Performance Analysis and Comparison of Various Data Mining Techniques

Brahamjit Pannu
Computer Science & Engineering, Satpriya Group of Institution, M.D.University Rohtak, Haryana, India

Puneet Sharma
Assistant Professor, Computer Science & Engineering, Satpriya Group of Institution, M.D.University Rohtak, Haryana, India

Abstract – Data Mining is a powerful tool for academic intervention. Mining in education environment is called Educational Data Mining. Educational Data Mining is concerned with developing new methods to discover knowledge from educational database and can used for decision making in educational system. The data mining software applications involves various methodologies that have been developed by both commercial and research centers. The techniques have been used for industrial, commercial and scientific purpose. For example data mining has been used to analyze large datasets and establish useful classification and patterns in the datasets. This paper outlines observe which may establish if new data mining techniques will improve the effectiveness and accuracy of the classification and Clustering of Education dataset. The main objectives of our work are to investigate the performance analysis of different classification methods and Clustering Methods using the WEKA software for the Education dataset.

Index Terms – Data Mining, Discover knowledge, Educational Data Mining, Classification, Comparative Data mining, Classification Method, Weeka Tool.

1. INTRODUCTION

Data Mining is a process of extracting previously unknown, valid, potentional useful and hidden patterns from large data sets (Connolly, 1999). As the amount of data stored in educational databases is increasing rapidly. In order to get required benefits from such large data and to find hidden relationships between variables using different data mining techniques developed and used (Han and Kamber, 2006). Data Mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data come from educational environments. Application of data mining in education sector is an emerging trend [1,2] . The data mining terms, tasks, techniques and application can be used to developing data mining in education sector. Clustering and classification both are very useful to improve the performance on education sector.

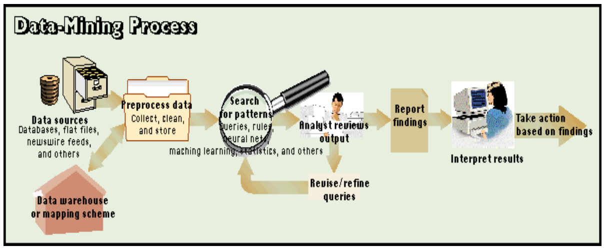

Data-Mining Process
Data sources
Databases, flat files, newswire feeds, and others
Preprocess data
Collect, clean, and store
Search for patterns
Queries, rules, neural nets,
maching learning, statistics, and others
Analyst reviews output
Revise/refine queries
Report findings
Interpret results
Take action based on findings
Data warehouse or mapping scheme

Fig 1.1 Process of Data Mining

2. DATA MINING TECHNIQUES

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or “mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery in database are frequently treated as synonyms. Data mining methods like prediction, clustering and relationship mining are mostly used in the field of marketing, agriculture and finance etc . These methods can be efficiently applied on educational data.

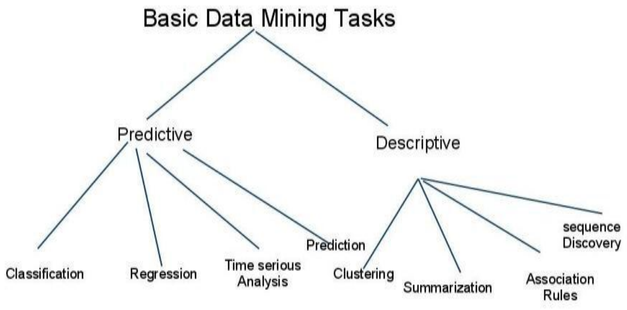


Fig 2.1 Data Mining Technique

3. CLASSIFICATION

Classification is the most commonly applied data mining technique, which employs a set of pre-classified attributes to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification .In learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified attributes to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier [3][4].

The Classification methods used for the comparative study are:
o Classification by decision tree induction
o Bayesian Classification
o Neural Networks
o Support Vector Machines (SVM)
o Classification Based on Associations

4. EDUCATIONAL DATA MINING

Education is an essential element for the betterment and progress of a country. It enables the people of a country civilized and wellmannered. Educational Data Mining is an emerging discipline concerned with developing methods for exploring the unique types of data that come from educational database. Mining in educational environment is called Educational Data Mining, concern with developing new methods to discover knowledge from educational databases (Galit, 2007) (Erdogan and Timor 2005) ,in order to analyze students trends and b ehaviors toward education(Alaa el-Halees , 2009). Lack of deep and enough knowledge in higher educational system may prevent system management to achieve quality objectives, data mining methodology can help bridging this knowledge gaps in higher eduction syatem.

