ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Application of Bayesian Networks for Customer Behaviour Analysis

Heena Timani [1], Dr. Mayuri Pandya[2]

Assistant Professor, School of Computer Studies, Ahmedabad University, Ahmedabad, Gujrat-India[1]

Head and Professor of Statistics Department, Maharaja Krishnakumar Sinhji Bhavnagar University, Bhavnagar, Gujrat-India [2]

ABSTRACT: In the digital age many business generate and collect huge amount of data in relatively short period of time with the help of powerful computers and resourceful statistical tools. Sensitive information from data can be extracted related to customer characteristics and their purchase pattern. The explosive growth of data requires a powerful computing tool to extract knowledge from the data. Data mining can help to discover complex relationship among various factors that can help in managerial decision making. Bayesian network are being applied for a range of problems in a variety of domains. Successful applications are being realized for example medical diagnosis, for traffic prediction, for technical trouble shooting and information retrieval. The main task of learning Bayesian Network from data is to find automatically directed edges between the nodes to identify network structure model that can be best describe the causality. To transform India into a digitally empowered society and knowledge economy, Digital India program is one of the flagship programs of government of India. "Cashless, Faceless, Paperless," is one of apparent role of digital India. To check the effect of demonetisation in India how people are moving towards cashless, we have conducted online survey through Google form in Ahmedabad City. Data generated through this survey is used in this paper. Data analysis is done using Bayesian approach to find categorical information from the data. For exploratory data analysis we have used R programming. We have measured the performance of various Generative Bayesian classification algorithm Bayesnet, Naive Bayes, WAODE, AODE, HNB and discriminative Classification Logistic regression and Simple regression function using in WEKA 3.6.2 open source software. We have compared various measures Kappa statistics, Receiver Operating Curve, Root Mean square error, Sensitivity, Specificity. While performing algorithms time taken by Generative Bayesian classification algorithm is less then discriminative Classification.

KEYWORDS: Bayesian Network, Classification, Data Analysis, Online shopping Customer Behaviour, Demonetization

i. INTRODUCTION

Data mining can be useful for many real word problem related to business. It can potentially improve decision making using knowledge discovery from data. Bayes decision theory is the accepted way to apply for classification. In classifier design task the importance of designing is so that it classifies correctly a given categorical variable correctly for a given attribute. It is straightforward to use Bayesian networks as classifier (Friedman and Goldszmit (1996) Friedman et al(1997) here the interest is generally on conditional probability densities of a particular random variable called class given that data on some of the other variables called attributes that are statistically related. Advantage of Bayesian network over the other classification of different regression models is that classification can be done with partial observation attribute. Further often regression models assume mutual independence covariates (features variables) but Bayesian network do not. Therefore we encourage applications of Bayesian network for classification task. In fact there is a lot of application of Bayesian network classifiers in various subject domains including business and decision making.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

ii. RELATED WORK

The current trend on consumer behaviour analysis has been recognized as the business problem rather than information technology (Hsieh and Chu, 2009).Analysis of customer behaviour is a costly execution of information technology. It requires business knowledge and detailed planning for successful adoption.
To understand Bayesian networks we have to model a situation in which causality plays an important role. Under uncertainty and incomplete information we can describe things probabilistically (Charniak, 1991).Bayesian networks provide a method for decomposing a probability distribution into a set of local distributions. The network characteristics specifies how to combine these local distributions to obtain the complete joint-probability over all the random variables represented by the nodes in the network model, making it an effective tool for solving classification and prediction problems (Haddawy, 1999).
In this study we have applied Bayesian network technique to analyze the relationships among customer online shopping behaviour. Bayesian Network is generated through machine learning tool WEKA Using Bayesian Classification techniques. We found association of variables using various cliques between variables. It Produce supportive probabilistic arguments for the final conclusion. A set of useful findings has been obtained for customer online shopping behaviour related to demographic factor. Bayesian classification are also used for pattern recognition. We have obtained pattern from the categorical data.
.

