Predicting Purchasing Behavior of Customer by Analyzing Cluster of Customers

Pragya Shukla, Mohit Jain

Abstract— In today's shopping environment, variations in choices are increasing among customers. Personal and transactional data of customers can be analyzed to derive useful patterns and results on the basis of which marketing strategies are can be decided effectively. The paper presents a new methodology in which different clusters of the customers are created on the basis of demographic attributes of customers. Data mining is applied on the data of clusters to understand the similarity in purchasing behavior of customers of a cluster. In this study a technique of data mining known as association rule mining is used to uncover interesting associations between a cluster of customers and products which are frequently purchased by this cluster. Frequently purchased products are founded by using Apriori algorithm. If frequently purchased products by a cluster are determined than the product choice of target customer who belongs to the same cluster can be predicted with the assumption that the customers having the same demographic backgrounds will have higher probability of having similar choices. These results are helpful for a marketing manager to make effective external influences on the customers and are also helpful to manage inventory.

Index Terms— Customer Behavior, Cluster of Customers, KDD, Association Rule Mining, Apriori Algorithm, Frequent Item Set.

1 INTRODUCTION

IN this Present age of information it is easy to store and manage the transactional and personal information of the customers in database.Applying the technique of data mining on this database can lead us to useful knowledge.

Customers are clustered based on certain demographic data, with an assumption that people of similar demographic backgrounds are expected to behave in a consistent way [2]. Mining the database of the different clusters of the customers will generate the useful patterns and results which can be used to understand the purchasing behavior of the customers more effectively [1].

The study of customer behavior is to know about the product and the services that customers purchase and how these purchases influence their daily life [6]. In trying to understand customer behavior we must identify various factors like external influences, internal processes that will affect product, brand and service purchase decisions [10].

External influences can be divided into two categories:

a. Firm's marketing efforts like product, promotion, price, place etc.
b. Customer's culture like religion, ethnicity, references groups, social class etc.

Internal processes are as:

a. Psychological processes like motivation, perception, attitude, knowledge etc.
b. Decision making like problem recognition, information search, judgment, decision making behavior etc. [7].

Internal processes are responsible for the decision made by the customer and these internal processes can be affected by external influences that are made on the customers. There are different customers with different purchasing behavior and choices so if we can predict the customer's behavior and changes that occurs in the customer's behavior than it can be used to make more effective external influences on a customer and they can customize their promotion and marketing strategies for a particular customer or a cluster of customer [6].

The paper is organized as follows: Section 2 describes the process of Knowledge Discovery in Databases (KDD) and various steps of the process. Section 3 describes the Association rule mining. Section 4 describes Apriori algorithm. Section 5 describes the proposed algorithm for the cluster analysis of the customers. Section 6 provides conclusion of research. Section 7 provides the references.

2 KNOWLEDGE DISCOVERY IN DATABASES

A process known as Knowledge Discovery in Databases (KDD) is defined as a process of extracting the useful knowledge and patterns from large database. Data mining is a analysis step in this process [4]. KDD is an iterative, interactive process [11]. Various steps of KDD process are as follows:

2.1 Goal Identification

The focus of this step is to understand the domain for knowledge discovery. In this step we decide that what are the objectives to be accomplish [11][4].

2.2 Creatiing a Target Data Set

In this step we decide initial data set which is to be analyzed [11][4].

2.3 Data Preprocessing

Raw data is highly susceptible to noise, missing values and inconsistency. Raw data is pre-processed so as to improve the efficiency and case of mining process [11][4][3].

- *Pragya Shukla, Associate Professor in Computer Engineering Department in IET-DAVV, Indore, India, PH- +91 9425082663. E-mail: pragyashukla_iet@yahoo.co.in*
- *Mohit Jain, Rsearch Fellow in Computer Engineering Department in IET-DAVV, Indore, India, PH- +918871259482. E-mail: mohit.jain.121188@gmail.com*

2.4 Data Transformation
Instances and attributes are modified in target data. Normalization, conversion and smoothing of data takes place [11][4].

