Generating Customer Profiles for Retail Stores Using Clustering Techniques

Pramod Prasad, Research Scholar
Department of Computer Science and Engineering
G. H. Raisoni College of Engineering
Nagpur, India
Email: pmpramod@gmail.com

Dr. Latesh G. Malik, Professor
Department of Computer Science and Engineering
G. H. Raisoni College of Engineering
Nagpur, India

Abstract – **The retail industry collects huge amounts of data on sales, customer buying history, goods transportation, consumption, and service. With increased availability and ease of use of modern computing technology and e-commerce, the availability and popularity of such businesses has grown rapidly. Many retail stores have websites where customers can make online purchases. These factors have resulted in increase in the quantity of the data collected. For this reason, the retail industry is a major application area for data mining. This paper elaborates upon the use of the data mining technique of clustering to segment customer profiles for a retail store. Retail data mining can help identify customer buying patterns and behaviours, improve customer service for better customer satisfaction and hence retention.**

Keywords – data mining, clustering, customer segmentation

I. INTRODUCTION

As per a research conducted by Credit Analysis and Research Ltd. (CARE) in March 2011, the Indian retail industry has grown at a Compounded Annual Growth Rate (CAGR) of 13.3% for the period FY06-10. In FY2010-11 alone, this sector has generated revenue of over 12 lakh crore rupees [2]. Changes in consumer shopping habits and emerging technologies are bringing heavy transformation across the retail industry. Consumers are challenging the industry to adapt to the ways they live and shop today. Supported by emerging technologies, consumers have become more focused than ever on price and convenience. Hence, retailers have to be able to very clearly differentiate themselves through excellent customer service that is further enabled through technology. This is all the more important to avoid or reduce customer churn, since the cost of acquiring new customers is much higher as compared to that of retaining them [3]. The key to survive in this competitive industry lies in better understanding of customers. One of the approaches used to understand customers and identify their homogenous groups is customer clustering [1]. In practice, many retail players have adopted the approach of customer clustering to improve their marketing efficiencies and customer service.

The process of grouping a set of objects into classes of similar objects is called *clustering*. In data mining, clustering techniques look to segment the entire data set into relatively homogeneous groups or clusters. The data objects are clustered or grouped based on the principle of maximising the intra-class similarity and minimising the inter-class similarity i.e. clusters are formed so that objects within a cluster have high similarity with respect to one another, but are very dissimilar to objects in other clusters. Clustering techniques analyse data objects without consulting a known class label, unlike classification techniques, that analyse class-labelled data objects. The class labels are not present in the training data, simply because they are not known at the beginning. Clustering is used to generate such class labels. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived.

Although classification is also an effective means for distinguishing classes of objects, it is uses tasks that are often expensive to conduct. It involves collecting and labelling of a large set of training patterns that is used to model each group. Clustering performs the task in the reverse direction; it first partitions the set of data into groups based on similarity and then assigns labels to the comparatively small number of groups. An important advantage of a clustering-based process is that it is adaptable to changes and helps single out useful features that distinguish different groups.

Sections 2 and 3 of this paper describe the clustering techniques of k-means and expectation maximisation and how they can be applied in customer segmentation. Section 4 details our use of the Weka data mining tool for generating clusters from a sample dataset.

II. K-Means Clustering

Cluster analysis is an important human activity. Through automated clustering, dense and sparse regions of an object space can be identified thereby leading to the discovery of overall distribution patterns and interesting connections among data attributes. An important business application of clustering is to help organisations discover distinct groups in their customer bases and characterise customer groups based on purchasing patterns. Such clustering is also called *segmentation* since large data sets are partitioned into groups as per their similarity. Cluster analysis techniques mainly focus on *distance based methods*; the prevalent being the iterative partitioning method of *k-means clustering*. Given a database of *n* objects or data tuples, this method constructs *k* partitions of the data, where each partition represents a cluster and $k \leq n$. The *k* groups together generally satisfy the following requirements: 1) each group must contain at least one object, and 2) each object must belong to exactly one group. Each cluster is represented by the mean value of the objects in it. Although this technique was in use for long, it was first published by Stuart Lloyd in 1982 [4].

