Dissection of Customer Purchasing Behavior with Demographics Using Soft Clustering Approach

1Gautham Manohar, 2P.K. Prafulla Premnath and 3L. Nitha

1Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India. gautham.manohar95@gmail.com

2Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India. prafullapremnath4@gmail.com

3Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India. nitha.leeladharan@gmail.com

Abstract

The customers in an area are segmented into various groups to understand their product preference from that particular region. The collected data consists of various factors which are based on demographic characteristic, satisfaction of service and purchasing behavior of the customers. The dissection of customer purchasing behavior is analyzed using frequency and monetary attributes of RFM (Recency, Frequency, Monetary) analysis. Analyzed results help supermarkets to uncover the purchasing behavior of customer. Local supermarkets use this information to establish them in a better manner by fine-tuning their marketing strategies to incorporate the customer’s expectation and behavior. For arriving at this segmentation, we use a Soft Clustering method based on Expectation Maximization (EM) clustering approach. Because of this reason that the hard clustering reveals less promising results; the proposed Soft Clustering can be used more effectively than finite mixture model and we also incorporate a hard clustering method based on K-means. A comparison is made with the soft clustering method for evaluating the performance of these algorithms. Our result shows that K-means Clustering algorithm performs efficiently than EM Clustering algorithm.
1. Introduction

The style of shopping are continually changing in today’s world. Supermarkets need to maintain better relationship with their customers. Understanding the customer’s needs, desires, and behaviors is important in CRM (customer relationship management). Bhatnagar and Ghose [1] proposed a customer service and satisfaction model in which the cost of a product is not only the major factor to purchase a product but also good customer service is very important. That is, segmentation based on their behavior of purchasing, demographic characters and satisfaction with services is crucial for establishing a market plan. This also helps for building and sustaining the marketing strategies and thereby super markets can set their objectives for profit-making.

Segmentation is a technique used to group customers according to their similarity in purchasing behavior. Different market segments may have different product preferences and behavioral patterns. According to Amit, et al. [2], Clustering can be used for analysis and predicting customer product preferences for providing services based on their needs and it also helps in market segmentation. The clustering techniques can be used to divide the customers into segments and the results are evaluated to improve the customer satisfaction and loyalty. The customers are grouped [3] according to the purchasing frequency and money spend during their purchase (RFM analysis) and also by means of customer satisfaction with services. Other information, such as their demographic characteristics can also be accommodated to identify the product preferences and purchasing behavior of each customer groups.

Expectation–Maximization (EM) clustering algorithm is used for dividing the customers into segments. It creates clusters through cross validation and it produces probabilistic clusters by using statistical methods. So, EM is the best method for soft clustering approach. We also describe a hard clustering approach for analyzing the performance of two algorithms: Expectation–Maximization (EM) and K-means clustering algorithm using the same dataset. This paper proposes an effective soft clustering method for predicting the purchasing behavior of the customers and also investigates the performance of two clustering models in data mining.

We organized the paper as follows. In section I the concepts of CRM, RFM and customer segmentation are introduced. In section II literature review is discussed. Section III includes segmentation model and proposed system. In section IV, we present the results and analysis of our experiments. The conclusions are presented in section V.

2. Literature Review
Reference [3] shows Customer relationship management (CRM) is a marketing technique used to resolve quantitatively the outstanding customers through RFM analysis. CRM mainly focuses on the idea of right time, right channel, right price, and the right customers. Reference [4] proposed that RFM model preserves valued customers by predicting the customer purchasing behaviour and it also helps for decision making in the marketing segment. CRM provides opportunities for using information [5] to analyze customers and create customer loyalty. The valued customers are identified for providing basic information about their behavior that is used for deploying the target marketing and more personalized marketing which in turn helps for developing an effective CRM [6].

