EXPLORATION COMPARATIVE OF SUPERVISED LEARNING ALGORITHMS IN CLASSIFICATION MINING

Mrs.M.Porkizhi[1], Mrs. S.Nagarathinam[2],
Assistant Professors of CS & IT,
Nadar Saraswathi College of Arts and Science, Theni. TamilNadu, India.

Abstract

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Classification is one of the main technique in data mining which will be used to which is the process of verdict a model that describe the data classes or Concepts. These techniques are applied in learning algorithms such as Decision tree (DT), Support Vector Machines (SVM), Naive Bayes (NB) and Artificial Neural Network (ANN) and these methods can handle both numerical and categorical attributes. This study will be implementing in Rapid Miner tool and it will be applied in Titanic dataset. In this paper, four classification algorithms comparatively test to find the optimum algorithm for this dataset. This study described the performance analysis of classification algorithm based on the correct and incorrect instances of data classification. The comparison will be taking the following parameters such as Precision, Recall, F-Measure, Accuracy and Root mean squared error.

Keywords:

Data mining, Classification, Decision tree, Rapid Miner, Precision, Recall.

Correspondence: M.Porkizhi, Contact No: (+91)8508545105,

Email id: porkizhiraja007@gmail.com

I.INTRODUCTION

Data mining is the process, finding required knowledge from the large amount of database. Data mining having two types of learning which are supervised learning and unsupervised learning. Classification is one type of techniques in data mining which is based on supervised learning. Supervised learning often also called directed data mining the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables. The goal of the analysis is to specify a relationship between the dependent variable and explanatory variables the as it is done in regression analysis. Two set has performed here such as training and testing. In training set, contains a collection of records. Each record contains a set of attributes; one of the attributes is the class. Find a model for class attribute as a function of the values of other attributes. Testing set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it. Classification technique is

a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data). The proposed model architecture diagram shown below:-

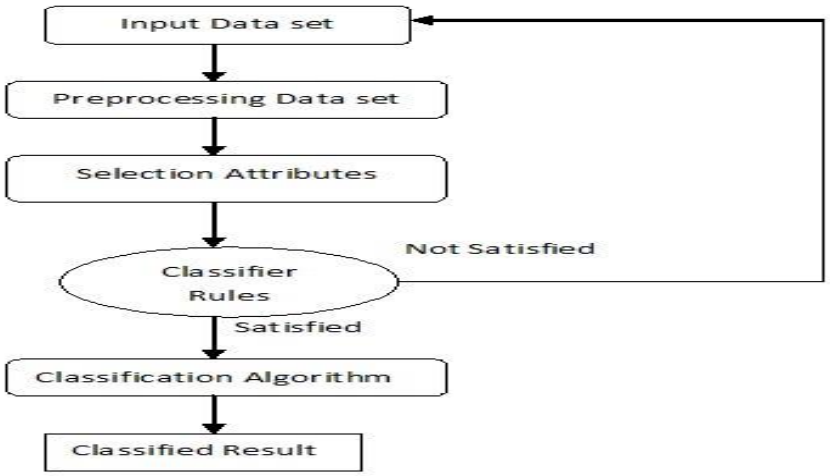

Fig 1: Proposed architecture

In proposed model architecture can be explained how the classifier rules will be act as main process of the classification process. Input dataset will be passed data to preprocessing dataset, which will reduce the noise and inconsistent data. Selection attributes part used to pick the necessary attributes to perform further process. Classifier rules check whether the incoming selection attributes could be satisfy the particular criteria or not. Classification algorithm has been processed only satisfied selection data. Finally, the result will be displayed in the form of classified manner. Classification techniques can be split up in to four types of supervised learning algorithms such as Decision Tree, Support Vector Machine, Naïve Bayes and Artificial Neural Network. The main focus is to compare the performance analysis of those four supervised learning using secondary dataset.