2.5 Data Mining
Data mining is the analysis step of the KDD process [11][4].

2.6 Interpretation and Evaluation
We examine the output from the previous step whether it is useful and interesting or not [11][4].

2.7 Taking action
If the discovered knowledge is interesting and useful than it can be used to solve the problem [11][4].

3 Association Rule Mining
Association rule mining is a technique of data mining. Its main objective is to find interesting relations among variables, patterns which have higher frequency, associations among sets of items in the transaction databases and data repositories. For example in a shop of online books there are some suggestions after purchasing of some books. These suggestions include some books which are somehow associated with books which are already purchased [8].

An Association rule describes the association among the items. It can be used to show that when some items are purchased in a transaction than others are purchased too. An Association rule can be represented as X→Y here X is antecedent Y is consequent. An Association rule is valid if it satisfies the user defined threshold support and confidence. Antecedent is an item found in data. Consequent is an item found in combination with antecedent [13].

3.1 Support
A transaction T supports an item set I if I is contained in transaction T. Support for an item set I is defined as the ratio of the number of transactions that contain I to the total number of transactions. Let say number of transactions that contain item set I are M And total number of transactions are N than the support S can be calculated as S=M / N [14][9].

A term support count is also used sometimes in place of support. Support count for an item set I can be defined as number of transactions that contain I.

3.2 Frequent Item Set
If support value of an item set I in the transactional database is greater than the specified threshold value of support than the item set is frequent [14][9].

3.3 Confidence
The confidence of association rule X→Y is defined as the ratio of the support for item set XUY to the support for item set X. If XUY is a frequent item set and confidence for the rule X→Y is greater than or equal to threshold value of confidence than we can say that X→Y is a valid association rule [14][9].

4 The Apriori Algorithm
Apriori algorithm is one of the best known algorithms for finding frequent item sets. This algorithm employs an iterative approach known as level wise search. It uses important property called Apriori property is used to reduce the search [7].

All the non-empty subsets of frequent item sets must also be frequent this property belongs to a special category of properties called Antimonotone. If a set can't pass a certain test than all of its supersets will fail the same test this property is called Antimonotone.

There are mainly two operations takes place when we apply Apriori algorithm on a data set. In first operation it generate candidate item set and in second operation it generate frequent item set by removing all infrequent item set. Various terms which are used in Apriori algorithm are D = transaction set, min_sup = predefined minimum support, C_k = candidate item set of size k, L_k = it represent frequent item set of size k [12].

Apriori algorithm:

Algorithm Apriori (D, min_sup)

Input: - D, min_sup

Output: - Generation of all frequent item sets

- L1 = large item set of size 1 generate from the candidate set C1.
- for(k= 2; $L_{k!}$= ∅; k++)
- do
- {
- C_k = Apriori_gen (L_{k-1})
- L_k = (C € C_k | | C.sup> min_sup)
- }
- Answer = L_k
- END

4.1 Illustrating Example
Consider predefined minimum support = 0.5 and predefined minimum confidence = 0.5.

Table 1 represents the transactional data set with item sets A, B, C, D, and E.

TABLE 1
TRANSACTIONAL DATA SET

Transaction ID	Items Bought
001	{A,B,C}
002	{A,C}
003	{A,D}
004	{B,E,F}

Generating frequent item set by using Apriori algorithm on the transactional data set represet by table 1.

Iteration 1:

Generating candidate item set C1 from a transactional data set. Candidate item set C1 is represented by table 2.

TABLE 2
CANDIDATE ITEM SET C1

Items	Support
{A}	0.75
{B}	0.50
{C}	0.50
{D}	0.25
{E}	0.25
{F}	0.25

Generating frequent item set L1 by removing infrequent item set from C1. Frequent item set L1 is represented by table 3.

TABLE 3
FREQUENT ITEM SET L1

Items	Support
{A}	0.75
{B}	0.50
{C}	0.50

Iteration 2:

Generating candidate item set C2 from L1. Candidate item set C2 is represented by table 4.