RFM model is the best model for analyzing customer shopping characteristics and it makes use of point system. The customers are grouped into different segments by scrutinizing the major elements like Recency, Frequency, and Money [RFM] spent while shopping. The scores are given by the customers and that are used for segmenting them. RFM parameters (recency, frequency, monetary) are conformable to measure customer values and for segmenting them in different services areas [7]. RFM analysis is the most important tasks in database marketing. Based on customers’ archival purchasing behavior, RFM analysis can identify the valuable customer group [8]. As in [6] a count item is added beside RFM parameters and they proved that they make no difference to clustering result. So, in this paper, we are collecting our dataset based on RFM analysis.

Customers in the modern retail environments are segmented to examine how customers are using multiple channels and media focusing on different parameters including demographics [9]. To achieve optimal segmentation, the online customers are divided into different clusters based on their purchase data by using a soft clustering method [3]. Referring to their study, we use soft clustering method for generating customer segments. Reference [10] proposed a methodology for foreseeing the purchase probability. Their study focuses on extracting purchasing patterns and predicting probability of purchases. So, we extract purchasing behavior of customer using RFM model and predict the probability of purchase using EM clustering method. The Expectation Maximization (EM) technique is relatively slow while handling huge data, as in each iteration of the dataset they also consider the unseen variables [11]. Partition based clustering is done by means of distance between each data items and model based clustering allocates observations to appropriate clusters by estimating membership probabilities [12]. We use K-means as partition based clustering and EM as model based clustering.

3. Proposed System

The data for analysis is obtained from customer survey questionnaire and is shown in Table 1. The dataset includes customer demographic data such as age, gender and marital status; customer satisfaction with service; customer shopping
behavior like shopping frequency and money spend; and also the type of product they are frequently purchasing. The attributes and its values are represented in Table 2.

Table 2: Customer Survey Attributes

<table>
<thead>
<tr>
<th>DEMOGRAPHICS</th>
<th>AGE</th>
<th>GENDER</th>
<th>MARITAL STATUS</th>
<th>SHOPPING BEHAVIOR</th>
<th>MONEY SPENT</th>
<th>PRODUCT CATEGORY AND SATISFACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-15, 16-20, 21-25, 26-30, 31-35, 36-50, over 50</td>
<td>Male, Female</td>
<td>Single, Married</td>
<td>FREQUENCY 1-2, 3-4, 5-7, 8-10, more than 11</td>
<td>0-2000, 2000-3999, 4000-5999, 6000-7999, 8000-9999, above 10000</td>
<td>SATISFACTION Highly Satisfied, Satisfied, Neutral, Unsatisfied, No opinion</td>
</tr>
</tbody>
</table>

Two major steps are involved in the analysis. Firstly, estimating preferences based on customer behavior and secondly, comparing Soft Clustering and Hard Clustering using EM clustering algorithm and K-Means Clustering algorithm.

Step 1: Estimating Preferences

Data preprocessing and clustering was performed by WEKA (Waikato Environment for Knowledge Analysis) tool. First we group the customers according to their product preferences by using Soft Clustering approach and it is done by EM clustering algorithm. It is iterative in nature and finds maximum likelihood solutions. EM consists of two steps. First, the expectation step assigns members to clusters based on the parameters of probabilistic clusters. Second,
the maximization step identifies new clusters or parameters that maximize the expected likelihood in probabilistic model based clustering. Hence the clusters can be identified efficiently. The clusters are visualized in WEKA. So, we can easily identify the product preferences for each customer category.

Step 2: Comparison of Soft Clustering and Hard Clustering Approaches

Soft clustering and hard clustering are the two approaches involved in the form of Clustering techniques; if the data objects can belong to more than one cluster, it is referred to as a Soft Clustering technique. Otherwise, in hard clustering, the data objects belong to exactly one cluster. Hard clustering is done with the help of K-Means clustering in WEKA. Cluster members are identified by calculating centroid (here product category) for each group and then assign each customers to the group with the closest centroid. The final comparison result includes number of iterations performed by each algorithm, time taken to build the model, and how many incorrectly clustered instances are formed. According to the result of comparison, the efficient algorithm is identified.

4. Experimental Results

In our first experiment, we identify customer segments based on their age group (Fig. 1) and their purchasing behavior that are analyzed using EM soft clustering approaches which is shown in Fig. 2.

