steps. Hence, the KDD process is highly interactive and iterative.

Data Mining Application Examples

The areas where data mining has been applied recently include

Science (astronomy, bioinformatics, drug discovery), Business (advertising, Customer modelling and Relationship management),

- E-Commerce,
- Fraud detection,
- Health care,
- Investments
- Manufacturing,
- Sports/ entertainment,
- Telecom (telephone and communications),
- Targeted marketing),
- Web (search engines),
- Government (anti-terrorism efforts, law enforcement, profiling tax cheaters) and
- Education etc.

In this paper, the performance of different classifiers on a real – time students' data with 610 instances and 19 attributes (Karnataka Education Department) is studied in detail. No work in this direction is available.

Section II deals with Data Mining Techniques. Section III deals with the Methodology. Section IV deals with Data Set description. Section V deals with Data experiments and results and finally conclusions are presented in Section VI.

Related Work

No doubt most of the related works have used Data Mining Techniques but their objectives and analysis are different. Some of the recent works are from [2-11].

Data Mining Techniques

Here is a brief account of two of the most popular data mining techniques: Regression and Classification.

Regression

This is the most widely known and the oldest statistical technique that is utilized by the data mining community and essentially makes use of a dataset to develop a mathematical formula which fits the data. So whenever we want to use the results for predicting the

future behavioural patterns, all you need to do is just take the new data, and apply it to the formula that has been developed, and you will get your prediction.

The greatest limiting factor of this technique is that it works well with only quantitative data that is continuous, such as age, speed, or weight. But if you need to work with data that is categorical, where there is no significant order, such as gender, name, or color, it is better to use a different technique.

Classification

If we need to work with categorical data, or a combination of categorical data and continuous numeric, classification analysis will meet your requirements. This technique has the capability to process a more extensive variety of data compared to regression and is therefore increasing in popularity.

In addition, the output it provides can be interpreted more easily. Rather than the complex mathematical formula that the regression technique provides, in this you will be provided a decision tree which requires a sequence of binary decisions.

Classification, Clustering and Nearest Neighbour

Before we get into the specific details of each method and run them through WEKA, One should understand the type of data and goals that each model strives to accomplish. In the present work we compare three new regression models by considering a real time example.

All the real-world examples all revolve around a local BMW dealership in order to increase sales and how it can increase sales. The dealership that is interested in increasing the future sales has all the past information stored and employs data mining to accomplish the object.

Classification

Supposed we are interested to know how likely is person X to buy the newest BMW M5? By creating a classification tree (a decision tree), one can determine the likelihood of this person to buying a new M5. The possible nodes of the tree would be age, income level, and current number of cars, marital status, kids, homeowner, or renter. Thus attributes of this person can be used against the decision tree to determine the likelihood of his purchasing the M5.

Clustering

Suppose we are interested to know the age groups which like the silver BMW M5. The data can be mined to compare the ages of the purchasers of the cars and their colors. From this data, it could be found that the age groups(22-30 years, for example) has a higher propensity to order a certain color of BMW M5s. Similarly, it can be shown that a different age group (55-62, for example) orders silver BMWs. During the process of data mining, user is able to determine the patterns by studying the clusters formed around different colors.

Nearest Neighbour

Suppose we are interested to knowing the time of purchase of BMW M5 and the available options for buying at the same time. Normally purchasers buy the matching luggage also. Using this data, the sales can be increased and the study comes under market basket analysis.

Methodology

The present work employs the following classifiers for the data set considered.

In this work we use the WEKA DT tools, which include 7 different and independent algorithms for constructing decision trees. In the following we give a brief description of each algorithm. WEKA is a Java Software package for data mining tasks developed by the University of Waikato, New Zealand.

J48- J48 is the WEKA implementation of the C4.5 algorithm (Quinlam, 1993) [12]. Given a data set, it generates a DT by recursive partitioning of the data. Depth – first strategy is used to grow the tree and the algorithm computes the information. This process is repeated for each new node until a leaf node has been reached.

J48graft- J48graft generates a grafted DT from a J48 tree [13]. The grafting technique (Webb, 1999) adds nodes to an existing DT with the purpose of reducing pre- diction errors.

