A Comparative Analysis of Supervised Machine Learning Methods using Disaster Datasets

Mullapudi Raghu Ram
Dept. of I.T., ANITS
Visakhapatnam

Potnuru Sai Anish
Dept. of I.T., ANITS
Visakhapatnam

Burada Basant
Dept. of I.T., ANITS
Visakhapatnam

ABSTRACT
Supervised machine learning is one of the machine learning task that generates required function from the training data which is labelled. The aim of supervised machine learning is to build or construct a model that makes predictions by using the function inferred from the labelled training data**.** This paper put a light on how the supervised machine-learning techniques are used to build a predictive model from the dataset of titanic disaster and also a comparative analysis of supervised machine learning methods like Random Forests and Decision Trees are implemented. In this work, with a training dataset containing features or labels like sex, age and class, survivors are predicted from the four test datasets. And from the observations of results a comparative analysis of both supervised machine learning methods namely Decision Trees and Random Forests is implemented.

Keywords
Supervised machine learning, Decision Trees, Random Forests.

1. INTRODUCTION
As Data mining is actually a part of Knowledge discovery process, it involves the use of sophisticated data analysis tools to discover the valid patterns and valid relationships among the attributes or features that are present in large data set. Statistical models, mathematical algorithms and machine learning methods are the tools that are used in data mining. Evidently, data mining consists of more than collection and managing data, it also includes analysis and prediction**.**

Classification and prediction are two types of data analysis that comes under supervised machine learning methods which can be used to derive models by describing important data classes or to predict future data trends where the instance are given with known labels, unlike unsupervised learning is of unknown labels. This analysis can help us in understanding of the large data in a better way. Classification predicts categorical labels, while prediction models or generates a continuous valued functions. Classification technique is capable of processing a wider variety of data than regression and classification is increasing its popularity. Each instance in the dataset used by supervised or unsupervised learning method is represented by a set of attributes which may be categorical or continuous.

In Classification the model is built from the training set made up of database instances and associated class label. The resulting model is then used to predict the required class label of the testing instances or samples of testing datasets in which the values of features which are being predicted are known, which are further used to find the accuracy of the prediction models.

This work concentrated on the comparative study of some very well-known classification algorithms like Decision tree and Random Forests. A comparative study would definitely bring out the advantages and disadvantages of one method over the other. This would provide the guideline for research issues which in turn help other researchers in developing algorithms for applications of data mining which are not available.

2. MATERIALS AND METHODS USED

2.1 Data Set
The data used for this work was provided online by encyclopedia. The training set consists of 891 passenger samples and also their labels are associated of whether the passenger is survived or not. For each passenger, the details such as the fare, class in which passenger was travelled, port of embarkation in which the passenger was boarded and age of the passenger and other important attributes such as name, sex, etc. are given. For the test data, there are four datasets with 104 samples each with the same attributes as of the training dataset.

The given dataset is not complete, therefore many fields for different samples or instances of dataset were marked empty such as the fields of age, embarked port and fare, etc. However, all samples or instances of the dataset was provided with minimal information about gender and passenger class which are required for the work. To facilitate the data for generating models, all missing values are replaced with the mean of the remaining data of that particular attribute from the provided dataset.

2.2 Supervised Machine Learning Methods

2.2.1 Decision Trees
A decision tree is a method used for classification and prediction and for facilitating decision making in sequential decision problems. In this work, the training data set is analyzed by visualizing the data set and finding the best patterns in dataset in order to construct decision trees. Here, three decision trees are constructed by using different group of labels, the combination of labels are as described in Table 1. The three decision trees refer to fig 1,2,3 are subjected to test the accuracy of generated models using test datasets and the results used for further analysis.

2.2.2 Random Forests
Random Forests is an ensemble machine learning method for classification and regression that builds many decision trees at training time. In this work the training data set is subjected to random forests classifier then it generates or builds many decision trees at training time gives out the best decision tree that was constructed and it was subjected over test datasets. Here, the training dataset is subjected to random forest classifier by using different group of labels, the combinations of labels are as described in Table 1. The three models are

evaluated by using the test datasets and the accuracy of models are calculated and the results are used for further comparative analysis.

