International Journal of Computational Intelligence and Informatics Vol. 4: No. 2, July – September 2014

Naïve Bayes Classification Technique for Analysis of Ecoli Imbalance Dataset

P Manikandan
Department of Computer Science
Bharathiar University
Coimbatore-046
manimkn89@gmail.com

D Ramyachitra
Department of Computer Science
Bharathiar University
ICoimbatore-046
jaichitra1@yahoo.co.in

***Abstract*- The classification technique is a systematic approach to build classification models from an input data set. The techniques include rule-based classifiers, decision tree classifiers, support vector machines, neural networks and Naive Bayes classifiers. Every technique employs a learning algorithm to discover a model that best fits the relationship among the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly forecast the class labels of records it has never seen before. Therefore, a key objective of the learning algorithm is to construct models with good generality capability. That is the models that accurately predict the class labels of previously unknown records. In this paper we are analyzing the performance of 3 classifiers algorithms namely Naïve Bayes, Instance Based K-Nearest Neighbor (IBK) and J48 Decision Tree. From the experimental results, it is found that Naïve Bayes technique performs better than the other two techniques. We use the ecoli protein datasets for calculating the performance by using the cross validation parameter. And finally we find out the comparative analysis based on the performance factors such as the classification accuracy and execution time is performed on all the algorithms.**

Keywords- Classification, Naïve Bayes, Instance Based K-Nearest Neighbor (IBK), J48 Decision Tree, Ecoli, Imbalance Dataset.

I. INTRODUCTION

Generally, Classification refers to the task of assigning objects to one of various predefined categories, is a determined problem that encompasses many different applications. The examples include categorizing cells as malignant or benign based upon the results of MRI scans, detecting spam email messages based upon the message header and content and classifying galaxies based upon their shapes. Classifiers are used to enhance the performance of given datasets. To construct or training a classifier is the process of creating a function or data structure that will be used for determining the missing value of the class attribute of the new unclassified instances. There are large numbers of learning schemes for classification and regression numeric prediction - like decision trees, instance-based classifiers, support vector machines, Bayes decision schemes, neural networks etc. Numerous attribute selection methods and evaluation methods exists like cross-validation and bootstrapping, and preprocessing techniques.

In this paper an analysis is made to find out which test option is the best for classifier algorithm called IBK, Naïve Bayes, and J48 decision tree. In the test option there are four kinds of parameter like training set, supplied test set, cross validation and percentage spilt. We use the cross validation parameter to calculate the data set values. This paper uses the ecoli protein dataset for comparison of those algorithms. And our paper is structured as follows. Section 2 describes the literature review, Section 3 describes the methodology for the ecoli protein dataset and Section 4 describes our experimental result. And finally Section 5 gives the conclusion and future work.

II. LITERATURE REVIEW

Pablo Bermejo, et al., presented a proposal that is based on the combination of the NB classifier with incremental wrapper feature subset selection (FSS) algorithms. The advantage of this approach is analyzed both theoretically and experimentally, and the results show a striking speed-up for the embedded FSS process [1]. Li-Min Wang, et al., proposed a novel algorithm, Self-adaptive NBTree, which induces a hybrid of decision tree and Naive Bayes. The Naive Bayes node helps to solve overgeneralization and overspecialization problems. The experimental results on a variety of natural domains indicate that Self-adaptive NBTree has clear advantages with respect to the generalization ability [2].

C.K. Chan, et al., compared numerically to the conventional preprocessing approaches such as data elimination, averaging, imputation to treat missing values. The efficiencies were confirmed by the classification accuracies through BayesNet, Lazy Kstar, Decision table and Part method classifiers [3]. B. Kavitha, et al., presented the classifying methods ID3, J48, Naive Bayes and OneR Their result shows that ID3 and J48 method carry the highest accuracy and sensitivity with 7 and 14 attributes. The Naive Bayes holds the highest degree of specification for all three dimensionalities [4]. Himadri Chauhan, et al., presented the comparison of different classification techniques to detect and classify intrusions into normal and abnormal behaviors. They used the J48, Naive Bayes, JRip, and OneR algorithms [5].

III. METHODOLOGY

Using the classification technique we find the best algorithm for the ecoli protein dataset. The flow diagram for the comparative analysis is shown in Fig. 1.

A. Dataset

The ecoli protein datasets has been collected from the Keel Repository database. This dataset contains 336 instances and 8 attributes. The data mining tool weka is used for analyzing the performance of these classification algorithms.