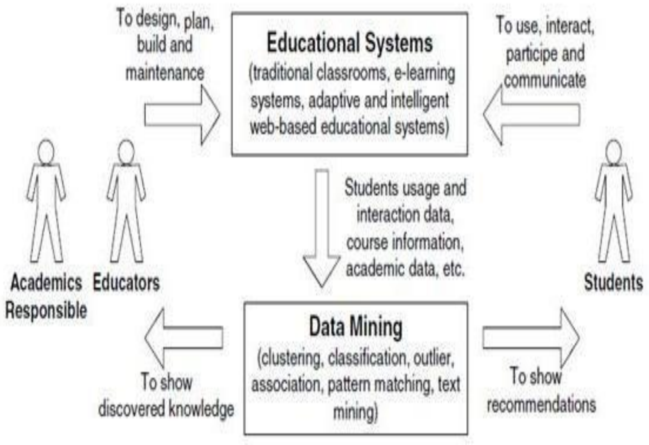

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Academics Responsible
Educators
Students usage and interaction data, course information, academic data, etc.
Students
To show discovered knowledge
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show recommendations

Fig 4.1 Data mining applications in the education sector

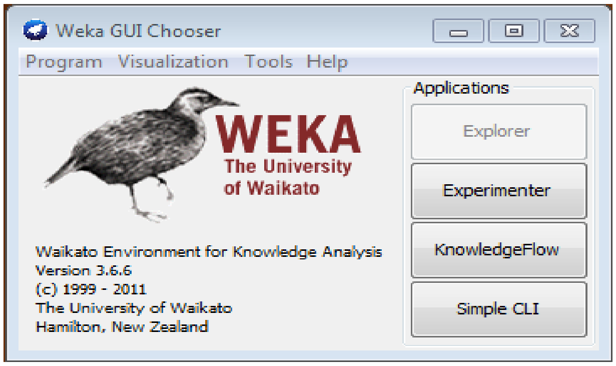
5. METHODOLOGY

To compare and analysis of data mining techniques we used weka tool. It is open sources gui tool. firslty we open the weka tool than select the data set .in second step we choose data mining techniques than any one method of that techniques .perform the method and check the result . Similarity run other method and check result than compare these result and find out better one of them.

5.1. Classification on WEKA

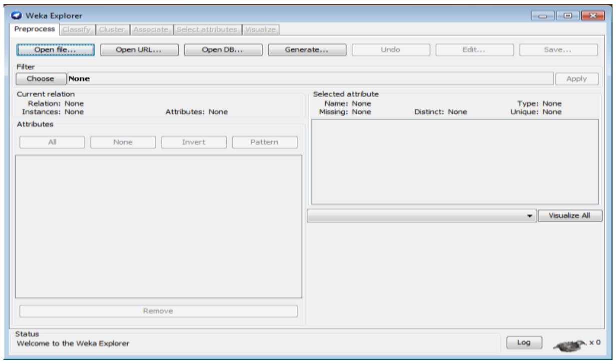
Working model of the proposed technique has following phases:

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at theUniversity of Waikato, New Zealand. It provides an interface where a dataset can be selected and a data mining technique can be applied on the dataset.The Weka GUI Chooser (class weka.gui.GUIChooser) provides a starting point for launching Weka's main GUI applications and supporting tools. If one prefers a MDI ("multiple document interface") appearance, then this is provided by an alternative launcher called "Main" (class weka.gui.Main). The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and four menus.



Waikato Environment for Knowledge Analysis
Version 3.6.6
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand

Fig 5.1.1 Weka user Interface



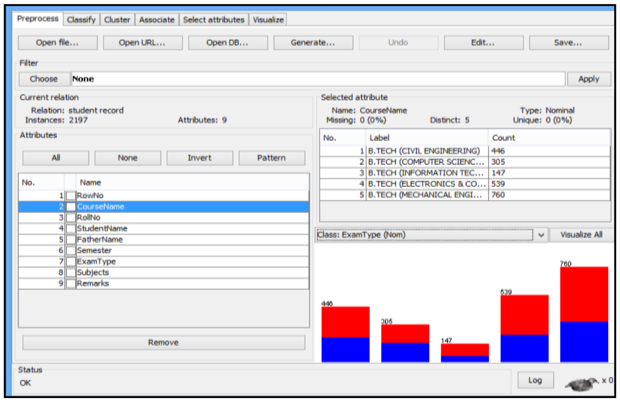
Filter: None

Relation: None
Instances: None
Attributes: None

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

Welcome to the Weka Explorer

Fig 5.1.2 Weka Explorer



Filter: None

Current relation
Relation: student record
Instances: 2197 Attributes: 9

Attributes

No.	Name
1	RowNo
2	CourseName
3	RollNo
4	StudentName
5	FatherName
6	Semester
7	ExamType
8	Subjects
9	Remarks

