An Analysis of Different Classifier on Soybean Dataset

Sandeep Sonali
CSE/IT, MITS GWALIOR

Abstract- *The paper presents an effort that has been carried out to make a performance evaluation of J48, Naive Bayes, IBk, SMO and Bayes Net classification algorithm. Naive Bayes algorithm relies on probability and J48 algorithm relies on a decision tree. The paper also denotes a comparative evaluation of various classifiers such as J48, Naive Bayes, IBk, Bayes Net and SMO in the context of Soybean datasets. The experiments are carried out using Weka tool developed by Waikato University. The results demonstrate that the efficiency of Bayes Net and SMO is good.*

Keywords- Classifiers, data mining

I. INTRODUCTION

Data mining involves the use of various sophisticated data analysis tools for discovering previously unknown, valid patterns and relationships in huge data set. These tools are nothing but the machine learning methods, statistical models and a mathematical algorithm. [1]. Data mining consists of over assortment and managing the info, it also includes analysis and Prediction. Classification technique in data mining is capable of processing a wider variety of data than regression and is growing in popularity. There are various data mining techniques are pre-processing, association, classification, pattern recognition and clustering [2]. Classification and association are the popular techniques used to predict user interest and relationship between those data items which has been used by users. Classification strategies include Bayesian network, J48, call tree, Neural Network, Decission Tree approach etc. Particularly this work is concerned with classification techniques [4,5,6,8,11]. The rest of the paper is organized as follows: Section II covers the literature review, Section III covers methodology, and finally in section IV summarize the comparative results.

II. LITERATURE REVIEW

A. Bayes Net classifier: Bayes Net classifier is based on the Bayes theorem. So, in Bayes Net classifier conditional Hill Climbing, Tabu Search, Simulated Annealing, Genetic Algorithm [7] and K2 such a different type of algorithms is used to estimate conditional probability in Bayes Net. In Bayes Net, the output of can be visualized in terms of a graph.

B. Naïve Bayes classifier: The name Naïve Bayes itself suggest that it is the updatable or improved version of Naïve Bayes. A default precision used by this classifier when build Classifier is called with zero training instances is of 0.1 for numeric attributes and hence is also known as an incremental update[2].

C. J48 Classifier : J48 a can be called an optimized implementation of the C4.5 or improved version of the C4.5. The output given by J48 is the Decision tree. A Decision tree is the same as that of the tree structure having different nodes, such as the root node, intermediate nodes, and leaf node. Each node in the tree contains a decision and that decision leads to our result as a name is decision tree. Decision tree divides the input space of a data set into mutually exclusive areas, where each area having a label, a value or an action to describe or elaborate its data points. Splitting criterion is used in the decision tree to calculate which attribute is the best to split that portion tree of the training data that reaches a node [1].

D. IBk Classifier: Simple instance-based learner (IBk) that uses the class of the nearest training instances for the class of the test instances. A value of 0 signifies no limit to the number of training instances [3].

E. SMO Classifier: Sequential Minimal Optimization (SMO) is one way to solve the SVM training problem that is more efficient than standard QP solvers. SMO uses heuristics to partition the training problem into smaller problems that can be solved analytically [13].

III. METHODOLOGY

WEKA

The full form of WEKA: Waikato Environment for Knowledge Learning. Weka is a computer program that was developed by the student of the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains [3]. Data preprocessing, classification, clustering, association, regression and feature selection these standard data mining tasks are supported by Weka. It is an open source application which is freely available

Dataset:

The dataset Soybean used in this paper has been taken from the UCI Machine Learning Repository [12]

Steps to apply classification techniques on the dataset and get the result in Weka:

Step 1: Take the input dataset.

Step 2: Apply the classifier algorithm on the whole data set.

Step 3: Note the accuracy given by it and time required for execution.

Step 4: Repeat step 2 and 3 for different classification algorithms on different datasets.

Step 5: Compare the different accuracy provided by the dataset with different classification algorithms and identify the significant classification algorithm for the dataset

IV. RESULTS AND DISCUSSION

A comparison of classifiers for different datasets based on the accuracy and time taken for execution is made. Accuracy is defined as the no of instances classified correctly [9]. It is observed from table 4.1 and 4.2, Bayes Net performed well with Soybean database, and SMO classifier outperformed with Soybean dataset in terms of correctly classified instances.

