International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Improving the Automobile Purchasing Behavior of Customer: Classification Techniques

S.Kavitha, S.Manikandan

Abstract: ***Data mining (DM) is the automate detection of relevant pattern from the database. E-Commerce is a very famous as well as frequently used new technique in the real world applications. DM is an automate detection of relevant patterns from large amount of information repositories. E-Commerce is a Killer-domain for data mining. DM is often a complex process and may require a variety of steps before some results are obtained. To predict behaviors and future trends many tools are available in DM, also allowing the businesses to make proactive pathways for the customer. In this research work, it is taken online shoppers purchasing vehicle data set and find accuracy in terms of its purchasing behavior using some of the classification algorithms. The classification algorithms namely Bayes Net and NavieBayse are utilized for the analysis and a comparative study of both the algorithms are carried out. Finally, the performance of the chosen algorithm is suggested for analyzing the vehicle data set based on the purchasing behavior of the customer and predicts some accuracy.***

Key words: Classification Algorithms, Bayes Net, Naïve Bayes Algorithms.

I. INTRODUCTION

Understanding the needs of customer is challenging task to the company. Also we are not able to predict their buying behavior, one of the way is we can conclude from their past purchases. Company also tries to influence the customer to purchase their products. Psychology of buying process is widely helps to improvise your businesses. Try to involve the customer for the checking the quality, fixing price, because they are the potential customers, so we must retain the customers, so keep in touch with always, sometime announcing incentives, gifts gift voucher for the occasion periods. Invite the selective customers into plant for visiting the product manufacturing process which is one of the way of advertising products. The first purchase is not the last one, so bringing them into long term relationship. Jayendra Sinha (USA), Jiyeon Kim (USA) [1], the results of this study give the insight idea of online retailing in India – specifically factors affecting Indian consumersonline buying behavior. Although the convenience risk seemed to be the only factor significantly affecting Indian consumers online purchases, when looking at male and female perceptions, there were different factors affecting male and female consumers behaviors. Perceived risk is significant for male but not for female. Dr. Sankar RajaGopal[2] identified the high-profit, high-value and low-risk customers via the data mining techniques, customer clustering has been studied using IBM Intelligent Miner. Aditya Kumar Gupta &ChakitGupta[5], the main contribution of their paper is lied in the focusing important issues how to improve decision making and to optimize relationships with Customer in highlycustomer based businesses. This study identifies major problems of customer behavior and that has direct impact on the sale and production. The rest of the paper is structured as follows. Section II discusses about the data set and methods used for this research work. The experimental results are explained in section III. Finally, section IV concludes the research work.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author

S.Kavitha, Research Scholar, Research & Development Centre, Bharathiar University, Coimbatore, TamilNadu, India.

S.Manikandan, Research Supervisor, Prof & Head, Department of CSE,Sriram Engineering College, Chennai, TamilNadu, India.

II. MATERIALS AND METHODS

Classification which means that estimation of associations or we called as separation or ordering of objects into classes. This word is difficult to design precisely. Let we see the brief introduction about the various algorithm.

a) Naïve Bayesian Classification:

The following Figure1represents the structure of a Naive Bayes classifier, it is quite different from decision approach, in a Naïve Bayesian classification has a hypothesis it classified that the given data belongs to class. This model use a link to directed from input to output, it shows the model's simplicity, it doesn't have any interaction between inputs, but it has indirect input via output. The directed link from output to input is not necessary for all Bayesian Network model.

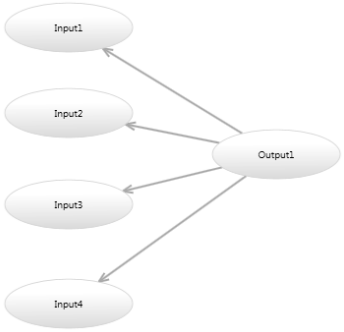

Figure1: Naive Baised classifier

NAIVE based method is based on the work of Thomas Bayes (1702-1761).Bayes was a British Minister and his theory was published only after his death. This theorem has some notations which is defined as, P(A) is the possibility the event A will occur and P(A/B) means that the probability that event A will happen when the event B already happened.The Bayes theorem is

P(A/B) = P(B/A)P(A)P(B)

In this model the probability belongs to any one of the classes namely C1, C2, C3 and etc. using the probability P(Ci/X), this probability helps to compute all the three classes, however, we assign X to the class that have been higher order probability. The P(Ci/X) will be calculated by the below specified formulas,

P(Ci/X) = [P(X/Ci)/P(Ci)]/P(X) (1)

P(Ci/X) is a probability of the object X belongs to class Ci.