II. SUPERVISED LEARNING ALGORITHMS

Classification techniques can be compared on the basis of predictive accuracy, speed, robustness, scalability and interpretability criteria. In this study, four supervised learning algorithms were compared.

- ✓ Decision Tree
- ✓ Support Vector Machine
- ✓ Naïve Bayes
- ✓ Artificial Neural Network

DECISION TREE

A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. It has two goals: producing an accurate classifier and understanding the predictive structure of the problem.

SUPPORT VECTOR MACHINE

Support Vector Machine is used to separate the two classes by a function which is induced from available examples. The goal is to produce a classifier that will be worked well on unseen examples, i.e. it generalizes well. This linear classifier is termed the optimal separating hyper plane. Spontaneously, it would expect this boundary to generalize well as opposed to the other possible boundaries. An SVM outputs a map of the sorted data with the margins between the two as far apart as possible.

NAIVE BAYES

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. A naïve Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large dataset.

ARTIFICIAL NEURAL NETWORK

A neural network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the inter unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns. ANN process records one at a time and learn by comparing their classification of the record with known actual classification of the record. The errors from the initial classification of the first record is fed back into the network and used to modify the networks algorithm for further iterations. Neurons are organized into layers: input, hidden and output.

III. DATA COLLECTION

The dataset for this study “Titanic Training” has been collected from the in built dataset of the Rapid Miner tool because the dataset has been already preprocessed and found less number of missing value, noisy data. Therefore the result obtained will be more accurate and performance of the classifier also will be more efficient. Totally 916 samples with 7 attributes we collected and used in this study report. In Titanic dataset, the following attributes were used such as age group, sex, number of siblings, number of children’s on board, passengers fare and survived. This dataset has been analyzed and statistical calculation has been done on it to classify the survived and non-survived category.

IV. TECHNOLOGY USED

Rapid Miner is an open source data mining tool. This tool provides data mining and machine learning procedures, including data loading and transformation. This process will be based on ETL concept (Extract,

Transform and Load). It performs data preprocessing, and visualization, predictive analytics and statistical modeling, evaluation and deployment. Rapid Miner provides 99% of an advanced analytical solution through template based framework that speed delivery and reduce errors by nearly eliminating the need to write code.

V. PERFORMANCE ANALYSIS FOR SUPERVISED LEARNING ALGORITHMS

The following measurements will be used to analyze the best algorithm for given dataset. The goal of classification technique based algorithm, to find the optimum solution algorithm for "Titanic training" dataset. The dataset details shown below:

Titanic Training Dataset	
Attributes	7
Instances	916
Total Value of Dataset	916

Table 1: The Dataset Used in analysis

Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix contained measures such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The table 2 represented how the diagonal elements will be formed in the dataset.

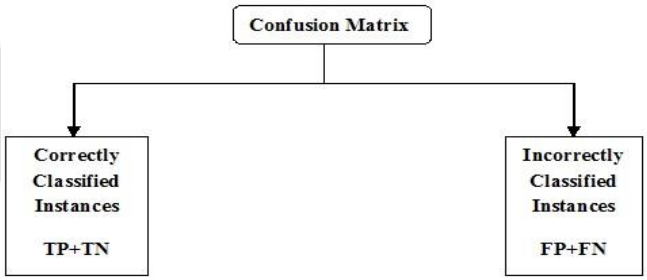

Fig 2: Confusion Matrix

Actual Vs Predicted		Predicted	
		Survived	**Not Survived**
Actual	**Survived**	True Positive **(TP)**	False Negative **(FN)**
	Not Survived	False Positive **(FP)**	True Negative **(TN)**

Table 2: Actual Vs Predicted Confusion Matrix

Total number of instances = Correctly classified instance + Incorrectly classified instance

Correctly classified instance = TP + TN

Incorrectly classified instance = FP + FN

The following measurement will be implemented by this dataset. The measurement has been shown in table 3.

