Using Ensemble Learning and Association Rules to Help Car Buyers Make Informed Choices

Akarsh Goyal
VIT University
Computer Science and Engineering
Vellore, India
akarsh.goyal15@gmail.com

Saurabh Thakur
VIT University
Computer Science and Engineering
Vellore, India
thakursc1@gmail.com

Rahul Chowdhury
VIT University
Computer Science and Engineering
Vellore, India
chowdhuryrahul5@gmail.com

ABSTRACT

Cars are an essential part of our everyday life. Nowadays we have a wide plethora of cars produced by a number of companies in all segments. The buyer has to consider a lot of factors while buying a car which makes the whole process a lot more difficult. So in this paper we have developed a method of ensemble learning to aid people in making the decision. Bagging, boosting and voting ensemble learning have been used to improve the precision rate i.e. accuracy of classification. Also we have performed class association rules to see if it performs better than collaborative filtering for suggesting item to the user.

Keywords

Ensemble Learning, Clustering, Classification, Class Association Rules, Accuracy

1. INTRODUCTION

A car is a wheeled, self-powered motor vehicle used for transportation used by people in their day to day living. We all like cars. Cars are everywhere and have become a necessity in our lives especially for those living in metropolitans due to its fast-paced and monotonous way of livelihood. They are ubiquitous in today's world. Nowadays they have become so common in every household that it has become extremely difficult for us to think of our existence without them even for one day. Having a nice luxurious car is a sign of opulence, pride and a higher standard of living for people. People take a lot of things into consideration while buying a car like safety features, leg-space, capacity, mileage, maintenance and price. Safety is a factor which cannot be done without. There are many car evaluators like magazines, TV-shows, online blogs, newspapers who give first-hand knowledge about cars taking the above mentioned features into consideration. But a lot of raw data is generated in the process which goes relatively unused. So here we introduce the concept of using ensemble learning to classify cars based on their suitability as considered appropriate by buyers. This could be put to great use. By using it the evaluation process becomes a lot easier for those who specialize in evaluating the cars. The study determines the acceptability based on factors like doors, no. of persons, leg-space and safety.

Ensemble Learning is a sub-concept present in the much bigger domain of data mining. Data Mining [3], [10] is the process of discovering patterns in large datasets by using methods from various fields of interest. It has many area of applications and it could take the form of either clustering, classification, association rules or regression. The various concepts which are a part of it are k-means, k-median, KNN, Artificial neural networks, Naïve Bayes, random forest etc.

Ensemble learning [4] uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. It refers to a concrete finite set of alternative models and typically allows for much more flexible structure to exist among those alternatives. So here we will be using many base classifiers and combining their results by using ensemble methods to get much more accuracy than could be achieved by these individual classifiers when they are used alone.

We have also applied class association rule, specifically Apriori. Apriori is used to identify strong rules is the database based on the attributes. The knowledge gained from it is used for making decisions.

We have also discussed the idea of collaborative filtering which could be used by the car buyers as a recommender system for buying cars and making decisions.

2. RELATED WORK

A study conducted by [3] based on using data mining in order to combine multi-attribute decision making with expert systems is given. It also deals with interactive knowledge acquisition and knowledge explanation. In [10] proper use of car evaluation dataset is demonstrated by developing a new machine learning method. This method effectively decomposes the problem into smaller, less complex problems. In [5] different collaborative filtering algorithms have been discussed. In [7] boosting has been described. Related work in [8] applies bagging [2] and boosting to perform linear discriminant analysis. This paper also gives a highlight on how mini sample size properties of the base classifiers affect their combination. In [1], Apriori algorithm has been improved by elimination of unimportant candidate keys with factors like set size and set size multiplicity.

BDAW '16, November 10-11, 2016, *Blagoevgrad, Bulgaria*
© 2016 ACM. ISBN 978-1-4503-4779-2/16/11...$15.00
DOI: http://dx.doi.org/10.1145/3010089.3010093

3. DATASET USED

The dataset used in this paper was taken from UCI dataset repository where various datasets are available for public use. The dataset was donated to UCI repository by M. Bohanec and V. Rajkovic.