TABLE 4
CANDIDATE ITEM SET C2

Items	Support
{A,B}	0.25
{B,C}	0.25
{A,C}	0.50

Generating frequent item set L2 by removing infrequent item set from C2. Frequent item set L2 is represented by table 5.

TABLE 5
FREQUENT ITEM SET L2

Items	Support
{A,C}	0.50

Following item sets are frequent {A}, {B}, {C}, {A, C} with support value 0.75, 0.5, 0.5, 0.5 respectively. Now let's check validity of the association rule A→C Support for AUC is 0.5 which is equal to threshold support value and support for A is 0.75 which is above the threshold support value which concludes that item sets AUC and A are frequent.

Confidence value for rule A→C is ratio of support value for AUC and support value for A which is 0.66. Confidence value of the rule A→C is greater than the threshold value of confidence and it can be concluded that the rule A→C is valid.

5 THE PROPOSED ALGORITHM

The Proposed algorithm can be explained in following steps:-

Step1:- Select an attribute or composite attributes on the basis of which cluster of the customers are can be formed.

Step2:- Select all the transactions which are made by these clusters.

Step3:- Apply Apriori algorithm on these selected transactions and obtain the frequent item sets which are frequently purchased products.

Step4:- In the previous step products which are frequently purchased by a cluster is obtained. Customers of this cluster are more interested in purchasing these products. So if there is a target customer who belongs to the same cluster will be more interested in purchasing these products.

5.1 Illustrating Example

The Database contains personal information and transactional information of the customers.Here table 6 represents the personal details of the customers and table 7 represents the transactional records of the customers. The customers are divided into different age groups as follows.

- Group A having age from 11-18 years.
- Group B having age from 19-26 years.
- Group C having age from 27-34 years.
- Group D having age from 35-42 years.
- Group E having age from 43-50 years.
- Group F having age 51 and above.

A cluster named D1 is created in which all the customers which are having age group value B and gender value male are included. We considered the transactions made by this cluster D1 on the day 16/11/2012 which are represented by table 8 and applied Apriori algorithm with assumed minimum support value =.2 and minimum confidence value=.2.

A cluster named D1 is created in which all the customers which are having age group value B and gender value male are included. We considered the transactions made by this cluster D1 on the day 16/11/2012 which are represented by table 8 and applied apriori algorithm with predefined minimum support value =.2. Apriori algorithm uses iterative approach. In this example in the first iteration candidate item set C1 which is represented by table 9 is generated from the transactional data. C1 contains each item set with one element in it. Support value is calculated for each item set of C1. Frequent item set L1 which is represented by table 10 is generated by pruning all infrequent item sets from C1. Infrequent item sets are those item sets which are having their support value less than predefined minimum support value.

In second iteration candidate item set C2 which is represented by table 11 is generated from L1. In C2 each item set contains two elements. Frequent item set L2 which is represented by table 12 is generated by removing all infrequent item sets from C2.

In third iteration candidate item set C3 which is represented

by table 13 is generated from L2. In C3 each item set contains three elements. Frequent item set L3 which is generated by removing infrequent item sets from C3. L3 is represented by table 14. Than no furthur iteration is possible because in L3 there is only one item set so no more candidate item sets are possible.

