Identifying Buying Patterns: A Data Mining Approach

Chaitra S N
MCA
Global Institute of Management Sciences,
Bangalore

Ashok M V
Associate Professor,
Global Institute of Management Sciences,
Bangalore

***Abstract*—Analyzing and understanding customer behaviors and characteristics is the foundation of the development of a competitive customer relationship management (CRM) strategy, so as to acquire and retain best customers and maximize customer value. The objectives of this paper is to identify best customers based on monetary value using clustering; to classify the products purchased that contribute to the monetary value identified during clustering, into different categories; and to analyze the buying behavior of the customers. A data mining approach is used. This problem is solved using three phases. In the first phase, K-means algorithm for clustering, decision tree for classification in the second phase and association rules for analyzing the consumer behavior in the third phase are used. Data from a departmental store consisting 1000 samples are collected. Best customers identified was 60 in first phase; customers identified mainly spend on food, groceries and beverages in second phase and customer who buys food items it was found that, he buys groceries as well in the third phase.**

Keywords: CRM, Data mining, K-means, Decision tree, Classification, Clustering, Monetary value, Association rules.

I. INTRODUCTION

Customer data and information technology (IT) tools outline the foundation upon which any successful CRM strategy is built. In addition, the rapid increase of the Internet and its associated technologies has greatly increased the opportunities for marketing and has transformed the way relationships between companies and their customers are managed. Analytical CRM invoke to the analysis of customer characteristics and behavior's so as to support the organization's customer management strategies. Data mining tools are a popular means of analyzing customer data within the analytical CRM framework.

II. PROBLEM STATEMENT

Normally hundreds of customers visit departmental stores daily; over a month, more than a lakh customers visit the store. In order to retain and attract customers, identifying the best customers and their buying behavior is the primary objective. Then classifying the products purchased in to categories is the second objective. Third objective is to identify the hidden patterns among the products purchased.

III. RELATED WORKS

A. The Recency–frequency–monetary analysis model

RFM method is also applied to segment markets, for customer value analysis Kaymak,(2001)[1] and to measure the strength of customer relationship Schijns et al., (1999) [2], R.T. Rustet et al.,2004[3] and is also found to be effective for clustering the TCS. F. Newell et al., (1994)[5]. According to him there are two types of studies of the RFM model. However, Stone et al., (1995)[6] contradicted this and indicated that the three variables have different weights that are dependent on the specific industry.(Hughes et al.,(1994) [4])

K-means approach belongs to one kind of multivariate statistical analysis that cut samples apart into K primitive clusters. This approach or method is especially suitable when the number of observations is more or the data file is enormous .Wu, 2000[15]. K-means method is widely used to segmenting markets. (Kim et al., 2006[16]; Shin & Sohn, 2004 [17]; Jang et al., 2002[18]; Hruschka & Natter, 1999[19]; Leon Bottou et al., 1995 [20]; Vance Fabere et al., 1994[21].

Decision tree is a classification algorithm used as a valuable tool for the description, classification and generalization of data. surveys regarding existing work on decision tree construction were conducted, attempting to identify the important issues involved, directions the work has taken and the current state of the art Sreerama K. Murthy et al., (1998)[7] also advantages of DTC's over single stage classifiers, the subjects of tree structure design, feature selection at each internal node, and decision and search strategies were discussed Safavian, S.R st al., (1991)[9]. Efforts were made to apply and evaluate the algorithm to various domains such as university records, producing human-readable graphs that are useful both for predicting graduation, and understanding factors that lead to graduation Elizabeth Murray et al., (2005)[8],internet shopping mall to detect the change of classification criteria in a dynamically changing environment Jae kyeong kin et al.,2005[10],telecommunication to improve customer retention Luobin, 2007[ ], and mobile commerce to identify the factors influencing customer satisfaction and loyalty Jeewon Choi et al.,2008[ ] ,K L choy, 2006[12]

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository. Discovering association rules is one of the most important task in data mining. Many efficient algorithms have been proposed. Close algorithm by Nicolas Pasquier et al., (1999)[14](Rakesh Agrawal et al., (1993) [22];Rok Rupnik et al.,2007[23]; Akash Rajak et al., 2007[24].In our problem association rules has been applied to recognize customer buying pattern.

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

IV. DATA DESCRIPTION

TABLE I. DATABASE DESCRIPTION

Variables	Description	Possible Values
C_id	Id of the customer	{Text}
Product	Product name	{Text}
I_no	Number of items	{1, 2, 3, 4, 5...}
Discount	Discount in amount for each item	{ 1% - 100% }
Amount	Amount for items the customer bought	{1, 2, 3, 4, 5...}
Date	Date of the bill	{1, 2, 3, 4, 5...}
Bill_no	Bill number	{1, 2, 3, 4, 5...}
Age	Age of the customer	{1, 2, 3, 4, 5...}

a) ***C_Id*** *– ID of the customer. It can take any string values ranging from A-Z, 0-9.*

b) ***Product*** *– represents the name of the product. It can take only text values ranging from A-Z.*