Measurement Name	Formula
True Positive Rate	TP/ (TP+FN)
True Negative Rate	TN/ (FP+TN)
False Positive Rate	FN/ (TP+FN)
False Negative Rate	FP/ (FP+TN)
Precision	TP/TP+FP
Recall	TP/ TP+ FN
F-Measure	2*Recall*Precision/ (Recall+Precision)
Accuracy	TP+TN/ (TP+FP+TN+FN)
Misclassification Error Rate	1 – Accuracy

Table 3: Measurement Formula

VI. EXPERIMENTAL WORK AND ANALYSIS

The repository data contains 7 attributes and 916 instances respectively. Rapid Miner tool have been applied on "Titanic Training" data set taking cross validation for performance evaluation of the different supervised learning algorithms. In table 4, the confusion matrix and performance vector will be categorized for supervised learning algorithms.

ALGORITHM NAME	ACCRACY OF ALGORITHM	PERFORMANCE VECTOR
Decision Tree	accuracy: 79.04% +/- 3.51% (mikro: 79.04%) true Yes / true No pred. Yes: 234 / 77 pred. No: 115 / 490 class recall: 67.05% / 86.42%	PerformanceVector: accuracy: 79.04% +/- 3.51% (mikro: 79.04%) ConfusionMatrix: True: Yes No Yes: 234 77 No: 115 490
Support Vector Machine	accuracy: 78.93% +/- 3.42% (mikro: 78.93%) true Yes / true No pred. Yes: 239 / 83 pred. No: 110 / 484 class recall: 68.48% / 85.36%	PerformanceVector: accuracy: 78.93% +/- 3.42% (mikro: 78.93%) ConfusionMatrix: True: Yes No Yes: 239 83 No: 110 484
Naïve Bayes	accuracy: 78.17% +/- 4.58% (mikro: 78.17%) true Yes / true No pred. Yes: 236 / 87 pred. No: 113 / 480 class recall: 67.62% / 84.66%	PerformanceVector: accuracy: 78.17% +/- 4.58% (mikro: 78.17%) ConfusionMatrix: True: Yes No Yes: 236 87 No: 113 480
Artificial Neural Network	accuracy: 81.01% +/- 3.64% (mikro: 81.00%) true Yes / true No pred. Yes: 245 / 70 pred. No: 104 / 497 class recall: 70.20% / 87.65%	PerformanceVector: accuracy: 81.01% +/- 3.64% (mikro: 81.00%) ConfusionMatrix: True: Yes No Yes: 245 70 No: 104 497

Table 4: Accuracy and Performance from Rapid Miner

Table 5 reveals confusion matrix for mentioned four algorithms, which maps the actual and predicted values for the respective algorithms.

Actual	Predicted							
	Decision Tree		Support Vector Machine		Naïve Bayes		Artificial Neural Network	
	Survived	Not Survived	Survived	Not Survived	Survived	Not Survived	Survived	Not Survived
Survived	234	77	239	83	236	87	245	70
Not Survived	115	490	110	484	113	480	104	497

Table 5: Classifiers for Confusion Matrix

In table 6, depicts instances correctly predicted vs. instance incorrectly predicted with accuracy and total execution time taken by each algorithm. The accuracy of ANN is greater than other examined techniques but time taken to make model is greater than other respective algorithms and also observed that total time taken to make a model is minimum for NB model.