P(X/Ci) is probability of getting attribute value X, its belongs to Ci.

P(Ci)is a probability of any object which belongs to a class Ci without any other information.

b) Bayesian Belief Network:

Bayesian networks are widely used visualize the probabilistic for a model of a particular domain with the following advantages. This network provides a simple way of using Bayes theorem into complex problem. To design a Bayesian network require three things such as random variables, conditional relationships and probability distributions. Bayesian network works like a probability theory and does not have a structure of black box which allows a rich in structure. This method helps us to mix the expert opinion and data to build a right model and also support for missing data while learning and classification.

A Bayesian networks is used to define with help of two components such as the first one would be a directed acyclic group and the second one is a set of conditional probability tables. It simplifies the computation, and it is most accurate when compare with other classifiers.

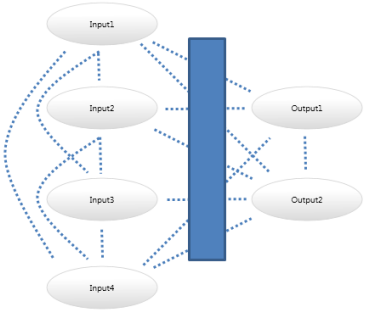

Figure2: Bayesian Belief Network classifier

Figure 2 shows the Bayesian Network with possible structure, its an another method of classification. In this figure the dotted lines depicts the potential links and the blue box tells us to add additional nodes and the links can be added simultaneously, the links usually represent between the input and output nodes. This network shows some scenario of learning and training tuples. The topology of network is given in advance or inferred from data. The variables of network can be observable or hidden in the training tuples. In this case the hidden data is referred to as missing values or incomplete data.

c) Statistical Measures

To calculate the different measures by using different formulas. Wish to calculate the proportion of the predicted positive values, should use the following formulas. TP is used by precision P which means that True Positive likely FP which is used for False positive rate, the formulas for TP and FP are,

$$P(\text{Precision}) = \frac{TP}{TP + FP} \quad (1)$$

Using this equation we can find the True Positive Rate (TPR), Recall or Sensitivity and also tends to correctly identified the positive proposition which is calculated by,

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

Here, FN is False Negative Rate.

This work has been taken three measures likely classified instances, incorrect classified instances and accuracy, in first one instances are correctly classified, The second measure means that are not correctly classified and third one Accuracy means that how many percentages are correctly classified, The incorrectly instances may consists some unnecessary data such as noisy data, in consistency data and out of scope data.Accuracy value is calculated by measured value that should be closed to true value

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

To calculate Accuracy (the total number correct predictions) with True Negative (TN). To find the sensitivity based on the percentage of positive records correctly classified out of all positive records.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

Out of all positive records to calculate the percentage of positive records by Specificity.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

To get the average if information with Retrieval precision and Recall metrics using F- Measure,

$$F = \frac{2 * \text{Re}call * \text{Pr}ecision}{\text{Pr}ecision + \text{Re}call} \quad (6)$$

Using two sets of categorized data to find the degree agreement by Kappa Statistics, which produced the different results that lies between 0 and 1 interval. If the value of result is high means that the Kappa is Stronger. The chance of agreement has normalized value.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (7)$$

Here P(A) is a percentage of agreement and P(E) ps a chance of agreement. If K=1, the agreement is perfect between the classifier and ground truth value. If K=0, there is a chance of agreement and Mean absolute error(MAE), MAE measures the prediction of eventual outcomes.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|fi - yi| = \frac{1}{n}\sum_{i=1}^{n}|ei| \quad (8)$$

The average of Absolute Error is a Mean Absolute Error, the average error e_i = | f_i - y_i |, in this formula where f_i is prediction and y_i is true value.