TABLE 6
CUSTOMER PERSONAL DATA

ID	Name	Age group	G	Occupation	Religion	Phone no.	Address	Marital status	Email address
01	Ashish Sharma	B	M	Student	Hindu	942455612	8-tilakpat rajwada Indore	Single	ashish.sharma@gmail.com
02	Rishab jain	D	M	Business man	Jain	8962266035	Mahidpr (M.P.)	Married	rishab.jain0016@gmail.com
03	Heena khan	B	F	Student	Muslim	9926081320	Khajrana Indore	Single	heena_khan_786@yahoo.com
04	Sanjna Jain	D	F	Housewife	Jain	8962266030	Bhawarkua Indore	Married	Sanjana.1992@gmail.com
05	Ranu Gangwal	B	F	Student	Hindu	9424587143	Shivnest township Indore	Single	ra.gang@gmail.com
06	Pankaj Soni	C	M	Business man	Hindu	8769277976	Sukhlia Indore	Married	soni22@yahoo.com
07	Yamini Gaur	B	F	Student	Hindu	9425510057	Rambagh, Rajwada Indore	Single	yamini94.ss@rediffmail.com
08	Ankit Nema	D	M	Business man	Hindu	9765584510	Siyaganj indore	married	nemaankit45@gmail.com
09	Ashish Sharma	B	M	Student	Hindu	9988772345	Kalani nagar Ind	Single	Shsharma11@yahoo.co.
10	Ankit Jain	B	M	Soft. Eng.	Jain	8976453567	Tilak nagar ind	Single	Jainankit02@gmail.com
11	Ankita Kukreja	B	F	Soft. Eng.	Hindu	8796453627	Rajwada Ind	Single	Kukja.009@gmail.com
12	Avni	B	F	Student	Hindu	9875643276	Shivnest Township Ind	Single	av.cool77@yahoo.com
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…

TABLE 7
CUSTOMER TRANSACTIONAL DATA

ID	Prod_type	Prod_name	Prod_brand	Prod_price /unit in Rs	Date of pur-chase	Time of purchase	Number of pieces
04	Cloth	Jeans	Lewis	1200	16/11/2012	10:30	1
07	Cosmetic	Cold cream	Nivia	45	16/11/2012	11:00	1
09	Cloth	Jeans	John player	1500	16/11/2012	11:15	1
09	Cloth	T shirt	Duke	1100	16/11/2012	11:15	1
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…

TABLE 8
CLUSTER D1 TRANSACTIONAL DATA

Serial no.	Transactions
01	{Cloths, Jeans, John player}
02	{Cloths, T shirt, Duke}
03	{Cosmetic, Hair gel, Setwet}
04	{Cloth, Jeans, John player}
05	{Cosmetic, Deodorant, Setwet }
06	{Cloth, Jeans, John player}
07	{Cosmetic, Hair gel, Setwet}
08	{Cloth, Jeans, John player}
09	{Cloth, Jeans, Koutons}
10	{Cosmetic, Deodorant, Setwet}
11	{Cloth, Jeans, John player}
12	{Cloth, T shirt, Duke}
13	{Cloth, Jeans, Lewis}
14	{Cloth, Jeans, Koutons}
15	{Cosmetic, Deodorant, Rexona}

Applying apriori algorithm on above transactional data

Iteration 1

Generating candidate item set C1 from transactional data of cluster D1. Candidate Item Set C1 is represented by table 9.

TABLE 9
CANDIATE ITEM SET C1

Items	Support
{Cloths}	0.66
{Jeans}	0.53
{John player}	0.33
{T shirt}	0.13
{Duke}	0.13
{Cosmetic}	0.33
{Hair gel}	0.13
{Setwet}	0.26
{Deodorant}	0.20
{Lewis}	0.06
{Koutons}	0.13
{Rexona}	0.06

Generating frequent item set L1 by removing infrequent item sets from candidate item set C1. Frequent item set L1 is represented by table 10.

TABLE 10
FREQUENT ITEM SET L1

Items	Support
{Cloths}	0.66
{Jeans}	0.53
{John player}	0.33
{Cosmetic}	0.33
{Setwet}	0.26
{Deodorant}	0.20

Iteration 2

Generating candidate item set C2 from frequent item set L1.

Candidate item set C2 is represented by table 11.

TABLE 11
CANDIDATE ITEM SET C2

Items	Support
{Cloths, Jeans}	0.53
{Cloths, John player}	0.33
{Cloths, Cosmetic}	0.00
{Cloths, Setwet}	0.00
{Cloths, Deodorant}	0.00
{Jeans, John player}	0.33
{Jeans, Cosmetic}	0.00
{Jeans, Setwet}	0.00
{Jeans, Deodorant}	0.00
{John player, Cosmetic}	0.00
{John player, Setwet}	0.00
{John player, Deodorant}	0.00
{Cosmetic, Setwet}	0.26
{Cosmetic, Deodorant}	0.20
{Setwet, Deodorant}	0.13

Generating frequent item set L2 by removing infrequent item sets from Frequent item set L2 is represented by table 12.