The categorization is as shown for the dataset as shown in Table I:

Table 1. Description of Dataset

Data Set Characteristics	Attribute Characteristics	Number of Instances
Multivariate	Categorical	Classification

Number of Instances	Number of Attributes	Missing Values
1728	6	No

The class values in the Car evaluation dataset are as shown in Table II:

Table 2. Description of class values

Class	Frequency
Acc	385
Good	70
Unacc	1207
Vgood	66
Total	1728

A recce was done on the dataset to identify different matches.

4. APRIORI ALGORITHM

The Apriori Algorithm is an influential algorithm for boolean association rules and is used in mining frequent itemsets. The main aim is to find the frequent itemsets: the sets of items that have minimum support . A subset of a frequent itemset must also be a frequent itemset itemsets. The candidates with minimum support calculated by repeatedly performing the iterations to get the epochs is done. This is used to get to the accurate result.

5. COLLABORATIVE FILTERING

The method [5] aims to identify similar users based on their interests and preferences which if similar are tagged together. This method is relevant as the opinion of a person who has similar taste as the buyer could be used by that person to buy something. Implicit or explicit methods could be used to obtain the opinion of the buyers. The approach assumed by CF has two important steps :

1. The views and tastes of users is taken and the phase s collaboration phase.
2. Predictions of users interest is made from the data taken through gleaning and the phase is filtering phase

Assume we have m persons and n items which can be arranged in a M*N matrix where each row is a person, each column is an item. Thus if a person i likes item j then M(i,j)=1,otherwise M(i,j)=0. K-means type clustering methods are used. Cluster representation provides the missing values.

The block diagram of the Collaborative Filtering Tool is given below. Let us consider each unit of this tool.

A. **Database / Data Warehouse**: Logged data is contained in it. Also the final recommendations to be shown to the user is stored in it.

B. **Transaction Analysis:** Periodical analysis of all the 'followed' information is done. The new transactions are identified and the changes are then propagated to the next stage.

C. **Matrix Update Unit**: Based on the analysis the matrices are periodically updated. These matrices can be updated periodically or after 'n' number of new transactions. An entry gets added to the person-item matrix for every new transaction.

D. **Collaborative Filtering Unit:** Person-based filtering tool works on the matrices and compute recommendations along with their strengths.

E. **Rating Unit**: Recommendations provided by the tool is classified as top recommendation.

F. **Sorting Unit**: Recommendations belonging to a particular class are sorted according to their strengths. These are then updated in the database.

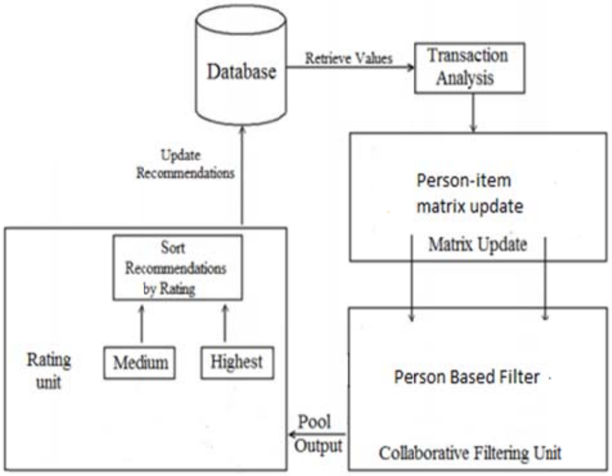

Figure 1. Block Diagram of Collaborative Filtering Tool

6. CLASSIFICATION TECHNIQUE

Different Classifiers are evaluated to find the effectiveness of those classifiers in the classification of Car Evaluation Data set. They are namely –