The k-means algorithm takes an input parameter, *k*, which represents the number of clusters that are required to be generated. It then partitions a set of *n* objects into *k* clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which is known as the cluster's *centroid*.

Algorithm *k-means*

```
Input – number of clusters $k$ and data set $D$ containing $n$ objects.

Output – A set of $k$ clusters

1) From $D$, randomly generate $k$ points as the initial cluster centres.

2) Assign each object to a cluster to which the object is the most similar, based on the cluster mean value and
   the object value.

3) Re-compute mean of each cluster from the objects in it and update the cluster means.

4) Repeat steps 2 and 3 till there is no change in clusters.
```

Consider a set of objects located in space as depicted in Figure 1(a). Suppose *k = 3*, then the k-means algorithm arbitrarily chooses three objects as the three initial cluster centres. Each object is allotted to a cluster based on the nearest cluster centre. This operation results in a distribution as shown in Figure 1(a).

In the next step, all cluster centres are updated using the objects they contain. This is done by recalculating the mean value of the current objects in the cluster. With the new cluster centres, the objects are redistributed to the nearest clusters. This operation results in a distribution as shown in Figure 1(b).

The process iteratively reassigns objects to clusters to improve the partitioning. This is called *iterative relocation*. Finally, when no redistribution of the objects in any cluster occurs, the process terminates. The resulting clusters are returned by the clustering process and they appear as shown in Figure 1(c).

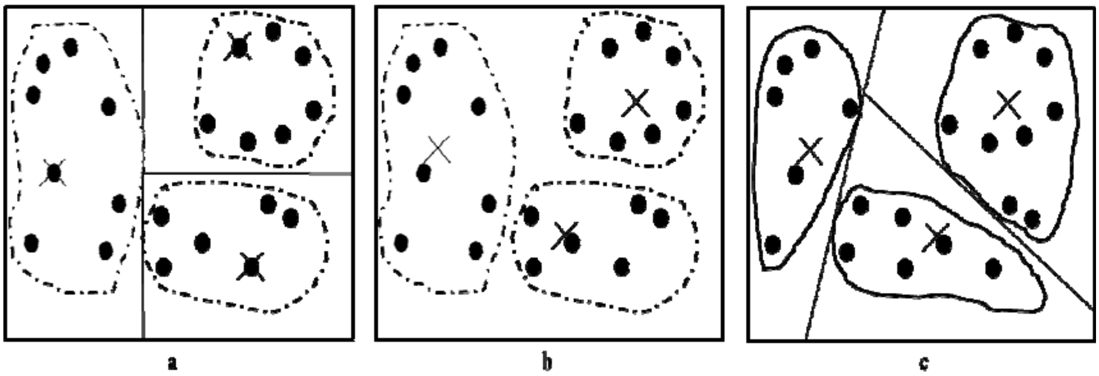


Figure 1. k-Means Clustering

The k-means algorithm has a computational complexity of *O(nkt)*. Although it is relatively scalable and efficient in processing large data sets, the necessity to specify the number of clusters to be generated (*k*) in advance, is seen as a disadvantage. Also, it is restricted when data with categorical attributes are involved. A variant of the k-means method, the Expectation-Maximisation algorithm is useful in such cases. It assigns each object to a cluster according to a weight representing its probability of membership rather than a strict distance measure. This is discussed in the next section.

III. Expectation Maximisation Clustering

Model-based clustering methods attempt to optimise the fit between the given data and a mathematical model. The assumption used is that the data are generated by a mixture of underlying probability distributions. Each cluster can be represented mathematically by a parametric probability distribution as shown in Figure 2. The problem is to estimate the parameters of the probability distributions so as to best fit the data. The Expectation-Maximisation (EM) algorithm, suggested by Dempster et al [5], is a popular iterative refinement algorithm that can be used for finding the parameter estimates. It is an extension of the k-means technique, but instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership. Hence, there are no strict boundaries between clusters. Therefore, new means are computed based on weighted measures.