Algorithm	Correctly classified Instances In (%)	Incorrectly classified Instances In (%)	Accuracy In (%)	Running Time Taken In Seconds
Decision Tree	79.04	21.00	79.04	3
Support Vector Machine	78.93	21.07	78.93	5
Naïve Bayes	78.17	21.83	78.17	2
Artificial Neural Network	81.00	19.00	81.01	15

Table 6: Performance Measure about Confusion Matrix

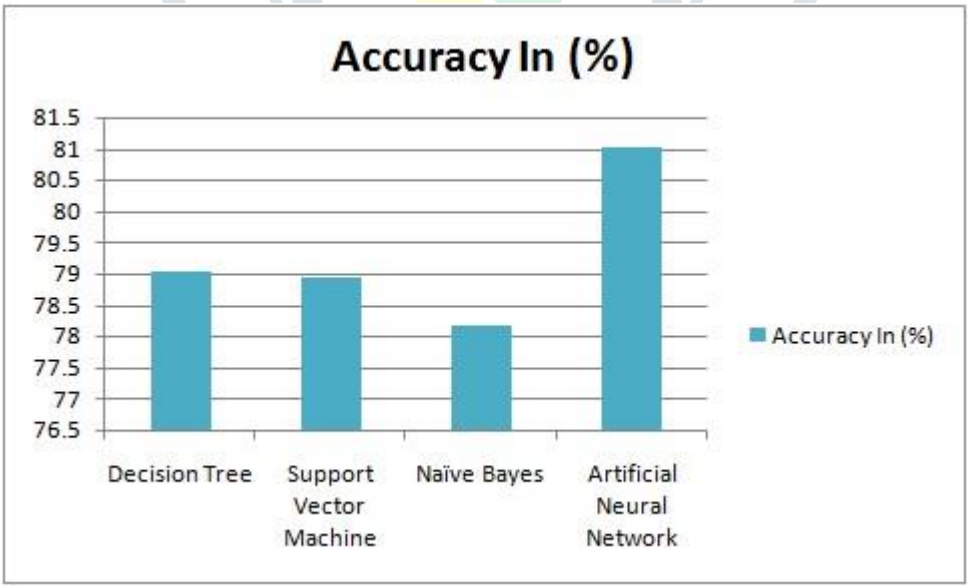

Fig 3: Accuracy

This chart has been shown like pictorial representation of comparison for accuracy to the algorithms. ANN was the highest accuracy compare with others.

In table 6, calculated True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) and False Negative Rate (FNR) as well as Precision and Recall. This Calculation has found the ANN is best among the DT, NB, SVM methods of classification.

Algorithm	TPR	FPR	TNR	FNR	PRECISION	RECALL
Decision Tree	0.75	0.25	0.81	0.19	0.67	0.75
Support Vector Machine	0.74	0.26	0.82	0.19	0.68	0.74
Naïve Bayes	0.73	0.27	0.81	0.19	0.68	0.73
Artificial Neural Network	0.78	0.22	0.83	0.17	0.70	0.78

Table 7: Performance Measure

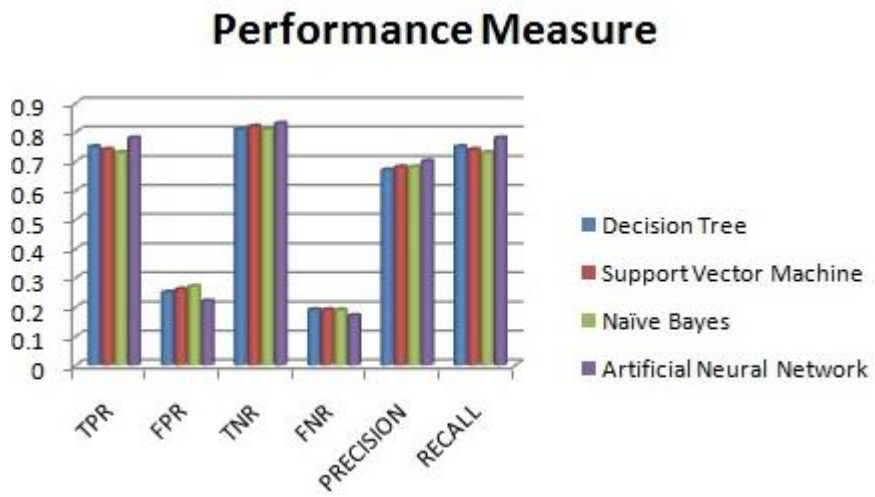

Fig 4: Performance Measure

Next we have also prepared error rate of each examined algorithm which is mentioned in Table 7. Hence they observed that ANN is showing minimum error rate than the other techniques, and SVM algorithm is showing maximum error rate.