1. **Naïve Bayes** - In probability theory, conditional and marginal probabilities are related by Bayes theorem for two random events. It is often used to compute probabilities given observations. Our goal is to build a classifier to predict its unknown class label based on Bayes theorem. A technique for constructing such classifiers to employ Bayes' theorem to obtain:

$$P(Ck|x) = \frac{P(x|Ck)P(Ck)}{\sum k\prime \, P(x|Ck\prime\,)P(Ck\prime\,)} \qquad (1)$$

A naive Bayes classifier [6] [9] assumes that the value of a particular feature of a class is unrelated to the value of any other feature, so that:

$$P(x|Ck) = \prod_{j=1}^{d} P(x\, j\, |Ck) \qquad (2)$$

2. **Decision Tree** - It is a commonly used method in data mining. The goal of this classifier is to create a model that predicts the value of a target variable based on several input variables. An example depicting the use of it is shown below. The interior nodes correspond to the input variables; then there are edges to children for each of the possible values of that input variable. The leaves represents the values of the target variables given the values of the input variables represented by the path from the root to the leaf. A decision tree is a simple representation for classifying examples. Data comes in records of the form:

$$(y,X) = (y_1 , y_2 , y_3 , ... , y_k , X) \tag{3}$$

Vector y is made up of of the input variables, y1, y2, y3 etc. The dependent variable, X, is the output variable.

3. **Support Vector Machine (SVM)** – It is a supervised learning algorithm. Classification as well as regression could be done easily using support vector machine. Plotting is done in this method where each record is depicted by a point in the n-dimensional space. Here n denotes the number of attributes we have. Then hyper-plane is found out which divides the two classes very well. This is how classification takes place in support vector machine.

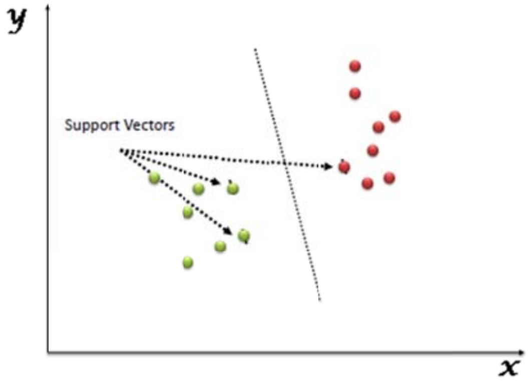

Figure 2. Support Vector Machine

The co-ordinates of individual observation denote the support vectors. It acts as a frontier which best segregates the two classes (hyper-plane/ line).

7. ENSEMBLE CLASSIFIER

The main function of ensemble classifiers is to improve the adherence level so that the classifiers are not variant in property. It combines various classifiers through different methods to achieve this functionality. The classification performance achieved by it is better than any base classifier used by it. In comparison with the base classifiers used the ensemble is very robust to noise. The method divides a complicated problem into various units and these units are much easier to compute upon and comprehend. They can adapt much easily to the data changes than the base classifiers involved. The precision of values predicted by it is much more. The ensemble classifier is termed successful if the individual classifiers are very much different to each other with respect to the misclassified ones. Four ways to get this change are that the base classifiers could be trained by using data which differ from each other, the parameters could differ, the attributes to coach the data could be different and the last one is the combination of different classifiers. The reasons for base classifiers performing badly as compared to ensemble classifiers are as follows:

- The training procedure of base classifiers might be having some faults.
- The knowledge provided by the data might be lesser than needed for utilization by a base classifier.

The various ensemble classifiers put to use in this paper have been elucidated below:

Bagging- One of the easiest technique used for ensemble is bagging [2] [8]. And yet it is the most successful methods for increasing the stability of classification problems. Like, decision tree algorithms which are weak classifiers and are not stable. A different tree is formed when there is a change in the position of the training point which leads to decrease in stability. It could be applied to any method. Finding the right model for large-dimensionality datas becomes an impossible task because they have huge complexity of the problem. So here bagging technique is very helpful. Variance of the classifier is reduced by using bagging technique. The technique goes as follows: let M be the size of the original training set. The training set is used many times by picking randomly M examples from the data and that too with replacement. A different model is trained by each of these data sets. Voting combines the output of the models to create a single output.