c) ***I_no****:– number of items taken by customer. It can take only the numeric values from 0 to 9.*

d) ***Discount****:– it is the discount given for the each item. It will be represented in percentage i.e., 1% to 100%.*

e) ***Amount****:- Total bill for the items purchased by customer and will be in rupees only.*

f) ***Date****: – date of purchase.*

g) ***Bill_no****:- Bill number generated and can take values from 0 to 9.*

h) ***Age****:- Age of the customer and the values range from 0 to 9.*

V. METHODOLOGY

Step 1: Data Collection

TABLE II. INPUT TABLE

Customer_id	Product	Item_no	Discount	Amount	Date	Bill_no
1	XXX	XXX	2%	90	2-1-2013	23
2	YYY	YYY	2%	780	2-1-2013	43
3	YYY	YYY	3%	3243	2-1-2013	23

This is an extract of the database obtained from the departmental stores with the fields or variables listed above. Data regarding the purchases for one year i.e., from April, 2013 to April, 2014 were collected from the retail stores in Bangalore.

Step 2: Data preprocessing

TABLE III. PREPROCESSED TABLE

Custer_id	Product	Amount	Date
1	XXX	90	23
2	YYY	780	43
3	YYY	3243	23

Preprocessing is done using two techniques

a) ***Chi-square test****: is applied to remove the useless variable that doesn't contribute to the result. From the above table 5.5.2 bill_no, item_no, and discount were removed.*

b) ***Min-max Normalization****: is applied to convert large data represented by RFM variables, to smaller data whose values range between 0 and 1.*

TABLE IV. AFTER NORMALIZATION

Custer_id	Amount	Date
1	0.0120	0.21
2	0.23	0.41
3	0.431	0.21

A. Three phase model

As has been stated above the problem is solved using three phases. The phase 1 is explained below.

Phase 1: Clustering using k-means algorithm**.**

Step 1: Preprocessed table will be the input for k-means.

TABLE V. COMPARISON OF DISTANCE BETWEEN THE CLUSTERS

Cluster	Cluster1	Cluster2	Cluster3	Cluster4
Custer 1	0	0.14443934448524	0.2267298015462	0.35704814361485
Custer 2	0.14443934448524	0	0.082290457060959	0.21260879912961
Custer 3	0.2267298015462	0.082290457060959	0	0.13031834206865
Custer 4	0.35704814361485	0.21260879912961	0.13031834206865	0

Comparison table given above compares the two clusters in terms of distance between them. Cluster 2- cluster 1 =0.144 given in row 1 column 3.Similarly the other values are calculated. This table is the resultant of application of k-means, incrementing value of k in every step by 1.

TABLE VI. CLUSTER DISTANCE TABLE

Cluster	Amount	Date
1cluster	0.12906990130718	0.032243491123228
2cluster	0.31706244963317	0.032243491123228
3cluster	0.59881666300362	0.032243491123228
4cluster	0.98927251985892	0.033813366265825

The first values in the shorter cluster distance field represents the distance between the cluster 1 and 3 similarly the second value viz., 0.357 represents the distance between 1 and 4. The other values in the table can be interpreted similarly.

From the above table it can be observed that, values in the ‘ shorter cluster distance’ attribute starts decreasing by larger extent i.e., from 0.357 to 0.123, after cluster 4..Hence it can be concluded that the maximum number clusters that can be formed is 4.

Phase 2: Classification using decision tree technique

The objective is to classify the elements of the cluster identified in the phase 1.

TABLE VII. INPUT FOR CLASSIFICATION

Number of cluster	The short cluster Distance
Cluster 3	0.2267298015462
Cluster 4	0.35704814361485
Cluster 5	0.123231231233
Cluster 6	0.231231231121
Cluster 7	0.123341324313

The above table is the output of the phase 1 which acts as the input for the phase 2.

Step1: Choosing the cluster

Step one of the decision tree algorithms will find, that cluster, whose values are highest in the fields viz., amount and date.

From the table we can observe that cluster 4 has highest values for amount and date fields viz., 0.98 and 0.033 respectively.

Step 2: Identifying the elements of the cluster.

Step 3: Identifying the category.

TABLE VIII. OUTPUT TABLE FOR DECISION TREE

Item	No.of items	Amount
Fd	73	63022.9
Gro	80	140766.4
B	9	452

The table explains the expenditure of the identified customers in the step2 on the categories mentioned in the table. By observation we can find that the customer has spent more on grocery.

Phase 3: Analyze the buying behavior of the customer using association rules.

TABLE IX. ASSOCIATION RULES FOR ANALYZING BUYING BEHAVIOR

Association rules:
if((age <=20)and(item_type==gro))
if((age <=20)and(item_type==b))
if(((age >20) and (age >35))and(item_type==fd))
if((age >20) and (age >35))and(item_type==gro))
if((age >20) and (age >35))and(item_type==b))
if(((age >35) and (age >45))and(item_type==fd))
if((age >35) and (age >45))and(item_type==gro))
if((age >35) and (age >45))and(item_type==b))
if((age <=20)and(item_type==fd))
if((age <=20)and(item_type==gro))
if((age <=20)and(item_type==b))
if((age >=46)and(item_type==fd))
if((age >=46)and(item_type==gro))
if((age >=46)and(item_type==b))

The association rules written above are self-explanatory. These rules are set to find the relationship in the buying behavior. This gives better results compared to regression methods of finding patterns.