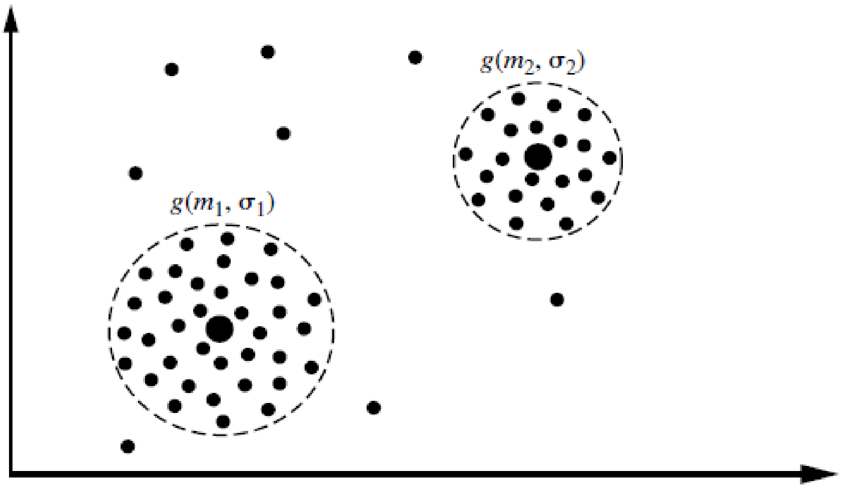


Figure 2. Two clusters corresponding to the Gaussian distributions $g(m_1,\quad {}_1)$ and $g(m_2,\quad {}_2)$, respectively

EM starts with an initial approximation of the parameters of the mixture model (collectively referred to as the parameter vector). It iteratively rescores the objects against the mixture density produced by the parameter vector. The rescored objects are then used to update the parameter estimates. Each object is assigned a probability that it would possess a certain set of attribute values given that it was a member of a given cluster.

Algorithm EM

Input – number of clusters *k* and data set *D* containing *n* objects.

Output – A set of *k* clusters

1) Make an initial estimate of the parameter vector by randomly selecting *k* objects to represent the cluster means or centre, as well as making guesses for the additional parameters.

2) Iteratively refine the parameters (or clusters) based on the following two steps:

(a) Expectation Step: Assign each object x_i to cluster C_k with the probability,

$$P(\mathbf{x}_i \in C_k) = p(C_k|\mathbf{x}_i) = \frac{p(C_k)p(\mathbf{x}_i|C_k)}{p(\mathbf{x}_i)},$$

where $p(x_i \mid C_k) = N(m_k, E_k(x_i))$ follows the Gaussian distribution around mean, m_k, with expectation, E_k i.e. this step calculates the probability of cluster membership of object x_i, for each of the clusters. These probabilities are the "expected" cluster memberships for object x_i.

(b) Maximisation Step: Use the probability estimates from above to re-estimate (or refine) the model parameters. For example,

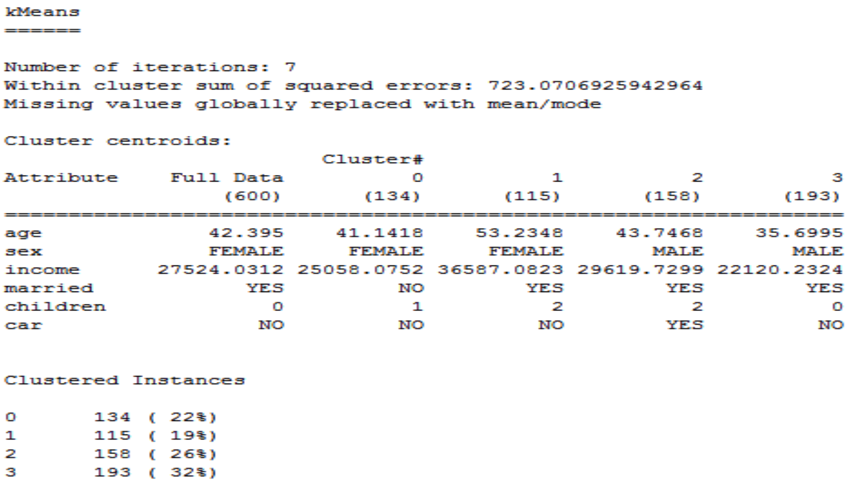
$$m_k = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i P(\mathbf{x}_i \in C_k)}{\sum_j P(\mathbf{x}_i \in C_j)}.$$

This step is the "maximisation" of the likelihood of the distributions given the data.