Selected attribute
Name: CourseName Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

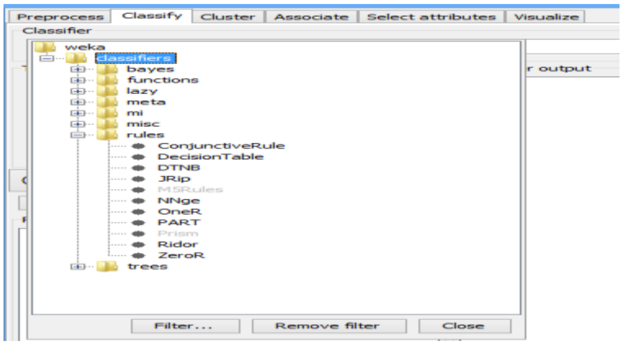
No.	Label	Count
1	B.TECH (CIVIL ENGINEERING)	446
2	B.TECH (COMPUTER SCIENC...	305
3	B.TECH (INFORMATION TEC...	147
4	B.TECH (ELECTRONICS & CO...	539
5	B.TECH (MECHANICAL ENGI...	760

Class: ExamType (Nom)

Status
OK

Fig 5.1.3 Weka selecting Dataset

We have many options shown in the figure5.3. We perform classification so we click on the classify button. After that we choose an algorithm which is applied to the data. It is shown in the figure 5.4. And the click ok button.



F ig 5.1.4 various Classification algorithms in weka

Results for classification techniques applied on Education Dataset on WEKA.

5.2. Bayesian Networks

A Bayesian Network (BN) is a graphical model. Which is used to provide relationships among a set of variable features. [5]. Fig 5.2.1 show the structure of Bayesian net.

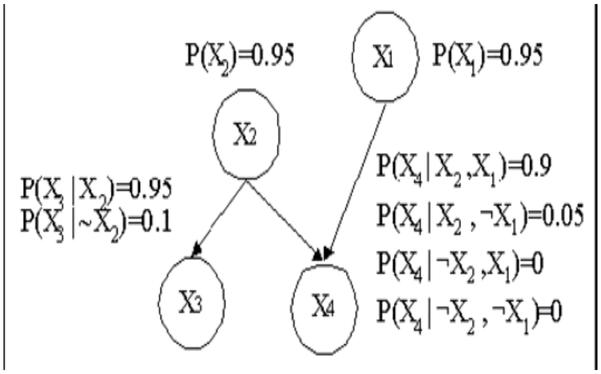


Fig 5.2.1 Basic structure of Bayes net

1.Accuracy: The measure of the Accuracy of the education dataset for BayesNet classifier technique is shown below with graph according to the Table No. 5.1.From the below Table No.5.1 shows the training size (%), total no. instances, correctly classified instances and Kappa Statistic and Fig 5.2.2. show the performance of Accuracy on education dataset. From the below Figure No.6.2 we can clearly see that the highest accuracy is 28.9773 % and lowest is 9.2593 % when the training size is 40% and 80% respectively. The accuracy decreases from 28% to 9%. The accuracy sometimes increases and sometimes decreases on the different types of training size. Here we can clearly see that the accuracy is increased when dataset is small split and when dataset is large split the accuracy is minimized.

Trainig Size (%)	Total no.of Instances (624)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Mean Absolute Error	Kappa Statistic
40	176	28.9773%	71.0227%	0.0344	-0.0187
50	147	23.1293%	76.8707%	0.4986	0.0273
60	118	16.1017%	83.8983%	0.0358	0.0237
70	82	17.0732 %	82.9268 %	0.1371	0.0103
80	54	9.2593 %	90.7407 %	0.0374	-0.0023

Table 1 result of BayesNet(Training)