iii. AIM AND OBJECTIVE

Converting India into less-cash society and promoting cashless transactions various modes of digital payments are increasing. The effect of Internet and E-wallet facilities are also increasing day by day. India will have 420 million mobile internet users by June 2017. There has been a year-on-year growth of 15 per cent in mobile internet users between October 2015-2016, said the report by the Internet and Mobile Association of India (IAMAI) and IMRB. This affects the online shopping behaviour in India.
To study the impact of demonetization on demographic and other factors of on-line shopping behaviour of consumers of Ahmedabad City a survey was conducted.
To study the effect of demonetisation / cashless economy in last six months, on online shopping online Air/Rail ticket booking, on line bill payment, recharge of various devices. To understand behaviour and attitude of customer towards online shopping after demonetization

iv. DATA DESCRIPTION

Data analysis of various categories of demographic and other factors like age, profession, online purchase products, use internet daily. Different financial services and variables impacting on-line shopping behaviour of consumers are also analysed. Data was generated through Google questioners. For online survey mail was send to 2000 people. The sample was taken from assuming high internet diffusion rate. From 2000 people 660 did not reply, 600 did not use any online shopping, 140 did not use online payment mode, 400 people answer the online questioner properly. The sample comprised of people who have done online shopping or used any other financial transaction through debit/credit card, E-wallet, Pay tm during last six months. Only those customers who do online payment using internet is part of the sample. Final sample for data analysis comprised of 400 instances and Attributes 15.
Data analysis is performed using Bayesian approach to extract knowledge from data and to identify causal relation between variables or attributes

v. METHODOLOGY

For classification task each data object has its origin the process of a given class with a certain probability and each process of a class generates objects with a certain probability called prior probability. To decide which class is to be predicted for a given object, it is necessary to determine the probability called the posterior probability of the object. It

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

describes the probability that an object has its origin in that particular class. To determine the posterior probability, the rule of Bayes is used. In Bayesian classifier training is very fast. Also the model designed is simple and intuitive. In Bayesian, classifier error is minimized subject to the assumptions of independence of attributes and data satisfying distribution model.

Aggregating One-Dependence Estimators (AODE) provide highly accurate classification. It is alternatively used inplace of navie bayes models and averaging over all of a small space. Independence assumptions are weaker then navie bayes. It is computationally efficient and provides high accurate classification on many learning task.

The main hurdle to leveraging the existing classification methods is that these assume record data with a fixed number of attributes. In general, people dealing with sequence data use to convert the data into non-sequential data and then apply traditional classification algorithm. Commonly used classification algorithms are decision trees, k nearest neighbors, support vector machines, and bayes classifiers. Here, we have used Bayesian classification algorithms. Though the methods to achieve classification vary based on the dataset and the model used, all the classifiers have something in common. The commonality is that they divide the given object space into disjunctive sections that are mapped to a given class. We have also applied Bayesian classification function to find the better classification accuracy.

vi. DATA ANALYSIS AND RESULT DISCUSSION

Exploratory Data Analysis and Data Visualization Using R programming

Figure 1 Attributes of Data

```
>   head(On.line.shopping.behaviour)
       Age   Reporter Gender                                Purpose
1 15 to 20    Student   Male                       Online Shopping
2 21 to 25   Director Female                       Online Shopping
3 26 to 30 Contractor   Male                       Online Shopping
4 31 to 35    Builder   Male Online Recharge (Mobile or other)
5 36 to 35     Doctor Female Online Recharge (Mobile or other)
6 21 to 25  Jounalist Female                 Online Bill payment
                      Time Location WebsiteEShop            PurchaseItem
1 Morning 8:00 to Noon 12:00     Home     Flipkart Clothing and accessories
2 Morning 8:00 to Noon 12:00   Office         Ebay                   Bookss
3 Morning 8:00 to Noon 12:00  College       Jabong       Electronic items
4                             Other       Amazone                Jewellery
5 Morning 8:00 to Noon 12:00     Home        Tradus     Computer peripherals
6         Night 12:00 to 4:00   Office FutureBazaar                    gifts
```