The EM algorithm is simple and easy to implement. In practice, it converges fast but may not reach the global optima. Convergence is guaranteed for certain forms of optimisation functions. The computational complexity is linear in *d* (the number of input features), *n* (the number of objects), and *t* (the number of iterations).

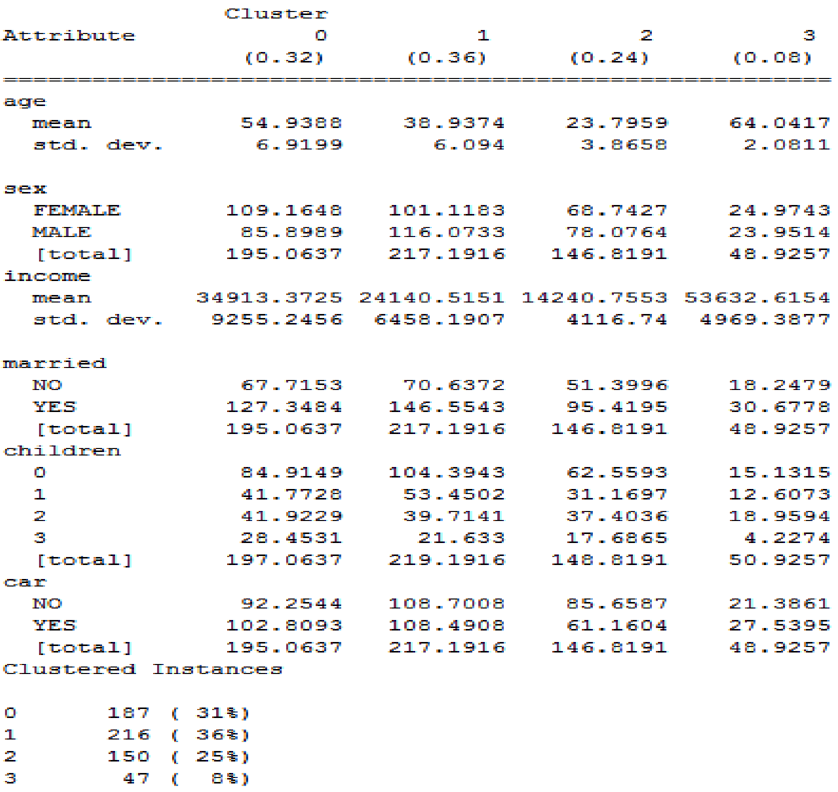
IV. Implementation of Clustering in Weka

For our experimentation, we tested the working of the k-means and EM algorithms in Weka (Waikato Environment for Knowledge Analysis) data mining tool. The data set provided was in attribute-relation file format (ARFF) through a file named “customers.arff”. The file contained records of 600 customers with data regarding eleven attributes of categorical nature. Of these, the selected attributes were age, gender, income, marital status, number of children and car owned. For both algorithms, the number of clusters specified was 4. The results for both algorithms are shown in Figures 3 and 4 respectively.



```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 723.0706925942964
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute     Full Data            0           1           2           3
                  (600)        (134)       (115)       (158)       (193)
========================================================================
age              42.395      41.1418     53.2348     43.7468     35.6995
sex              FEMALE       FEMALE      FEMALE        MALE        MALE
income       27524.0312   25058.0752  36587.0823  29619.7299  22120.2324
married             YES           NO         YES         YES         YES
children              0            1           2           2           0
car                  NO           NO          NO         YES          NO


Clustered Instances

0       134 ( 22%)
1       115 ( 19%)
2       158 ( 26%)
3       193 ( 32%)
```