B. Classification

In this paper we have analyzed the classification algorithms to predict which of the algorithm is most suitable for the ecoli protein dataset. In these classifications we compare three algorithms namely IBK, Naïve Bayes and J48 decision tree to find out which one fits effectively for the ecoli protein dataset.

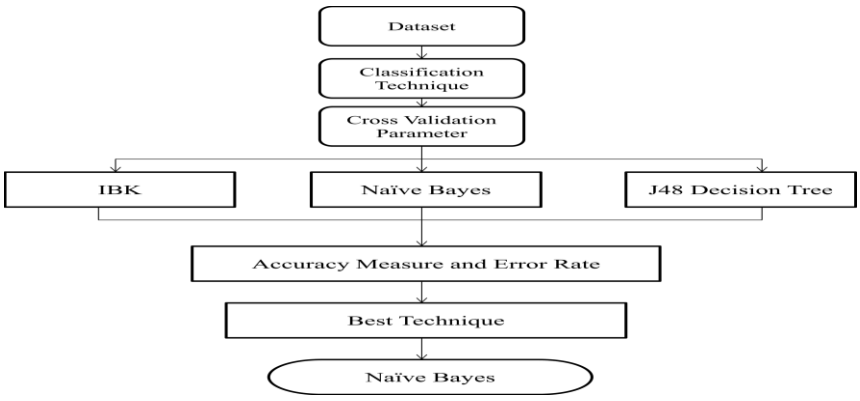

Figure 1. Comparative analysis

The classification algorithms are listed below.

1. Instance Based K-Nearest Neighbor (IBK)
2. Naïve Bayes
3. J48 Decision Tree

1) IBK

The IBK algorithm is a k-nearest-neighbor classifier that uses the same distance metric. The number of nearest neighbours can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. The distance function is used as a parameter of the search method. The remaining thing is the same as for IBL that is, the Euclidean distance; other options include Chebyshev, Manhattan, and Minkowski distances [4].

2) Naïve Bayes

The Naive Bayes classifier is a straightforward probabilistic classifier stand on applying Bayes' theorem with strong naive independence assumptions. A more expressive term for the underlying probability model would be "independent feature model". An inclusive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests [6].

3) J48 Decision Tree

The J48 algorithm builds the decision tree from labeled training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. The measure to compare the difference of impurity degrees is called information gain. The attribute with highest normalized information gain

is used to make the decision. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class [7].

IV. EXPERIMENTAL MEASURES

In this paper we calculate the experimental measures by using the performance factors such as the classification accuracy and execution time. And also we find out the accuracy measure and error rate to determine the best algorithm for the ecoli protein dataset. The performance factors for these classification algorithms are listed in Table 1 and the accuracy measure by class for the classifier algorithms is depicted in Table 2.

From the experimental results, it is inferred that for the cross validation parameter for Naïve Bayes algorithm, the Precision, F-Measure, TP rate values and the ROC value gives better results for the ecoli protein dataset. The performance factors for the classification algorithms are shown in Fig. 2 and the accuracy measure for the classifiers is shown in Fig. 3.

TABLE 1. PERFORMANCE FACTORS FOR THE CLASSIFICATION ALGORITHMS

Algorithms	TP Rate	Precision	F-Measure	ROC Curve	Kappa value	Execution Time
J48 Decision Tree	0.842	0.839	0.84	0.92	0.7826	0.03
Instance Based K-Nearest Neighbor (IBK)	0.804	0.799	0.801	0.878	0.7295	0
Naïve Bayes	0.851	0.861	0.851	0.96	0.7965	0.02

TABLE 2. ACCURACY MEASURES FOR CLASSIFICATION ALGORITHMS

Algorithms	Correctly Classified	Incorrectly Classified
J48 Decision Tree	84.22	15.78
Instance Based K-Nearest Neighbor (IBK)	80.36	19.64
Naïve Bayes	85.11	14.88