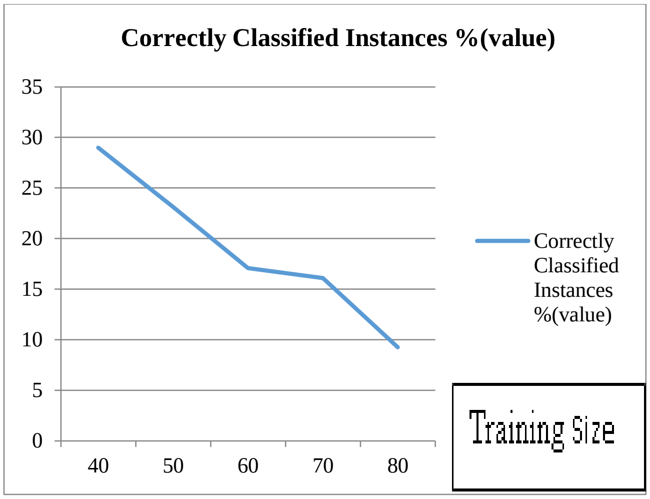


Fig 5.2.2 performance Accuracy on education data set.

2. Kappa Statistic: The measure of the Kappa Statistic of the Education dataset for Bayes Net classifier techniques is shown below with graph according to the Table 1.

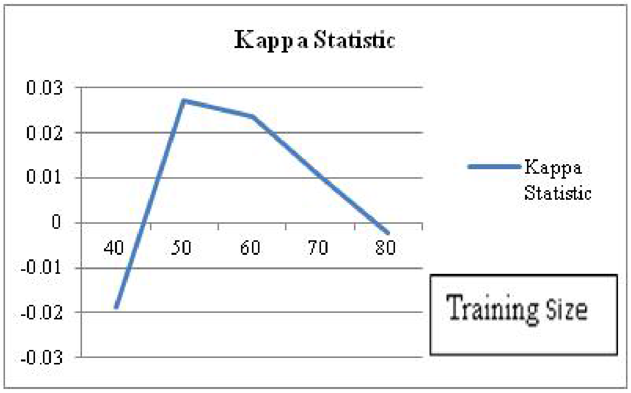


Fig 5.2.3 Show the Kappa Statistics**.**

From the Table 1 and Fig 5.2.3 show the performance of Kappa Statistic on education dataset. We can see value of Kappa Statistic is increase -0.0187to 0.0273.

3. Mean Absolute Error: The measure of the Mean Absolute Error of the mushroom dataset for Bayes Netclassifier techniques is shown below with graph according to the Table 5.1

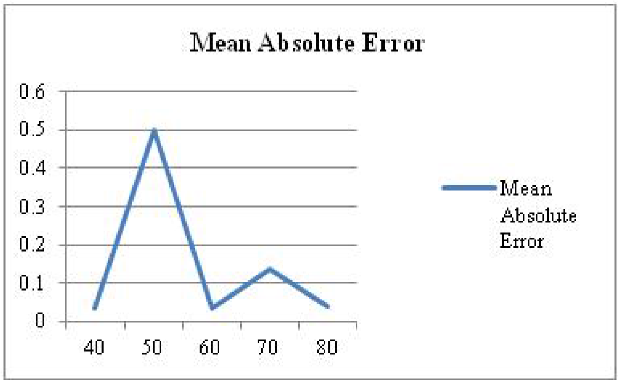


Figure 5.2.4 Show the Mean absolute Error

From the Table 1 and Fig 5.2.4 show the performance of the Mean Absolute Error on education dataset.

5.3. Naïve Bayes

Naives is a simple form of Bayes net which is represented DAG with one parent and many children. Its use with a strong assumption of independence among child nodes in the context of their parent. [6] . Table 5.2 shows the resultant measure the performance of the Naïve Bayes classifier techniques on the education dataset.

S. No.	Training Size (%)	Total no. of Instances (624)	Correctly Classified Instances % (value)	Incorrectly Classified Instances % (value)	Mean Absolute Error	Kappa Statistic
1	40	176	65.3409%	34.6591 %	0.0312	0.0993
2	50	147	63.2653%	36.7347 %	0.0315	0.0903
3	60	118	51.6949%	51.6949 %	0.0324	0.0469
4	70	82	39.0244%	60.9756 %	0.0332	0.0061
5	80	54	31.4815%	68.5185 %	0.0349	-0.0544

Table 2 result of algorithm Naive Bayes (Training)

Similarly, like Bayes net we can measure the accuracy, Kappa Statistic, mean absolute error in educate data set for Naïve Bayes.

5.4. Decision Trees (DT'S)

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Selection of a certain branch depends upon the outcome of the test [7]

Table 3 shows the resultant measure the performance of the classifier techniques on the education dataset which have the total no. of instances is 624 and attributes are 16.