In the above figure 1 the sample data is displayed using R Programming. Here first six sample attribute values are displayed. It shows attribute Age, Reporter, Gender, Purpose of Shopping, Time span, Location ,Website, purchased item category wise.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Figure 2 Summaries and Structure of Data

```
> names(`online.Shopping.behaviour..(2)`)
 [1] "Age"           "Reporter"      "Gender"        "Purpose"       "Time"          "Location"
 [7] "WebsiteEShop"  "PurchaseItem"  "Income"        "Frequency"
> summary(`online.Shopping.behaviour..(2)`)
       Age               Reporter      Gender                                  Purpose
 15 to 20:64   Pharmacist    : 12   Female:192   Online Bill payment              : 56
 21 to 25:80   Secretary     : 12   Male  :208   Online Recharge (Mobile or other): 48
 26 to 30:80   Biologist     : 10                Online Shopping                  :212
 31 to 35:80   Builder       : 10                Online study material            : 44
 36 to 35:16   Manager       : 10                Online Ticket Booksing           : 40
 45 to 50:16   Media Planner: 10
 50 above:64   (Other)       :336
                    Time         Location        WebsiteEShop                    PurchaseItem
 Morning 8:00 to Noon 12:00:132   College: 64   Flipkart     :66   Clothing and accessories:74
 Noon 12:00 to 4:00         : 78   Home   :112   Jabong       :60   Bookss                   :66
 Evening 4:00 to 8:00       : 67   Office :128   E-bay        :57   Electronic items         :62
 Night 8:00 to 12:00        : 56   Other  : 96   beststylish :56   Jewellery                :62
 Night 12:00 to 4:00        : 10                 Amazone      :46   Computer peripherals     :60
```

```
> str(online.Shopping.behaviour.)
'data.frame':   400 obs. of  10 variables:
 $ Age         : Factor w/ 7 levels "15 to 20","21 to 25",..: 1 2 3 4 5 2 6 7 3 2 ...
 $ Reporter    : Factor w/ 84 levels "Accountant","Advertising Manager",..: 76 25 19 8 26
 42 52 20 46 55 ...
 $ Gender      : Factor w/ 2 levels "Female","Male": 2 1 2 2 1 1 1 1 2 2 ...
 $ Purpose     : Factor w/ 5 levels "Online Bill payment",..: 3 3 3 2 2 1 1 1 1 1 ...
 $ Time        : Factor w/ 20 levels "","After 3:00 (midnight)",..: 12 12 12 1 12 15 9 19
 6 16 ...
 $ Location    : Factor w/ 4 levels "College","Home",..: 2 3 1 4 2 3 1 4 2 3 ...
 $ WebsiteEShop: Factor w/ 18 levels "Amazone","beststylish",..: 7 4 11 1 17 8 9 18 14 15
 ...
 $ PurchaseItem: Factor w/ 10 levels "Bookss","Clothing and accessories",..: 2 1 5 8 3 7
4 10 6 1 ...
 $ Income      : Factor w/ 7 levels "Rs 10,000 to 15,000",..: 1 4 7 3 7 1 7 3 7 4 ...
 $ Frequency   : Factor w/ 4 levels "1 - 4 times",..: 1 3 2 2 4 2 1 2 3 2 ...
```

In the above figure 2 the data frame of 400 observations and 10 variables sample is displayed. Summary of various categorical variables and structure of the data shown in figure 2. These six product categories were identified from the exploratory study which comprises of Airline/train reservations, Banking & other financial services, Electronics/Mobile phones, Books/Magazines/ Software/Hardware/DVD/CD, Dresses/Apparels/ Footwear/ Jewelers /cosmetics. Mostly all the age group who had good experience of online purchase answer very well about online purchase experience they use to purchase generally branded products, electronic items , CD,DVD, reading material, computer, online banking and on line ticket booking. There is strong relationship between age, gender and online shopping frequency is obtained.

There are various classification algorithms are available in WEKA, Open source software for Bayesian classification WAODE, AODE, HNB, NaiveBayes, Bayesnet. We have compared all the algorithm doe accuracy purpose AODE achieves highly accurate classification by averaging over all of a small space of alternative naive-Bayes-like models that have weaker.From the AODE algorithm prior probability of how frequently online purchase is done by a customer in last six months is obtained. Majority of the classifier gives 90% above classification accuracy..

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Table No 1 Prior Probability Table Using AODE classifier

Frequency of online shopping	Prior Probability
1 - 4 times	0.2
5 - 8 times	0.16
9 -14 times	0.34
15 -times or above	0.48

Table no (1) shows the priory probability of the selected attribute frequency of shopping for classification using weka AODE classification algorithm.