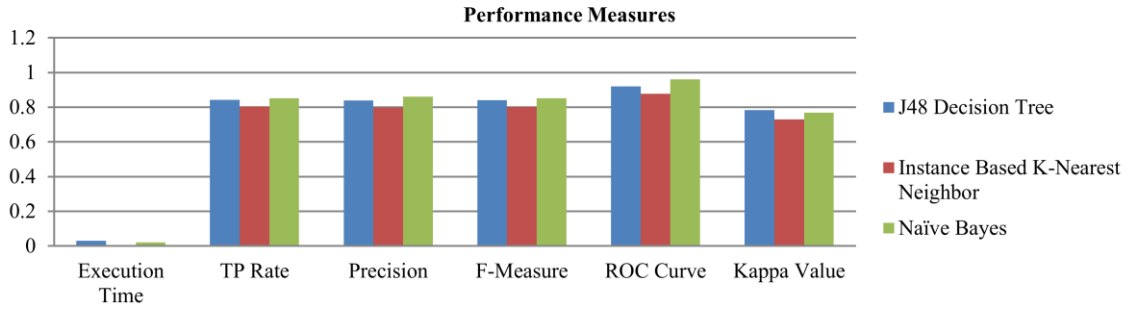

Figure 2. Performance Measures for the Classifier algorithms

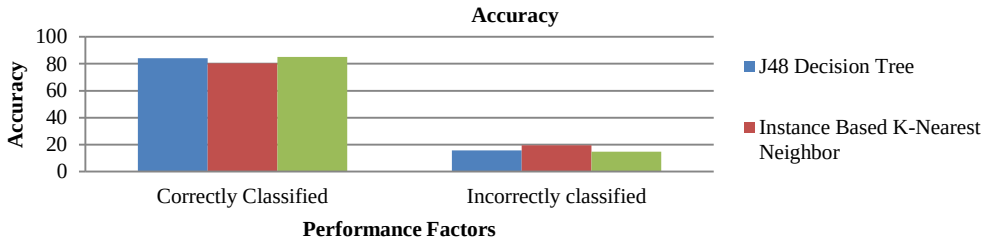

Figure 3. Accuracy Measure for the Classifier algorithms

For IBK algorithm it is inferred that for the cross validation parameter, the Precision, ROC, F-Measure, TP Rate values gives poor results than other algorithms. The Error rate measure for the classification is depicted in Table 3. And also Accuracy error rate measure for the classifier is shown in the Fig. 4.

For J48 Decision Tree algorithm it is inferred that for the cross validation parameter, the ROC value, TP Rate, Precision, F-Measure values gives better than IBK and poor results when compared to Naïve Bayes for the ecoli dataset.

TABLE 3. ERROR RATE MEASURE FOR CLASSIFICATION ALGORITHM

Algorithms	Mean Absolute Error	Root Mean Squared Error
J48 Decision Tree	0.0486	0.1851
Instance Based K-Nearest Neighbor (IBK)	0.0535	0.2189
Naïve Bayes	0.0434	0.1653

TABLE 4. ERROR RATE MEASURE FOR CLASSIFICATION ALGORITHMS

Algorithms	Relative Absolute Error	Root Relative Squared Error
J48 Decision Tree	26.59	61.34
Instance Based K-Nearest Neighbor (IBK)	29.24	72.56
Naïve Bayes	**23.71**	**54.78**

Figure 4. Accuracy error rate measure for classification algorithms

Figure 5. Accuracy error rate measure for classification algorithms

The experiment was carried out to the ecoli protein datasets by using the cross validation parameter. From the results it is inferred that the Naïve Bayes algorithm performs well as compare to the IBK and J48 Decision Tree. The Naïve Bayes algorithm gives more correctly classified instances compare to others. Also the error rate for Naïve Bayes algorithm is less compared to others.

The Table 5 and 6 shows the accuracy measure for the classifiers for various percentage splits and cross validation respectively. From the results it is found that, all the classifiers perform well for 76% split and the Naïve Bayes outperforms well than other classifiers. And also for cross validation the Naïve Bayes performs better than the remaining classifiers. The Fig. 6 and 7 shows the performance comparison of accuracy for the classifiers for different cross validation and percentage split respectively.

TABLE 5. PERFORMANCE COMPARISON OF ACCURACY FOR THE CLASSIFIERS FOR DIFFERENT PERCENTAGE SPLIT

Algorithms\Percentage Split	**56**	**66**	**76**	**86**	**96**
J48 Decision Tree	77.02	78.95	86.41	85.11	76.92
Instance Based K-Nearest Neighbor (IBK)	82.43	82.46	82.71	78.72	76.92
Naïve Bayes	**80.40**	**82.46**	**87.65**	**82.98**	**76.92**

TABLE 6. PERFORMANCE COMPARISON OF ACCURACY FOR THE CLASSIFIERS FOR DIFFERENT CROSS VALIDATION

Algorithms\Percentage Split	**5**	**10**	**15**	**20**	**25**
J48 Decision Tree	82.44	84.23	83.33	82.14	83.03
Instance Based K-Nearest Neighbor (IBK)	80.95	80.35	80.06	80.95	81.25
Naïve Bayes	**85.11**	**85.11**	**85.71**	**85.41**	**84.82**