S. No.	Trainin g Size (%)	Total no. of Instances (624)	Correctly Classified Instances %(value)	Incorrectl y Classified Instances %(value)	Mean Absolut e Error	Kappa Statistic
1	40	176	79.5455%	20.4545 %	0.014	0
2	50	147	78.9116%	21.0884 %	0.0141	0
3	60	118	80.5085%	19.4915 %	0.0141	0
4	70	82	80.4878%	19.5122 %	0.014	0
5	80	54	77.7778%	22.2222 %	0.0146	0

Table 3 Simulation result of algorithm decision

We can measure the accuracy. Kappa statistic, mean absolute error in education data for decision tree. The accuracy sometimes increases and sometimes decreases on the different types of training size.

5.5. Comparison of Different Classification Techniques

We compare different-2 techniques of classification in term of there accuracy, kappa statistics and mean square error.

5.5.1. Comparison for accuracy

S. No.	Training Size (%)	BayesNet (%)	NaiveBayes (%)	Decision tree(%)
1	40	28.9773 %	65.3409 %	79.5455 %
2	50	23.1293 %	63.2653 %	78.9116 %
3	60	16.1017 %	51.6949 %	80.5085 %
4	70	17.0732 %	39.0244 %	80.4878 %
5	80	9.2593 %	31.4815 %	77.7778 %

Table 4 Comparative result of classification techniques

Now we are describe the experimental results which is obtained from the various classification techniques and comparison with each other. The best techniques identified from each classifier then compared with other classifiers to discover what classifier is best to be used for classification of education dataset. We used here these techniques for the comparison on the education dataset and find the best techniques.

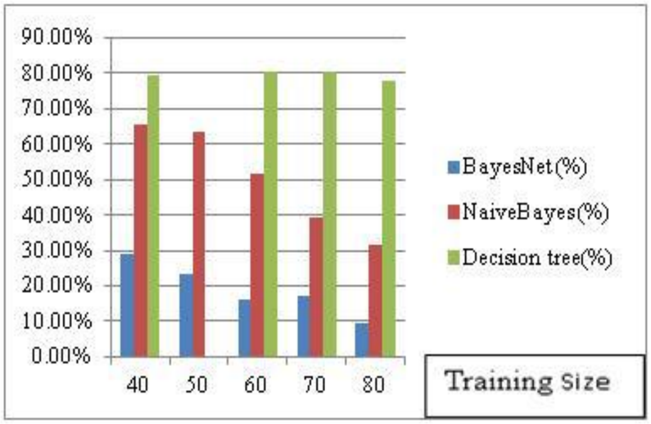


Fig 5.5.1 Comparison between parameters for Accuracy

Based on the above Figure and Table we can clearly see that the highest accuracy is 28.9773 %and lowest accuracy is 9.2593 %in the BayesNet classifiers. And the highest accuracy is 65.3409 %and lowest accuracy is 31.4815 %%in the NaiveBayes classifiers and the highest accuracy is 80.5085 % and lowest accuracy is 77.7778 %in the Decision tree classifiers. From the above graph we can clearly see that the accuracy rate of Decision tree classifier is the best among these three classifier techniques.

5.5.2. Comparison for Mean absolute Error

Based on the above Figure 5.5.2 and Table 5 we can clearly see that the highest Mean absolute error is 0.4986and lowest Mean absolute error is 0.0344% in the BayesNet classifiers. And the highest Mean absolute error is 0.0349and lowest Mean absolute error is 0.z0312in the NaiveBayes classifiers and the highest Mean absolute error is 0.0146% and lowest Mean absolute error is 0.014% in the Decision tree classifiers. From the above graph we can clearly see that the Mean absolute error rate of Bayes net classifier is the best among these three classifier techniques.

S. No.	Training Size (%)	BayesNet (%)	NaiveBayes (%)	Decision tree(%)
1	40	0.0344	0.0312	0.0146
2	50	0.4986	0.0315	0.0141
3	60	0.0358	0.0324	0.0141
4	70	0.1371	0.0332	0.014
5	80	0.0374	0.0349	0.0146