Table 2 Confusion Matrix of Frequency of online purchase in last six months.

a	b	c	d	On line Purchasing Frequency in last Six Months Using AODE classifier Classifier
80	0	0	0	a = 1 - 4 times
0	64	0	0	b=5 - 8 times
0	0	192	0	c = 15 times or above
0	0	0	64	d =9 - 14 times

Table 2 shows the confusion matrix of online shopping frequency in last six months. Here all the diagonal entry shows the frequency of shopping 80 customer had done 1-4 times shopping in last six months,64 customer had done 5-8 times shopping in last six months192 is the highest number of customer had done online shopping 15 times or more in last six months. All the off diagonal entries are zero it shows 100% correct classification using 10 cross fold AODEsr Classification algorithm of WEKA .

Table 3 Various measures of classification accuracy

Comparison of various Bayesian Classifiers (Using Open Source Weka Software 3.6.2)					
10-fold cross-validation	**WAODE**	**AODE**	**HNB**	**NaiveBayes**	**Bayesnet**
Kappa statistic	0.9628	0.9505	0.9628	0.908	0.988
Mean absolute error	0.006	0.017	0.006	0.0255	0.0199
Root mean squared error	0.0526	0.0713	0.0524	0.1017	0.0982
ROC Area	0.998	0.998	0.998	0.996	0.999
Time taken to build model (Second)	0.02	0.0	0.0	0.0	0.0

Table 3 shows the comparison of various Bayesian classification algorithms. Time taken to build the model is very less it shows the Bayesian classification algorithms perform very fast.

Figure No 3. Comparison of Various Bayesian Classification Algorithm

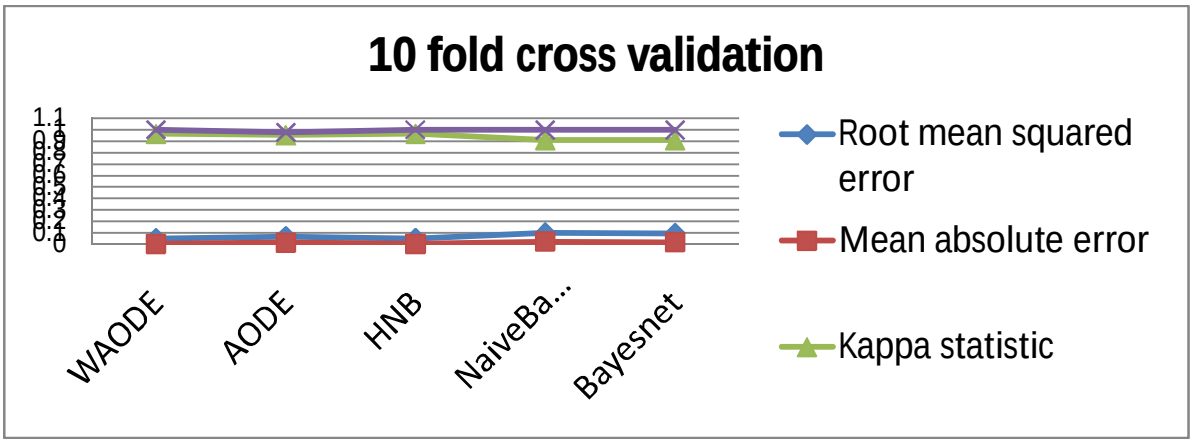
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017



In the above figure 3 we have compared various accuracy measure using different Bayesian classification algorithms**.** The value of Receiver Operating Characteristic curve and Kappa statistic is very near to 1 and Root mean squared error and mean absolute error are almost zero which shows the best classification accuracy.