This algorithm identifies regions of the multidimensional space of attributes not occupied by the training examples, or by the misclassified training examples, and takes into consideration the alternative branches for the leaf containing the region in question. Aa new test will be performed which generates new branches and leads to new classifications.

BFTree- BFTree (Best-First decision Tree; Haijian Shi, 2007) has a construction process similar to C4.5. TO build up a node, C4.5 uses a fixed order while BFTree employs the best-first order. The building of nodes is done normally from left to right and leads to the longest possible path (a path is the way from the root node to a leaf).

FT- FT (Functional Trees; Gama, 2004) combines a standard univariate DT, say C4.5, with linear functions of the attributes by means of linear regressions. However, a univariate DT uses in a node only simple value tests on single attributes, In a node leaf, DT can use linear combination of different attributes.

REPTree- REPTree is no doubt a fast decision tree learner that builds a decision/regression tree by using information gain/variance as the criterion to select the attribute to be tested in a node.

Decision Stump- Decision Stump is a simple binary DT classifier consisting of a single node (based on one attribute) and two leaves. The attributes used by the other trees are tested and the one giving the best classifications is chosen to use in the single node.

Data Set Description

The data set used in the present investigation for our classification example will focus on the students in various regions in Karnataka. The description is as follows:

- % Data on schools in Karnataka districts from the Department of Education
- % no of instances = 610
- % no of attributes = 16

- % Variable definitions
- % number of students in district = school(1:nobs,1);
- % index of concentration = school(1:nobs,2);
- % expenditures per pupil = school(1:nobs,3);
- % class size = school(1:nobs,4);
- % passed on all 10th grade proficiency tests = school (1:nobs,5);
- % big city = school(1:nobs,6);
- % small city = school(1:nobs,7);
- % suburban = school(1:nobs,8);
- % southeast region = school(1:nobs,9);
- % central region = school(1:nobs,10);
- % northcentral region = school(1:nobs,11);
- % northeast region = school(1:nobs,12);
- % northwest region = school(1:nobs,13);
- % southcentral region = school(1:nobs,14);
- % southwest region = school(1:nobs,15);
- % dropout rate = school(1:nobs,16);

Experiments and Results

This section contains the results of the experiments conducted on the students data sets mentioned earlier. This data set consists of 610 instances with 16 attributes. As mentioned earlier, the main objective of the present analysis is to predict the test classifier that will enable to take effective decisions.

[Table-1], presents the number of accurately and inaccurately classified instance, mean absolute error, Kappa statistics, the time to build up the model and the accuracy with regard to the classifiers viz; J48, J48Graft, BFTree, Decision Stump, Bayes Net, Naive Bayes, Naive Bayes updateable. It is found that

- Decision Stump is the best classifier with accuracy 97.918.
- Bayes Net, Naive Bayes and Naive Bayes Updateable are the next best classifiers with accuracy (=94.7541) and
- The other classifiers performed equally well.

Thus, it can be concluded that the classifiers considered here are most suited and accuracies are also remarkable. These results are also in accordance with the values of the Kappa statistics and mean absolute error.

[Table-2] and [Table-3] present the confusion matrices for the seven classifies mentioned above. The correctly and incorrectly classified instances for each classifier are available from these tables and are summarized in [Table-1]. A glance at these matrices yields some useful information with regard to the analysis.

Table 1- Accuracy prediction for different classifiers

Classifier						
Algorithms	Correctly Classified Instances	In Correctly Classified Instances	Mean absolute Error	Kapa Statistics	Time to build (in Sec)	Accuracy
Decision Stump	579	31	0.0183	0.2294	0.02	97.918
Bayes Net	578	32	0.0123	0.2045	0.03	94.7541
Naïve Bayes	578	32	0.0113	0.1862	0.02	94.7541
NaiveBayes Updateable	578	32	0.0113	0.1862	0	94.7541
J48 Graft	576	34	0.017	0.0714	0.05	94.4262
BFTree	576	34	0.0175	0.1733	1	94.4262
J48	575	35	0.0174	0.2373	0.01	94.2623