Table 1: Combinations of labels

Groups	Combination
1	Class , Sex
2	Class, Sex, Age, no of siblings
3	Class, Sex, Age, no of siblings, port of embarkation, fare

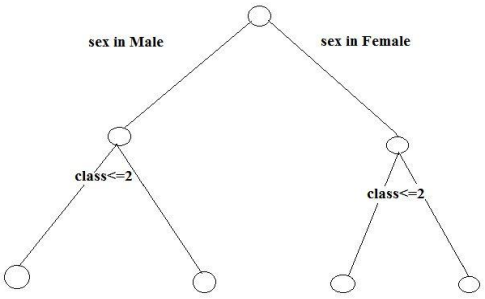

Fig 1: Decision tree of group1 combination labels

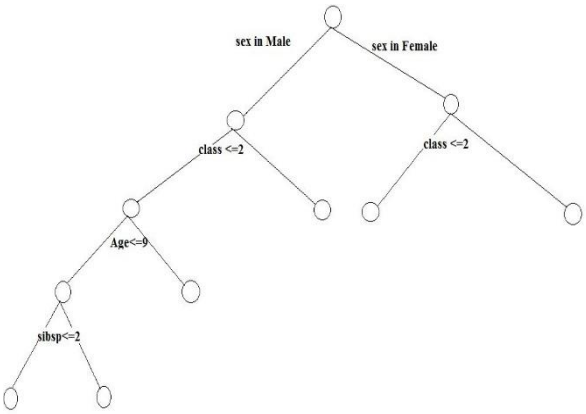

Fig 2: Decision tree of group2 combination of labels

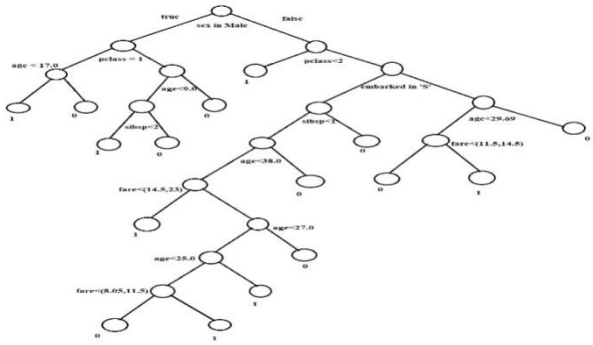

Fig 3: Decision tree of group3 combination labels

3. PROPOSED WORK

In this work the goal is to analyze the disaster datasets probably, in this work titanic dataset was used and make a comparative analysis of the Supervised machine learning methods like decision trees and random forests. Here, the generated or constructed models are used evaluate by using the test datasets and their respected results were used to make a comparative analysis of both methods would definitely bring out the advantages and disadvantages of one method over the other.

4. PERFORMANCE AND EVALUATION MEASURES

4.1 Accuracy

The relativeness of the measured value to the actual value is called accuracy. In these analysis the formula for calculation of accuracy of results is:

$$\text{Accuracy (a)} = (Cp / T) * 100$$

Where,

- Cp is the number if samples that were predicted correctly in a test dataset
- T is the total number of samples in a test dataset

5. RESULTS AND DISCUSSION

In this paper the titanic disaster dataset is used as training dataset for the supervised machine learning methods like decision trees and random forests. And the resultant models are used for predicting the survival of the samples in the test datasets and the results are subjected to conduct the comparative analysis of both decision trees and the random forests. From the results Table 2, an observation can be made that the models that generated by the combination of group three are having more accuracy percentages than the other combinations, by these a conclusion can be made that as if depth of the tree increases the accuracy rate is also increased by selecting the proper labels as the part of the combination that were used for generating the model. The results of both decision trees and the random forests yields a conclusion that the models that are generated by the random forests are high with accuracy rates than the decision trees with minimum number of labels as the part of the combination used for generating the model. And another observation could be that the decision trees having higher difference of accuracy percentages than the random forests as the number of labels in combination decreases, thus the random forest having lesser difference in accuracy when compared to the decision trees in Table 3.