RESULTS

- It is found that the number of best customers identified is 60, obtained by using□clustering.
- After mining 60 best customers, it was found that the customers mainly spend on the categories listed below.
- When customer aged 31 years, buys food items it was found that, he buys groceries as□well.

CONCLUSION

CRM is an ‘‘enterprise approach to understanding and influencing customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability”. The main objective was to identify best customers segments, as this would help retailers in designing new strategies for attracting customers, which was achieved by using K-means algorithm; these best customers were mined to unearth the categories of products, contributing to the monetary value. This led to another important objective our study, i.e., to find the hidden pattern and associations with regard to buying behavior. Association rules were written to find the pattern. This would help the retailer to arrange the associated products next to each other, and hence manage stock. Thus the problem considered was solved in three phases.

BIBLIOGRAPHY

[1] U. Kaymak “Fuzzy target selection using RFM variables”, in: Proceedings of the IFSA World Congress and 20th NAFIPS International Conference, vol. 2, 1038–1043, 2001

[2] M.C. Schijns, G.J. Schroder, ‘Segment selection by relationship strength’, 10 (3) 69– 79, 1996

[3] R.T. Rust, V.A. Zeithaml, K.N. Lemon, “Customer-centered brand management”, Harv. Bus. Rev,1–9, 2004

[4] A.M. Hughes, “Strategic Database Marketing, Probus Publishing Company”, Chicago, 1994,

[5] F. Newell, “The New Rules of Marketing: How to Use One-To-One Relationship Marketing to be the Leader in Your Industry”, McGraw-Hills Companies, New York, 1997.

[6] B. Stone Successful Direct Marketing Methods, NTC Business Books, Lincoln-wood, IL, 1995, 37–57.

[7] Sreerama K. Murthy, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery, 345-389 1998.

[8] Elizabeth Murray, Using Decision Trees to Understand Student Data, Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.

[9] Safavian S.R. , Landgrebe. D ‘A survey of decision tree classifier methodology”, Volume 21, Issue 3, pages 660 – 674.

[10] Jae Kyeong Kim, Hee Seok Song, Tae Seong Kim and Hyea Kyeong Kim,‘Detecting the change of customer behavior based on decision tree analysis’, Expert Systems, Volume 22, Issue 4, pages 193-205, 2005,

[11] Luo Bin, ‘Customer Churn Prediction Based on the

[12] K L Choy, Kenny K.H. Victor Lo, ‘Development of an intelligent customer supplier relationship management system: the application of case based reasoning’,

[13] Industrial management and data system, Vol. 103, Issue 4, pp 263- 274, 2003

[14] Jeewon Choi, Hyeonjoo Seol, Sungjoo Lee, Hyunmyung Cho, Yongtae Park, (2008) "Customer satisfaction factors of mobile commerce in Korea", Internet Research, Vol. 18 Iss: 3, pp.313 - 335

[15] Nicolas Pasquier, Yves Bastide, Rafik Taouil and Lotfi Lakhal, “EFFICIENT MINING OF ASSOCIATION RULES USING CLOSED ITEMSET LATTICES”, Information Systems Vol. 24, No. 1, pp. 25-46, 1999.

[16] Wu, M. L. (2000). Application practices of SPSS statistics. Song-Gun Bookstore Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systemswith Applications, 31(1), 101–107.

[17] Shin, H. W., & Sohn, S. Y. (2004). Product differentiation and market segmentation as alternative marketing strategies. Expert Systems with Applications, 27(1), 27–33.

[18] Jang, S. C., Morrison, A. M. T., & O’Leary, J. T. (2002). Benefit segmentation of Japanese pleasure travelers to the USA and Canada: Selecting target markets based on the profitability and the risk of individual market segment. Tourism Management, 23(4), 367–378.

[19] Hruschka, H., & Natter, M. (1999). Comparing performance of feed forward neural nets and k-means of cluster-based market segmentation. European Journal of Operational Research, 114(3), 346–353.

[20] Leon Bottou, Yoshua Bengio, “Convergence Properties of the K-Means Algorithms”, Advances in Neural Information Processing Systems 7, 1995.

[21] Vance Fabere, “Clustering and the Continuous k-Means Algorithm”, Los Alamos Science, 1994.

[22] Rakesh Agrawal, Tomasz Imielinski, Arun Swam, “Mining Association Rules between Sets of Items in Large Databases”, ACM SIGMOD Record, 1993

[23] Rok Rupnik, Matjaž Kukar, ‘Data Mining Based Decision Support System to Support Association Rules’, Elektrotehniški vestnik 74(4): 195-200, 2007