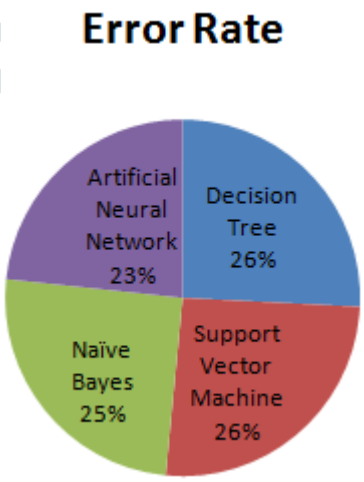

Fig 5: Error Rate

VII.CONCLUSION

This study has been analyzed and examine with the DT, NB, SVM and ANN methods using 916 samples from the Titanic training data set. The observations were noticed and discussed. In the discussion it

was found that the supervised method artificial neural network had maximum accuracy and minimum error rate. And also it was noticed that the running time of ANN is high. When compare with other algorithms, precision, recall and F-measure values has been elevated. According to running time, NB has run minimum time compare with other three algorithms. Error rate has been high in SVM. On the basis of accuracy measures of the classifiers and performance measures of classification used to easily understand the guidelines of result process. This result has classified in a category such as survived and non-survived.

In future, the processing speed can be reduced till more for ANN and error rate of SVM will be reduced with the help of modified algorithm. More similar studies on different data set for supervised learning approach are needed to confirm the above finding result.

VIII.REFERENCES

1. Han, J. and Kamber, M. Data Mining: "Concepts and Techniques", 2001 (Academic Press, San Diego, California, USA).
2. Tomoki Watanuma, Tomonobu Ozaki, and Takenao Ohkawa. ―"Decision Tree Construction from Multidimensional Structured Data"‖. Sixth IEEE International Conference on Data Mining – Workshops, 2006.
3. Micheline Kamber, Lara Winstone, Wan Gong, Shang Cheng, Jiawei Han, ―"Generalization and Decision Tree Induction: Efficient Classification in Data Mining"‖, Canada V5A IS6, 1996.
4. Caruana, R. and Niculescu-Mizil, A.: "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning, 2006.
5. X. Yang, Y. Guo, and Y. Liu, "Bayesian-inference-based recommendation in online social networks," Parallel and Distributed Systems, IEEE Transactions- April 2013.
6. G.Kesavaraj, Dr. S.Sukumaran, "A Study on Classification Techniques in Data Mining", IEEE-31661, July 4-6, 2013.
7. N. Cristianini and J. Shawe-Taylor. An Introduction to support vector machines and other kernel based learning methods. Cambridge University Press, 2000.
8. J. Platt. Fast training of support vector machines using sequential minimal optimization. In C. B. B. Scholkopf and A. Smola, editors, Advances in Kernel Methods | Support Vector Learning, MIT Press, 1999.
9. E. Osuna, R. Freund and F. Girosi. An Improved Training Algorithm for Support Vector Machines. To appear in Proc. of IEEE NNSP'97, Amelia Island, FL, 24-26 Sep., 1997.
10. Ackley, D.H., G.E. Hinton, and T.J. Sejnowski, "A Learning Algorithm for Boltzmann Machines," Cognitive Science, 1985.
11. German I.Parisi, Jun Tani and Cornelius weber"Lifelong learning of human actions with deep neural network self-organization", Neural Networks, December 2017.
12. Gullapalli V,"A stochastic reinforcement learning algorithm for learning real-valued functions", Neural Networks 3, 1990.
13. G. A. Carpenter and S. Grossberg. ART 3: "Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures", Neural Networks, 3(2):129–152, 1990.