Boosting- For increasing the performance of weak classifier by using some computations into a strong classifier boosting [6][7] is used. The model could be applied to both classification and regression. It averages the whole model's initial output. The training of the classifiers happens in a sequence. The first classifier learns from the entire dataset whereas the following ones learn from the learning sets using the performance of the first classifier as the reference. The examples which are classified incorrectly are highlighted and they are given more weightage so that they will have a higher chance of appearing in the learning class of the next classifier. The outcome is a set of machines which become talented enough to classify different parts of the datasets. The boosting technique used in this paper is Adaboost which is will transform weak predictors into strong ones. Freund and Schapire introduced the AdaBoost algorithm. The foundation of Adaboost algorithm is very strong and the predictions are very accurate.

Voting- It is a combination of many classifiers. The vote by majority classifiers is put to use by it. Accurate classification is achieved by it in this way. The limitations of equally good classifiers is settled out by this algorithm. Majority of the attribute which is the mode is the class label for the sample. Like if classifier 1 predicts class 1, 2 predicts 2 and 3 predicts 1 then the class label of the sample is 1.

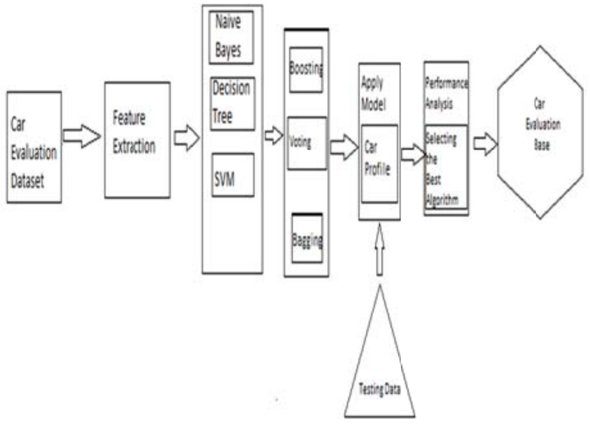

Figure 3. Ensemble Learning Process

Evaluation Criteria

Accuracy Classification

The accuracy of classification is given as follows :

Accuracy = (Correctly Classification of instances / Total Number of Instances)*100

8. RESULT

Class Association Rules- The association rules to each attribute of the class level attribute is produced by it. The class association rule applied here was Apriori. Table underneath presents the result of Apriri. Apriori outperforms Collaborative filtering technique. So we have only shown the result achieved on performing apriori.

Table 3. Apriori on the car dataset where it is false

No. of Cycles	14
Minmum Support	0.255 (438 instances)
Confidence Metric	0.9
Rules in number	12
Rules with Confidence level 0.9	12

Table 4. Apriori class association rules on complete car dataset – car (true)

No. of Cycles	17
Minmum Support	0.154 (259 instances)
Confidence Metric	0.9
Rules in number	19
Rules with Confidence level 0.9	11

Ensemble Learning- For getting the result from the classifiers we have used 10-fold cross validation technique. In this approach the data is partitioned into 10 subsets. One part is used for testing the data and the 9 parts are used for its training. Then the same process is repeated with the other 9 subparts in sequence. Then the result of each is taken and averaged to get the final output.

Overall, the performances of all the classifiers were good and 92.3 was the highest accuracy and the lowest was 56.04% .Table below shows that Adaboost when implemented with Naïve Bayes acts as a weak classifier and achieves the lowest accuracy, which is 56.04%. On the other hand, the SVM single classifier achieves the highest accuracy of 92.3%.