TABLE 12
FREQUENT ITEM SET L2

Items	Support
{Cloths, Jeans}	0.53
{Cloths, John player}	0.33
{Jeans, John player}	0.33
{Cosmetic, Setwet}	0.26
{Cosmetic, Deodorant}	0.20

Iteration 3

Generating candidate item set C3 from frequent item set L2. Candidate item set C3 is represented by table 13.

TABLE 13
CANDIDATE ITEM SET C3

Items	support
{Cloths, Jeans, John player}	0.33
{Cloths, Jeans, Setwet}	0.00
{Cloths, Jeans, Deodorant}	0.00
{Cosmetic, Deodorant, Setwet}	0.13
.	.
.	.
.	.

Generating frequent item set L3 by removing infrequent item

set from candidate item set C3. Frequent item set L3 is represented by table 14.

TABLE 14
FREQUENT ITEM SET C3

Items	Support
{Cloths, Jeans, John player}	0.33

The customers of cluster D1 have more frequently purchased the cloths and among the cloths they purchased Jeans of brand John player more frequently.
Cluster D1 is associated with every transaction of transactional data represented in table 8 so support value for {D1} is 1.00. Confidence value for association rule {D1} → {Cloths, Jeans, John player} is calculated as ratio of support ({D1} U {Cloths, Jeans, John player}) and support ({D1}) which is equal to 0.33.
So {D1} → {Cloths, Jeans, John player} is a valid Association rule because it's confidence value is greater than the predefined threshold value of confidence.

6 CONCLUSION

From the above experiment we can conclude that the male customer having age 19 to 26 are more interested in purchasing cloths and among cloths they like jeans of brand John player so if there are any new discounts or schemes on John player's product than organization can offer these to customers of cluster D1. Thus they can customize their services for a cluster of customers. If there is a target customer who belongs to cluster D1 then there is more probability that he/she will be more interested in purchasing the jeans of brand John player. Thus there are different clusters can be formed and by analysing the purchasing behavior of these clusters purchasing behavior of customer who belongs to these clusters can be predicted.

7 REFERENCES

[1] Chong Wang, Yanqing Wang, Discovering Consumer's Behavior Changes Based on Purchase Sequences. 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012).

[2] Vinay Prasad, Customer Spend Analysis: Unlocking The True Value of a Transaction.

[3] Mu-Chen Chena,*, Ai-Lun Chiub, Hsu-Hwa Changc, Mining changes in customer behavior in retail marketing, Expert Systems with Applications 28 (2005) 773–781.

[4] Usama Fayyad , Gregory Piatetsky - Shapiro , and Padhraic Smyth the KDD Process for Extracting Useful Knowledge from Volumes of Data.

[5] J. Han and M. Kamber, 2000, Chapter 6 Mining Association Rules in Large Databases.

[6] Hayden Noel, consumer behavior, published by AVA publishing SA 2009.

[7] Julianna Katalin Sipos Jiawei Han und Micheline Kamber. Data Mining – Concepts and Techniques. Chapter 5.2.

[8] Qiankun Zhao, Sourav S. Bhowmick, Association Rule Mining, A Survey Nanyang Technological University, Singapore.

[9] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, Mining Association Rules between Sets of Items in Large Databases, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120.

[10] Ray Wright, Consumer Behavior. Thomson Learning, Edition 2006.

[11] Richard J. Roiger, Michael W. Geatz, Data Mining A Tutorial-based Primer.

[12] Rakesh Agrawal, Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules, IBM Almaden Research Center 650 Harry Road, San Jose, CA 95120.

[13] Stephane Lallich, Olivier Teytaud, and Elie Prudhomme, Association rule interestingness: measure and statistical validation.

[14] Ramakrishnan Srikant, Rakesh Agrawal, Mining Generalized Association Rules, IBM Almaden Research Center San Jose, CA 95120 {srikant, ragrawal}@almedon.ibm.com.