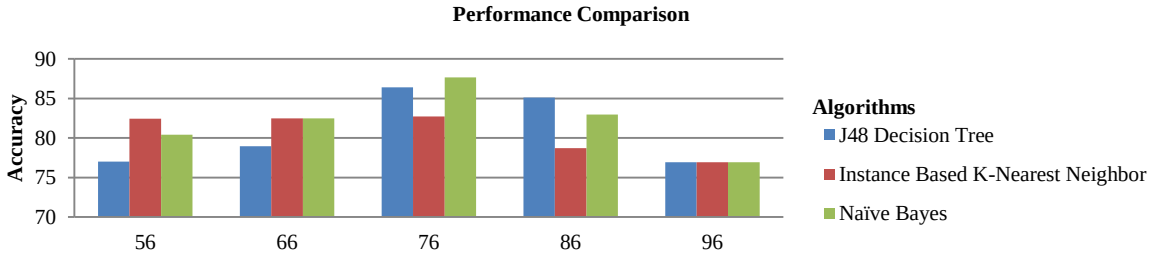

Figure 6: Performance comparison of accuracy for the classifiers for different percentage split

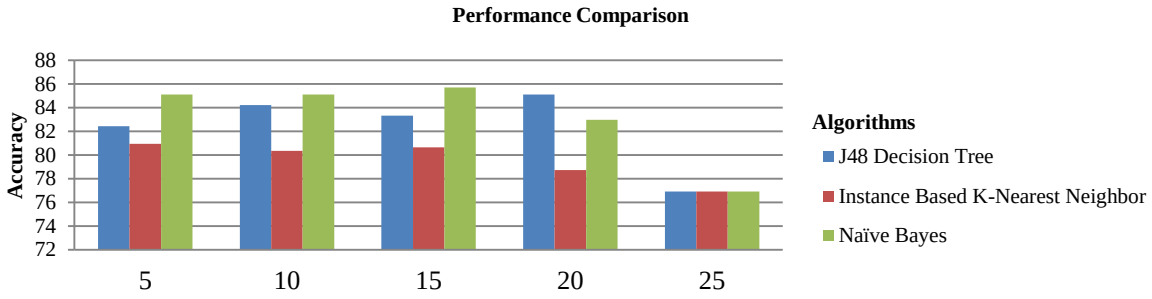

Figure 7: Performance comparison of accuracy for the classifiers for different cross validation

V. CONCLUSION AND FUTURE WORK

In this paper we analyzed the performance of 3 classifier algorithms namely IBK, J48 Decision Tree and Naïve Bayes. We used the ecoli protein datasets for calculating the performance by using the training set parameter. And finally we analyzed the algorithms by using the performance factors such as the classification accuracy and execution time. From the results, it is observed that the Naïve Bayes algorithm performs better than other algorithms.

In Future these classifications can be experimented on other datasets also. And in future we can modify the Naïve Bayes algorithm to obtain more effective results. And also the classification algorithms can be analyzed using parameters such as the cross validation, percentage split, and supplied test set.

REFERENCES

[1] Pablo Bermejo, José A. Gámez, José M. Puerta, "Speeding up incremental wrapper feature subset selection with Naive Bayes classifier", Knowledge Based Systems, vol. 55, pp. 140–147, 2014.

[2] Li-Min Wang, Xiao-Lin Li, Chun-Hong Cao, Sen-Miao Yuan, "Combining decision tree and Naive Bayes for classification", Knowledge-Based Systems, vol. 19, pp. 511–515, 2006,

[3] C.K. Chan, W.P. Loh, I. Abd Rahim, "Data Elimination cum Interpolation for Imputation: A Robust Preprocessing Concept for Human Motion Data", Procedia - Social and Behavioral Sciences, vol. 91, pp. 140–149, 2013.

[4] B. Kavitha, S. Karthikeyan, B. Chitra, "Efficient Intrusion Detection with Reduced Dimension Using Data Mining Classification Methods and Their Performance Comparison", Information Processing and Management Communications in Computer and Information Science vol. 70, pp. 96-101, 2010.

[5] Himadri Chauhan, Vipin Kumar, Sumit Pundir, Emmanuel S. Pilli, "Comparative Analysis and Research Issues in Classification Techniques for Intrusion Detection", Intelligent Computing, Networking, and Informatics, Advances in Intelligent Systems and Computing, vol. 243, pp 675-685, 2014.

[6] Caruana, R.; Niculescu-Mizil, A. (2006). "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning, pp. 161-168, 2006

[7] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, ISSN: 2319-5940, vol. 2, pp. 1925-1931, 2013.