The square root of mean square value is referred as the Mean Squared Error, it squares the errors before they are averaged, and RMSE will gives the relatively high weight with large errors, The RMSE of Ei is measured by

$$E_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(P(i,j) - T_j)^2} \quad (9)$$

Where P(i, j) is referred to as the value which is predicted by the individual program and i is fitness case, likely T_j is the target value for fitness case j. Receiver Operator characteristic(ROC) curve is used to define the performance of the binary classification algorithm, this curve randomly choose the positive instance which is ranked above randomly chosen negative one. Relative Absolute Error is the total absolute error maderelative to what the error wouldhave been if the prediction simply had been the average of the actual values. Mathematically, the relative absolute error E_i of an individual program is evaluated by the Equation:

$$E_i = \sum_{j=1}^{n}|Pi_j - T_j| \Big/ \sum_{j=1}^{n}|T_j - T''_j| \quad (10)$$

Where $P_{ij}$is the value predicted by the individual program i and for sample case j (out of n sample cases) and T_{ji}isthe target value for sample case j and $T_i$is given by the below formula: $Ti = \frac{1}{n}\sum_{j=1}^{n} T_j$ (11)

Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted values.

III. EXPERIMENTAL RESULTS

The customers purchasing automobile data set having 20 attributes and 9134 instances. Finally concluded from the results of classificationthe NaïveBaised was classified better than the Byes Belief Network. Bayes' theorem assumes that all attributes are independent and that the training sample is good sample to estimate probabilities. These assumptions are not always true in practice, as attributes are often correlated but in spite of this the Naïve Bayes method performs reasonably well. Other techniques have been designed to overcome this limitation. One approach is to use Bayesian networks that combine Bayesian reasoning with casual relationships with attributes. In this data set I was taken 6 attributes from the vehicle purchase database, these are the specific attributes purchaser those who are from rural, suburban, urban, and employment status namely

Table 1: Data Set Descriptions

S.No	Descriptions	Values
1	Gender	Female/Male
2	Location	Rural/Suburban/Urban
3	Employment Status	Employed/Unemployed
4	Marital Status	Single/Married/Divorce
5	Vehicle Class	Two door/four door/SUV/Luxury SUV/Sports Car/Luxury Car
6	Vehicle Size	Small/Med size/Large

In this data set I was taken 6 attributes from the vehicle purchase database, these are the specific attributes purchaser those who are from rural, suburban, urban, and employment status namely

Employed and unemployed, the major three classes of the data set is small, medium, large vehicle size.

Table 2: Results of two measures

Algorithms	TP Rate	FP Rate	Precision	Recall	F - Measure	ROC Area	Class
	0.848	0.777	0.721	0.848	0.78	0.604	Med size
	0.236	0.154	0.268	0.236	0.251	0.632	Small
	0.011	0.001	0.455	0.011	0.021	0.54	Large
Bayes Net	0.643	0.577	0.606	0.643	0.599	0.603	Weighted Average
	0.956	0.951	0.704	0.956	0.811	0.593	Med size
	0.046	0.041	0.211	0.046	0.075	0.626	Small
	0.003	0.004	0.086	0.003	0.006	0.547	Large
Naves Bayes	0.681	0.677	0.545	0.681	0.586	0.595	Weighted Average

Total Number of Instance : 9134

Table3. Error Reports

Statistics	Bayes Net	Navie Bayesd
Kappa Statistic	0.0738	0.0054
Mean Absolute Error	0.337	0.3227
Root Mean Squared Error	0.4022	0.3948
Relative Absolute Error	110.52%	105.83%
Root Relative Squared Error	103.00%	101.10%

Table 4: Performance of Accuracy

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
Bayes Net	64.34%	35.66%
Naive Bayes	68.13%	31.87%