Table 2- Confusion Matrix for Decision Stump

a	b	c	d	e	f	g	h	i	j	<-- classified as
576	1	0	0	0	0	0	0	0	0	a = '(-inf-7310.6]'
24	2	0	0	0	0	0	0	0	0	b = '(7310.6-14616.2]'
1	0	0	0	0	0	0	0	0	0	c = '(14616.2-21921.8]'
0	1	0	0	0	0	0	0	0	0	d = '(21921.8-29227.4]'
1	0	0	0	0	0	0	0	0	0	e = '(29227.4-36533]'
0	1	0	0	0	0	0	0	0	0	f = '(36533-43838.6]'
1	0	0	0	0	0	0	0	0	0	g = '(43838.6-51144.2]'
0	0	0	0	0	0	0	0	0	0	h = '(51144.2-58449.8]'
0	1	0	0	0	0	0	0	0	0	i = '(58449.8-65755.4]'
0	1	0	0	0	0	0	0	0	0	j = '(65755.4-inf)'

Conclusion

In this paper, experiments on students', data sets are (Karnataka) are conducted to study the behaviour of the different WEKA classifiers. The data set comprises 610 instances wit 16 attributes. The raw data is obtained from the education department and the results obtained are discussed in detail. These results provide an excellent platform for making effective decisions. Further, of all the classifiers Decision Stump is found to be an excellent performance with accuracy 97.918. No work in this direction is available although some literature exists with regard to the educational data.

Table 3- Confusion Matrix for NaiveBayesUpdateable

a	b	c	d	e	f	g	h	i	j	<-- classified as
576	1	0	0	0	0	0	0	0	0	a = '(-inf-7310.6]'
24	2	0	0	0	0	0	0	0	0	b = '(7310.6-14616.2]'
1	0	0	0	0	0	0	0	0	0	c = '(14616.2-21921.8]'
0	1	0	0	0	0	0	0	0	0	d = '(21921.8-29227.4]'
1	0	0	0	0	0	0	0	0	0	e = '(29227.4-36533]'
0	1	0	0	0	0	0	0	0	0	f = '(36533-43838.6]'
1	0	0	0	0	0	0	0	0	0	g = '(43838.6-51144.2]'
0	0	0	0	0	0	0	0	0	0	h = '(51144.2-58449.8]'
0	1	0	0	0	0	0	0	0	0	i = '(58449.8-65755.4]'
0	1	0	0	0	0	0	0	0	0	j = '(65755.4-inf)'

Acknowledgement

One of the authors Mr. Balaji K acknowledges Rayalaseema University, Kurnool, Andra Pradesh, India and Surana College PG Centre, Bangalore, Karnataka for providing the facilities for carrying out the research work.

References

[1] Susan P. Imberman (2009) *Effective use of the KDD Process and Data Mining for Computer Performance Professionals*, Thesis, College of Staten Island, New York.

[2] Srimani P.K. and Malini Patil (2011) 2*nd International Conference Methods and Models in Science & Technology*.

[3] Srimani P.K. and Malini Patil (2012) *International Conference on Intelligent Computational Methods*, Dubai.

[4] Srimani P.K., Srivathsa P.K. and Malini Patil (2012) *International Journal of Current Research*, 4(2), 183-190.

[5] Gowri Vijayakumar (2007) *Pupil- Teacher Ratio & the Accelerated programme for reading in Karnataka Schools*, Akshara Foundation.

[6] Gnardellis T. and Boutsinas B. (2001) *On Experimenting with Data Mining in Education*.

[7] Agathe Merceron and Kalina Yacef (2005) *Educational Data Mining: a Case Study*, ESILV- Pole Universitaire Leonard de Vinci, France.

[8] Govindaraju R. and Venkatesan S. (2010) *J. Psychology*, 1(1), 47-53.

[9] Brijesh Kumar Baradwaj (2012) *International Journal of Advanced Computer Science and Applications*, 2.

[10] Senaol Erdogan Mehpare Timor (2009) *Journal of Computing*, 1(1).

[11] Ramaswami M. and Bhaskaran R. (2009) *Journal of Computing*, 1(1), 7-11.

[12] Anil Rajput (2011) *International Journal of Compter Science & Security*, 5(2).

[13] Vasconcellos E.C., DeCarvalho R.R., Capelato H.V., Campos Velho H.F., Trevisan M. and Ruiz R.S.S. (2010) *Cosmology and Extragalactic Astrophysics*.