Table 4.1: Comparison of Accuracy for various classifiers

Name of classifier	*Soybean dataset*
	Accuracy
BayesNet	93.265
NaïveByes	92.9722
IBk	91.2152
J48	91.5080
SMO	94.282

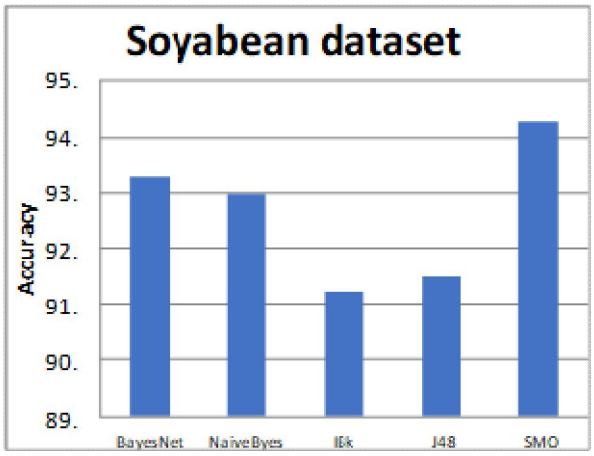

Figure: 4.1 show the graphical view of accuracy for various classifiers on Soybean dataset

Table: 4.2 Comparison of parameters for various classifiers

	CCI		ICI		KAPPA STATICS	MAE	RMSE	RAE	RRSE	TNI
	VALUE	%	VALUE	%	VALUES					
NAÏVE BAYES	637	93.265	46	6.735	0.9263	0.0085	0.0804	8.7916	36.6859	683
SMO	641	93.8507	42	6.1493	0.9326	0.0942	0.213	98.0165	97.1903	683
IBK	623	91.2152	60	8.7848	0.9036	0.0122	0.0879	12.71	40.1285	683
J48	625	91.5081	58	8.4919	0.9068	0.0135	0.0842	14.0484	38.4134	683
BAYESNET	637	93.265	46	6.735	0.9263	0.0085	0.0804	8.7916	36.6859	683

V. CONCLUSION AND FUTURE WORK

In this paper, we have compared the performance of various classifiers. Soybean data sets from benchmark dataset (UCI) are used for experimentation.

It is found that the performance of classification techniques varies with totally different knowledge sets. The SMO has given a good result with Soybean dataset. Our future work will focus on improvement of Classification Technique thereby improving the efficiency of classification in a decreased time. Also, a mix of classification techniques is accustomed to improve the performance.

REFERENCES

[1] V. Vaithiyanathan, K. Rajeswari, Kapil Tajane, Rahul Pitale, "Comparison of Different Classification Techniques using Different Datasets," in International Journal of Advances in Engineering & Technology, Vol. 6, Issue 2, pp. 764-768, May 2013.

[2] J. Han and M. Kamber, (2000) "Data Mining: Concepts and Techniques," Morgan Kaufmann.

[3] Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/

[4] W.A. Awad and S.M. ELseuofi, "Machine Learning Methods for Spam E-Mail Classification," International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[5] Giannotti, Fosca, Lakshmanan, Laks V. S., Monreale, Anna, pedreschi, Dino, and Wang, Wendy Hui, "Privacy-preserving mining of association rules from an outsourced transaction database," in IEEE Systems Journal, 385–395, 2013.

[6] Janez demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," Journal of Machine Learning Research, vol.7, 2006.

[7] Dr. Geraldin B. Dela, "Comparative Study of Data Mining Classification Techniques over Soybean Disease by Implementing PCA-GA," in International Journal of Engineering Research and General Science Volume 3, Issue 5, September-October, 2015.

[8] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, "Data Mining in Education: Data Classification and Decision Tree Approach," in International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012.

[9] D.Udhayakumarapandian, RM. Chandrasekaran, A. Kumaravel, " Modified Cross-Validation for Improving the Accuracy based on Randomized Partition over the Training and Testing Data Sets," International Journal of Computer Science and Mobile Computing, Vol. 5, issue. 9, 161 – 170, September 2016.

[10] Subhankar manna, Malathi G, "Performance Analysis of Classification Algorithm on Diabetes Healthcare Dataset," international general of research Granthaalayah, Vol.5, August 2017.

[11] Ritu Sharma, Mr. Shiv Kumar, Mr. Rohit Maheshwari, "Comparative Analysis of Classification Techniques in Data Mining using different Datasets," International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 12, pg. 125 – 134, December 2015

[12] UCI machine learning repository: https//archive .ics.uci.edu/m1/dataset.html

[13] Stack Exchange Portal: https://stats.stackexchange.com/questions/130293/svm-and-smo-main-differences