Table 5. The performance of bagging and boosting using 10-fold cross validation

Algorithm	Single Classifier	Accuracy Improvement			
		Bagging	%	Boosting	%
Naïve Bayes	60.3%	60.93%	0.63%	56.04%	-4.26%
Decision Tree	86.7%	86.96%	0.26%	86.90%	0.2%
SVM	92.3%	-	-	-	-

Table 6. The performance of voting and random forest using 10-fold cross validation

Algorithm	Single Classifier	**Accuracy Improvement**			
		Voting	%	Random Forest	%
Naïve Bayes	60.3%	89.8%	29.5%	-	-
Decision Tree	86.7%		3.1%	86.2%	-0.5%
SVM	92.3%		-2.5%	-	-

Tables above shows that the use of the bagging, boosting, voting and random forest algorithms did not improve the accuracy significantly. Only the use of voting was able to improve the accuracy, by 29.5% and 3.1% for Naïve Bayes and Decision Tree respectively. Others had a difference of only 1% accuracy from single classifiers.

We did not mention the accuracy for Boosting on SVM and Bagging on SVM. This is because SVM is a very strong classifier. So on applying Ensemble techniques on it, its accuracy will become very less as bagging and boosting improve accuracy for weak classifiers.

9. CONCLUSION

Ensemble Classifiers were used by us to improve the accuracy of single classifiers. We were able to achieve more than 92% accuracy of classification. No significant improvement was shown by the application of bagging, boosting and random forest. Voting was able to drastically improve the accuracy as compared to Naïve Bayes classifier. The performance of SVM was way better than the other two classifiers.

Also the rules used in Apriori could be used to suggest the user the cars which are apt for them. This could be done by combining the frequent item sets which are generated by applying Apriori and the choices of buyers similar to the user in question.

10. FUTURE WORK

We can further work on this research by doing feature engineering on the attributes used in the car evaluation dataset. This could be done by giving some attributes like safety and price much more weightage than others. Also we can apply fuzzification to the whole data by assigning membership values to the classes. Also we can use item-item based collaborative filtering which is much better than person-item based collaborative filtering by performing suitable modifications in the dataset. In addition to these we can also create some methods to reduce anomalies encountered during the classification.

11. REFERENCES

[1] Abaya S.: Association Rule Mining based on Apriori Algorithm in Minimizing Candidate Generation. International Journal of Scientific & Engineering Research, Volume 3, Issue 7, July-2012.

[2] Breiman, L.: Bagging predictors, Machine Learning 24 (2), pp. 123-140. (1996)

[3] Bohanec M. and Rajkovic V.: Knowledge acquisition and explanation for multi-attribute decision making. In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. pages 59-78, 1988.

[4] Claesen M., Smet F., Suykens J., Moor B.. EnsembleSVM: A Library for Ensemble Learning Using Support Vector Machines. Journal of Machine Learning Research, pages 141-145, 2014.

[5] Lee J., Sun M., Lebanon G. : A Comparative Study of Collaborative Filtering Algorithms. arXiv:1205.3193v1 [cs.IR] 14 May 2012.

[6] Lizotte D.J. ,Madani O. ,Greiner R. : Budgeted Learning of Naive-Bayes Classifiers. UAI. 2003.

[7] Sebban M.,Nock R. ,Lallich. S.: Stopping Criterion for Boosting-Based Data Reduction Techniques: from Binary to Multiclass Problem. Journal of Machine Learning Research, 3. 2002.

[8] Skurichina M. ,Duin R.P.W.: Bagging, Boosting and the Random Subspace Method for Linear Classifiers. Pattern Analysis & Applications, pages 121-135, 2002.

[9] Ting S.L., Ip W.H., Tsang A.H.C: Is Naïve Bayes a Good Classifier for Document Classification. International Journal of Software Engineering and Its Applications, Vol. 5, No. 3, July, 2011.

[10] Zupan B., Bohanec M., Bratko I., Demsar J.: Machine learning by function decomposition. ICML-97, Nashville, TN. 1997.