Table 5Comparison of classification techniques

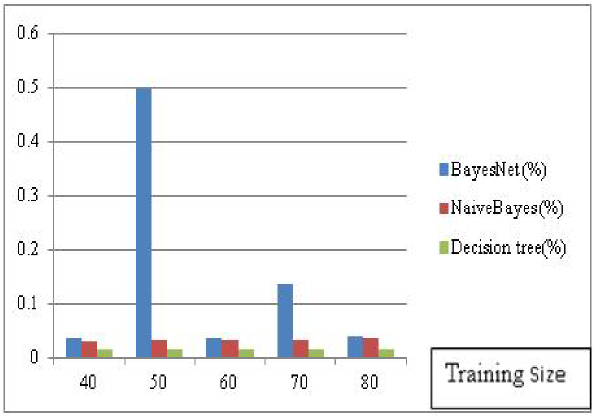


Figure 5.5.2 Comparison between parameters for Mean Absolute Error

5.5.3. Comparison for Kappa Statistic

S. No.	Training Size (%)	BayesNet (%)	NaiveBayes (%)	Decision tree(%)
1	40	-0.0187	0.0993	0
2	50	0.0273	0.0903	0
3	60	0.0237	0.0469	0
4	70	0.0103	0.0061	0
5	80	-0.0023	-0.0544	0

Table 6 Comparative result of classification techniques

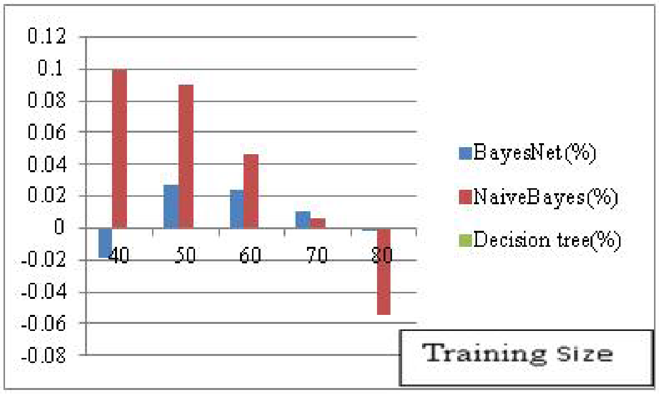


Figure 5.5.3 Comparison between parameters for Kappa Statistic

Based on the above Figure 5.5.3 and Table 6 we can clearly see that the highest Kappa Statistic is 0.0273and lowest Kappa Statistic is -0.0023in the BayesNet classifiers. And the highest Kappa Statistic is 0.0993and lowest Kappa Statistic is -0.0544in the NaiveBayes classifiers and the Kappa Statistic is zero in the Decision tree classifiers for all training size. From the above graph we can clearly see that the Kappa Statistic rate of Naïve net classifier is the best among these three classifier techniques.

6. CONCLUSION

Educations data sets are selected to study the relationship between the training dataset size and the accuracy, error rate. These datasets are used for training, three Bayesnet algorithms: NaiveBayes algorithm and ZeroR algorithm are used. Error rates are obtained for different split ratios. Furthermore, using these techniques is compared to see if there is any significant difference between the performances of the algorithms on the datasets. After analyzing the results of testing the algorithms we can say that every techniques perform best result according totheir parameters means if we take accuracy than decision tree is best but if we use mean absolute error than bayes netis better that other algorithms but if we take kappa statistics than naïve net perform better result.. Bayes network classifier has the potential to significantly improve the conventional classification methods for use in general education field.

REFERENCES

[1] Romero, C, Ventura, S. and Garcia, E., "Data mining in coursemanagement systems: Moodle case study and Tutorial". Computers & Education,Vol. 51, No. 1 pp.368- 384. 2008

[2] B. R.Sachin and S. M.Vijay, “A survey and Future Vision of data mining in educational field,”in Proc. International Conference on Advanced Computing Communication Technologies, pp. 96-100,2012

[3] Adrians, p., and D.Zantiuge. 1996. Data mining. Harlow, England:Addison Wesley.

[4] David Heckerman. 1995. A tutorial on learning with Bayesian N/W

[5] Thair Nu Phyu”Survey of Classification Techniques in Data Mining” Proceedings of the International MultiConference ofEngineers and Computer Scientists 2009 Vol I 2009, March18 - 20, 2009, Hong Kong

[6] S. B. Kotsiantis · I. D. Zaharakis · P. E. Pintelas” Machine learning: a review of classification and combining techniques” Published online: 10 November 2007 © SpringerScience+Business Media B.V. 2007

[7] A.Shameem Fathima1,D.Manimegalai2and Nisar Hundewale“A Review of Data Mining Classification TechniquesApplied for Diagnosis and Prognosis of the Arbovirus-Dengue” IJCSI International Journal of Computer ScienceIssues, Vol. 8, Issue 6, No 3, November 2011