Table 2: Results of Various Datasets,(DT refers to Decision Trees, RF refers to Random Forests)

Data sets	Combination Group 1		Combination Group 2		Combination Group 3	
	D.T	R.F	D.T	R.F	D.T	R.F
1	58.65	68.26	74.03	72.11	75.96	73.07
2	58.65	72.11	75.96	73.07	77.88	74.03
3	62.5	67.30	75.96	69.23	76.92	75.96
4	61.32	81.13	86.79	83.96	85.84	86.79

Table 3: Average differences of Accuracy

Datasets	Average difference of Accuracy	
	Decision Trees	**Random Forests**
1	8.665	2.405
2	9.615	0.96
3	7.21	4.33
4	12.26	2.83

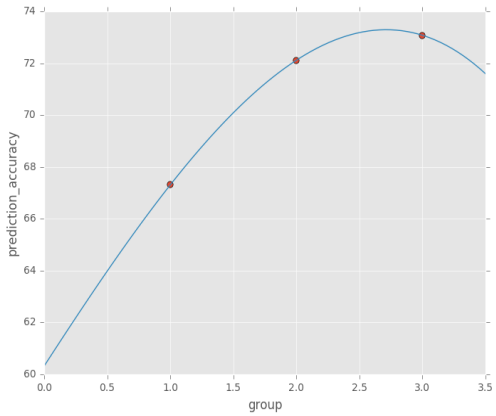

Fig 4: Random Forests Analysis

6. CONCULSION AND FUTURE WORK

In this study, the analysis of the titanic disaster dataset is carried out by using two supervised machine learning techniques namely Decision trees and Random Forests. By the comparative analysis of these methods (Fig 4 and Fig 5) one can notice that average difference of accuracy is less for random forests compared to that of decision trees. From the results, a conclusion that if the depth of the tree increases with a proper selection of labels for a combination there will be a substantial increase of accuracy. Also, the difference of accuracy for random forests is very less and also when compared to decision trees, random forests are high accuracy even with less numbers of labels in a combination. Consideration of the missing values in the training datasets which are very much essential for the construction of the Decision Trees and Random Forests and also performance of the two methods on missing attributes can be analyzed.

This study shows the importance of choosing important features and obtaining good data. It would be interesting to continue this analysis with other possible features like identifying the patterns in titles of passenger names and also with other supervised machine learning algorithms.

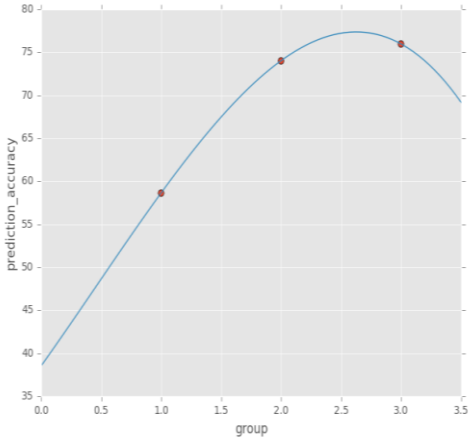

Fig 5: Decision Tree Analysis

7. REFERENCES

[1] A comparative analysis of machine learning methods for classification type decision problems in healthcare, Emanet et al. Decision Analytics 2014, 1:6, http://www.decisionanalyticsjournal.com/1/1/6

[2] Maimon, O., & Rokach, L. (2010). "Data Mining and Knowledge Discovery Handbook". (2nd, Ed.) Springer

[3] Kotsiantis, S. B. (2007). "Supervised Machine Learning: A review of classification techniques", Vol 160, No. 3, Frontiers in Artificial Intelligence and Applications

[4] Quinlan, J. R. (1987). "Generating production rules from decision trees". Proceedings of the 10th international joint conference on Artificial intelligence, pp. 304-307

[5] Tsang, S., Kao, B., Yip, K. Y., Ho, W. -S., & Lee, S. D. (2011, Jan). "Decision Trees for Uncertain Data". IEEE Transactions on Knowledge and Data Engineering, Vol 23, No. 1

[6] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (2006, March). "From Data Mining to Knowledge Discovery in Databases". The Knowledge Engineering Review, Vol 21, No. 1, pp. 1-24

[7] Breiman, L. 2001a. Random forests. Machine Learning 45:5