![Cluster segments based on age group](image)

Fig. 1: Customer Segments based on Age Group
The above result shows that the customers in the first cluster belonging to an age group of over 50 and are grouped into this cluster. Mostly, male customers visiting the supermarket, 8-10 times are belonging to the first cluster and their product preference is identified as cooking essentials and household supply items; also their Satisfaction attribute is identified as Satisfied. In the second cluster, mostly the female customers belong to an age group of 16-20 and the frequency of visits is less than two times. Their product preferences are identified as beauty care products and the Satisfaction rate is identified as neutral. In third cluster, mostly female customers coming under an age group of 26-30 are grouped and their product preference is cooking essentials. Their frequency of visit is between 5 to 7. In the fourth cluster, male and female customers in an age group of 0-15 are purchasing snacks and beverages, purchasing frequency is 3 to 4 times. Their satisfaction rate is marked as satisfied. In this way, we can easily identify each customer groups and their purchasing behavior.

We identify the product categories that are frequently purchased (Fig. 3) by each customer groups according to their buying frequency. In fig 3, the Y-axis denotes product grouping. Y-axis denotes age cluster of customers and the color of the cluster instances represents the buying frequency of customers. From Fig.3 we identified that more clusters are formed in the cooking essentials category. So from this clusters, we can easily figure out the group of customers belonging to an age group who are purchasing specific products more. For example, in the figure customers coming under the age group of 26-30, 31-35, 36-50 and above 50 are purchasing cooking essentials with the frequency of 5-7 or 8-10 times on monthly basis.
In the second experiment, we compare the Soft Clustering result with the result obtained by using K-means clustering algorithm involved in Hard Clustering approach. The EM clustering result is shown in Fig.4 and K-Means clustering result is shown in Fig.5.

The result (Table 3) shows that through 4 iterations, 9 clusters are formed in Soft Clustering approach. But in Hard Clustering, 4 iterations are performed for obtaining 7 clusters. So, for predicting the purchasing behavior and product preference of customers, we can choose a Soft Clustering approach. Because, it finds maximum likelihood solutions; here, the Log likelihood obtained is $-3.96849$. EM algorithm takes more time to build the model as compared to K-means. And also we identified that the incorrectly clustered instances are greater
in EM algorithm (24%) than K-mean algorithm (21%).

Table 3: Comparison of Two Clustering Approaches

<table>
<thead>
<tr>
<th></th>
<th>Soft Clustering EM Algorithm</th>
<th>Hard Clustering K-Means Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of clusters formed</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Number of iterations performed</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Time taken to build model</td>
<td>1.27 seconds</td>
<td>0 seconds</td>
</tr>
<tr>
<td>Incorrectly clustered instances</td>
<td>51.0 (23.7209 %)</td>
<td>45.0 (20.9302 %)</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-3.96849</td>
<td>NIL</td>
</tr>
<tr>
<td>Sum of squared errors</td>
<td>NIL</td>
<td>139.0</td>
</tr>
</tbody>
</table>

5. Conclusion

Customer satisfaction is vital for the growth of a company and therefore customer characteristics and purchasing behavior plays an important role in its economic aspect. We selected the customer product preference by various factors and the data is obtained through survey. The questionnaire is made by considering customer purchasing behavior (frequency, money spent), demographic characters (age, gender, marital status) and also by using their satisfaction rate. The proposed soft clustering method allows extensible large number of customers and also computationally feasible method for customer segmentation.

Firstly, Soft Clustering method can be used to group customers into different clusters. By analyzing the result, we conclude that it can produce more promising results than Hard Clustering where each customer can only be assigned to one cluster [3]. By using EM clustering algorithm, we grouped the customers and based on the results obtained, their product preferences are identified. Secondly, comparisons of both clustering methods are done. The result of Hard Clustering was obtained by using K-means Clustering algorithm. The comparison results shows that incorrectly clustered instances and the time taken to build the model are greater in Soft Clustering. So, we conclude that Hard Clustering is more efficient than Soft Clustering approach.

References