Table 4 Accuracy Measures of Classification Functions

Comparison of Various Bayesian Classifiers Functions (Using Open Source Weka Software 3.6.2)		
10-fold cross-validation	Logistic–R 1.0E-8 -M -1	Simple Logistic
Kappa statistic	0.8867	0.9813
Mean absolute error	0.014	0.0079
Root mean squared error	0.0923	0.0545
ROC Area	0.999	0.997
Time taken to build model (Second)	15.55	6.11

Table 4 Shows the Measures of accuracy using Bayesian classification functions Logistic regression and simple Logistic here time taken by classification function algorithms are higher than Bayesian classification. Which shows the deterministic approach take more time to perform compared to Bayesian classifier.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Figure No 5 Accuracy Measures of Classification Function
Comparison of Various Classifier Function (Using Open Source WEKA Software 3.6.2)

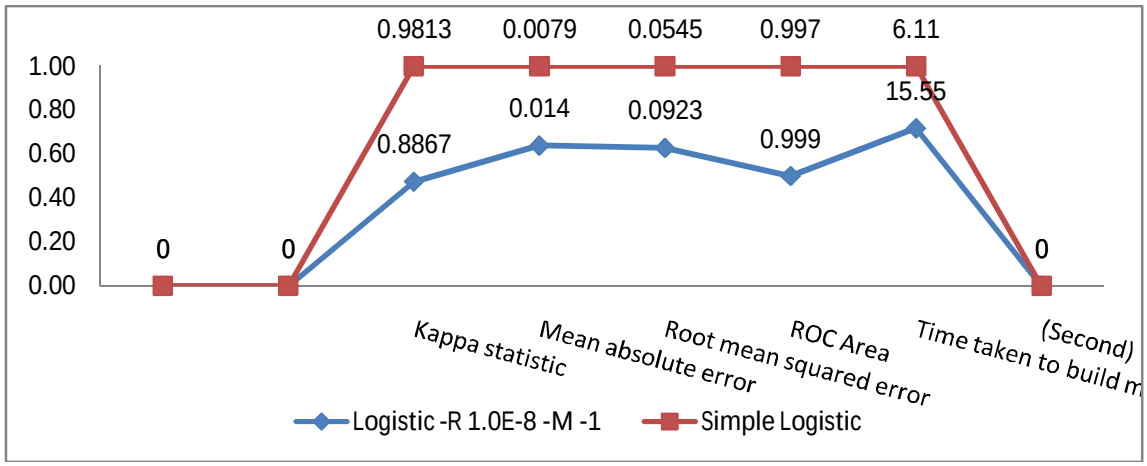
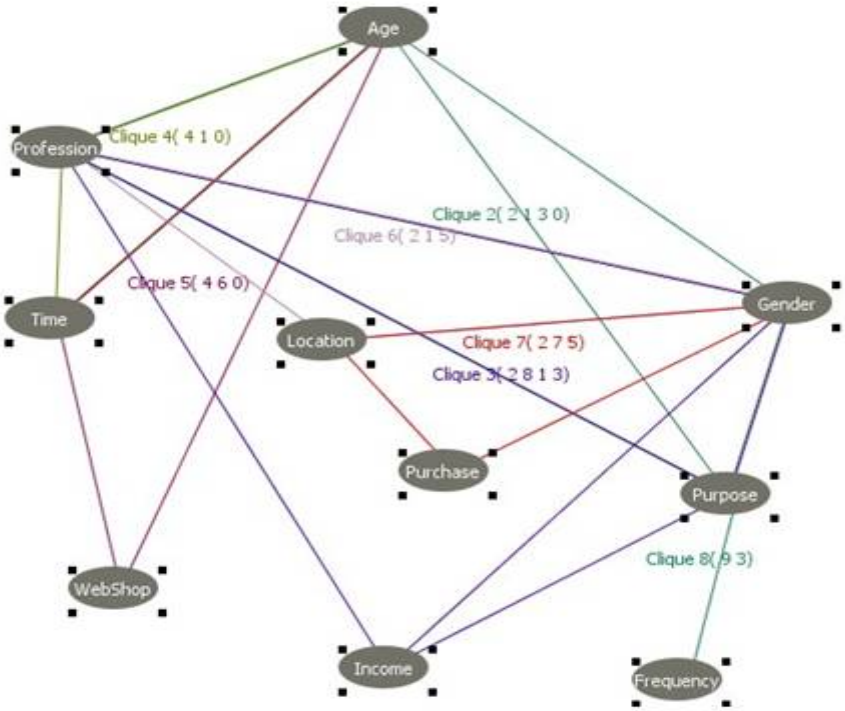


Figure 5 shows the various accuracy measures of classification functions. Simple logistic function perform better than logistic regression function for all the accuracy measures using 10 fold classification.