The accuracy is calculated based on addition of true positive and true negative followed by the division of all possibilities. Accuracy is measured and created using 10 fold cross validation method. Tenfold cross-validation is the standard way of measuring the error rate of a learning scheme on a particular dataset; for reliable results, 10 times the data set is executed by 10-fold cross-validation. In 10-fold cross validation method, randomly taken the data sets and divided in ten partitions with equal size.From the partition of nine sets used as training set and the remaining one is used as a test set. The performance is evaluated by comparing theMean Absolute Error, Root Mean Squared error, Receiver Operating Characteristic (ROC) Area and Kappa statistics. When we take the large data test sets this method provide a assessment of classifier with good results and also provide a poor result. The Table 4 depicts the error values of two classification algorithm. The classification of customers data is taken from the given training data set, Table 4 depicts accuracy which is calculated by classification algorithm. The accuracy of Bayes Net algorithm found to be 64.34%, Naive Bayes is 68.13%, and various measure results are depicted in Table 2. From those classification Med size car preferred 70.33% of people, small size car 18.98% of them preferred, the rest of large size car preferred 10.35% of people.

IV. CONCLUSIONS

In this research I used the classification techniques such as Naïve Bayes method, to separating the objects into classes. It was analyzed by conducting the experiments using the marketing analysis data set. The performance is known by the indicators such as accuracy, specificity, precision, and error rate. The data preprocessing method filter the missing data and noise data. For processing I used Naïve Bayes method. The data preprocessing method can improve the accuracy of the classifier because it removes the noise data or the missing values. To improve the overall accuracy use large data set and increase the number of cross fold validation.

REFERENCES

1. JayendraSinha (USA), Jiyeon Kim (USA) “Factors affecting Indian consumers, online buying behavior, Innovative Marketing”, Volume 8, Issue 2, 2012
2. Dr. SankarRajagopal, Enterprise DW/BI Consultant ,Tata Consultancy Services, Newark, DE, USA, “Customer Data Clustering Using Data Mining Technique”, International Journal Of Database Management Systems (Ijdms) Vol.3, No.4, November 2011
3. R.Deivaveeralakshmi, “A study on online shopping behaviour of customers”, International journal of scientific research and management (ijsrm)ISSN (e): 2321-3418
4. E.W.T. Ngai , Li Xiu , D.C.K. Chau,“Application of data mining techniques in customer relationship management: journal homepage:”www.elsevier.com/locate/eswa.

5. Aditya Kumar Gupta &Chakit Gupta,"Analyzing customer behavior using data mining Techniques: optimizing relationships with customer" International Journal Of Management Insight Vol. VI, No. 1; June, 2010
6. Krishna R.Kashwan, Member of IACSIT, and C.M.Velu, "Customer Segmentation using Data mining Techniques" Vol.5, No 6, December 2013.
7. DattatrayV.Bhate, M.Yaseen Pasha, "Analysing target customer behavior using datamining techniques for e-com."
8. N.R.SrinivasaRaghavan, "Data mining in e-commerce", Sadhana, vol 30, No 2, 2005.
9. Mohammad Ali Farajian,ShahriarMohammadi, "Mining the Banking Customer Behavior Using Clustering and Association Rules Methods", International Journal of Industrial Engineering & Production Research, December 2010, Volume 21, Number 4 pp . 239 245.
10. BelsareSatish and Patil Sunil, "Study and Evaluation of user's behavior in e-commerce Using Data Mining", www.isca.in.

AUTHORS PROFILE

S. Kavitha, Assistant Professor, Department of Computer Applications, D.G.Vaishnav College, Chennai, India, doing her Research in the domain of Data Mining, and Published 6 papers in National Conference, 1 paper at International Conference. Area of interest is Data mining, Computer Networks and Operation Research.

Dr. S. Manikandan M.E., Ph.D. is the Professor and Head of the Department of Computer Science and Engineering at Sriram Engineering College, Chennai. He has 20 Years of teaching experience. He has organized and conducted many workshops, seminars and conferences. He has written three books on Software Project Management, Professional Ethics and Data mining and Data Warehousing. He received Best Paper Presenter Award from Computer Society of India, Best Faculty Award from IIR Chennai and Best Administrator Award from Dr. Kalam Educational Trust, Chennai. His research areas of expertise are Image Processing, Data Mining and Computer Networks. He has published 76 papers in International Journals and Conferences. He guided 5 Ph.D. scholars and is guiding 7 more Ph.D. scholars.