Figure 6 Generated Bayesian Network Using Open Source WEKA Software 3.6.2



ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

Figure 6 explain the Bayesian Network Structure and association-relationship between various attributes. Here clique tree covers a Bayesian network and from the above Bayesian network we can say that union of the cliques is the set of variables in the Bayesian network and for any variable X in the Bayesian network, there is a clique that contains the variable and all its parents. That clique is called the family cover clique of X. Here total nine cliques are generated and it covers all the attribute of the data and it also shows the association/ relationship between all the attribute in the data. From the Bayesian network we can say that age, gender, profession are highly associated.

vii. CONCLUSION AND FUTURE WORK

There are very less research work on demonetization in India because it is the current issue. We have tried to find the effect of demonetization on online purchase pattern of customer. Because of demonetization cash flow was less and government encouraged various types of digital wallet during this period. Also the effect of free internet provided by Jio Mobile we found the hidden factor.

Using different Bayesian classification algorithms we found that majority of the classifier perform well with more than 90% accuracy under 10 cross fold classification. Bayesnet classifier gives the highest accuracy 99 %. Various results in the form of Confusion Matrix are also obtained. In table of confusion matrix the diagonal entries shows how frequently a customer purchased during the last six months using AODE Classifier. It gives 100% accuracy. All most all the detail accuracy measure like Precision, Recall, F-Measure, ROC are above 95%. From above result we can say that all Bayesian classifier machine algorithms give good accuracy for small data set . We have also used Bayesion function classification algorithm of logistic regression and Simple regression Our study is limited to Ahmedabad City only and we have taken the small sample size for analysis of customer behavior of online shopping. In future we can use the large data to find the online shopping behaviour of customer with various new features.

REFERENCES

[1] Agresti A ,Categorical Data Analysis Wiley, 2nd edition (2002).
[2] Nir Friedman, Dan Geiger, Moises Goldszmidt Bayesian Network Classifiers Machine Learning Kluwer Academic Publishers, Manufactured in The Netherlands. 29,131–163 (1997)
[3] Ben-Gal I, Ruggeri F., Faltin F. & Kenett R. Bayesian Networks, Encyclopedia of Statistics in Quality & Reliability, Wiley and Sons. (2007)
[4] David Heckerman and John S. Breese , Casual Independence for probability Assessment and Inference Using Bayesian Networks, IEEE Transactions on Systems, Man and Cybernetics.(1996)
[5] Dawn E. Holmes Laxmi C. Jain (Eds.) ,Studies in Computational Intelligence 156,Inovations in Bayesian Networks , Theory and Applications, Springer, ISBN 978-3-540-850663-6 (2008)
[6] Heena TImani, Mayuri Pandya, Data Analysis using probabilistic graphical models,4th IIMA International Conference on Advance Data Analysis ,Business Analytics and Intelligence , Conference Proceeding (2015)
[7] Heena Timani, Dhaval Panchal, Arzoo Shah ,Knowledge discovery from music meta data using semantic web,2nd International Conference on Emerging Research In Computing ,Information, Communication and Application ERCICA- Volume 3,Elsevier pp456-461(2014)
[8] Shenoy, Catherine, and Prakash Shenoy, A Bayesian network model of portfolio risk and return. Published in Abu-Mostafa, LeBaron, Lo and Weigend (eds.). *Computational Finance*, MIT Press, pp 87-106. (2000).
[9] Jiawei Han and Micheline Kamber Elsevier publication, Indian Reprint ISBN 978-81-32-05358 pp 282-327 (2011),
[10] M. Nikhita, J. Ashwini, P. Amritha and M.N. Lakshmi, Mining the E-commerce Data to Analyze the Target Customer Behavior International Journal of Industrial Informatics Volume 1, Issue 1, pp-03-06 (2011)
[11] Michael Ashcroft, Bayesian networks in business analytics, Proceedings of the Federated Conference on Computer Science and Information Systems ISBN 978-83-60810-51-4, pp. 955–961 (2012)
[12] Pearl, J. ,Fusion, Propagation and structuring in belief networks, Journal. Artificial Intelligence archive. Volume 29 Issue 3,pp241–288. (1986)
[13] Pearl, J. ,Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, Book, Publishers Inc. San Francisco, CA, USA ISBN:0-934613-73-76(1988)