Figure 2. Clusters generated using k-Means



```
                 Cluster
Attribute              0           1           2           3
                  (0.32)      (0.36)      (0.24)      (0.08)
============================================================
age
  mean           54.9388     38.9374     23.7959     64.0417
  std. dev.       6.9199       6.094      3.8658      2.0811

sex
  FEMALE        109.1648    101.1183     68.7427     24.9743
  MALE           85.8989    116.0733     78.0764     23.9514
  [total]       195.0637    217.1916    146.8191     48.9257
income
  mean        34913.3725  24140.5151  14240.7553  53632.6154
  std. dev.    9255.2456   6458.1907     4116.74   4969.3877

married
  NO             67.7153     70.6372     51.3996     18.2479
  YES           127.3484    146.5543     95.4195     30.6778
  [total]       195.0637    217.1916    146.8191     48.9257
children
  0              84.9149    104.3943     62.5593     15.1315
  1              41.7728     53.4502     31.1697     12.6073
  2              41.9229     39.7141     37.4036     18.9594
  3              28.4531      21.633     17.6865      4.2274
  [total]       197.0637    219.1916    148.8191     50.9257
car
  NO             92.2544    108.7008     85.6587     21.3861
  YES           102.8093    108.4908     61.1604     27.5395
  [total]       195.0637    217.1916    146.8191     48.9257
Clustered Instances

0       187 ( 31%)
1       216 ( 36%)
2       150 ( 25%)
3        47 (  8%)
```

Figure 3. Clusters generated using EM

From both figures, it is quite evident that, EM performs a better job at clustering categorical data as compared to k-means. For example, k-means generates age clusters centred around 35, 41, 43 and 53 which are not accurate as they do not encompass the entire range of customer ages. In comparison, EM generates age clusters centred around 23, 38, 54 and 64, which is more evenly distributed and has higher accuracy and significance. Similar comparisons may be made for the other attributes as well.

V. CONCLUSION AND FUTURE WORK

From our experimental results, we concur that K-Means performs comparably to EM. K-Means is popular industry standard used in clustering. Use of a business intelligence application incorporating this clustering mechanism to manage a retail business will provide retailers with means to segment customers and understand their behaviour and needs in a better way. It is important to provide powerful multidimensional analysis and visualisation tools, including the construction of sophisticated data cubes according to the needs of data analysis. Our current work is focusing on developing such an application that will include K-Means clustering to allow retailers to increase customer understanding and make knowledge-driven decisions in order to provide personalised and efficient customer service.

REFERENCES

[1] Girish Punj and David Stewart, "Credit Analysis in Marketing Research: Review and Suggestions for Application", Journal of Marketing Research, May 1983.

[2] Report on Indian Retail Industry - March 2011, Credit Analysis and Research Ltd., 2011

[3] A. Berson, S. Smith and K. Thearling, "Building Data Mining Applications for CRM", McGraw-Hill, 2002.

[4] Stuart Lloyd, "Least Squares Quantization in PCM", IEEE Transactions on Information Theory,Vol. IT-28, No. 2, March 1982.

[5] A. P. Dempster et al, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1, 1977.

[6] Tsukasa Ishigaki et al, "Customer-Item Category Based Knowledge Discovery Support System and Its Application to Department Store Service", IEEE Asia-Pacific Services Computing Conference, 2010.

[7] Maria Francesca Faraone et al, "Contextual segmentation: using context to improve behavior predictive models in e-commerce", IEEE Computer Society, 2010.

[8] Anil K. Jain, "Data clustering: 50 years beyond K-means", Elsevier Pattern Recognition Letters, September 2009.

[9] Indranil Bose and Xi Chen, "Exploring business opportunities from mobile services data of customers: An inter-cluster analysis approach", Elsevier Electronic Commerce Research and Applications, Vol. 9, 2010.

[10] Madhu Shashanka and Michael Giering, "Mining Retail Transaction Data for Targeting Customers with Headroom - A Case Study", Springer IFIP Advances in Information and Communication Technology, Vol. 296, 2009.

[11] E.W.T. Ngai et al, "Application of data mining techniques in customer relationship management: A literature review and classification", Elsevier Expert Systems with Applications Vol. 36, 2009.

[12] Shian-Chang Huang et al, "A case study of applying data mining techniques in an outfitter's customer value analysis", Elsevier Expert Systems with Applications Vol. 36, 2009.

[13] Tianyi Jiang and Alexander Tuzhilin, Improving Personalization Solutions through Optimal Segmentation of Customer Bases", IEEE Transaction on Knowledge and Data Engineering, Vol. 21, No. 3, March 2009.

[14] Indranil Bose and Xi Chen, "Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn", Proceedings of IMECS 2009, Hong Kong, Vol. I, March 2009.

[15] Chia-Cheng Shen and Huan-Ming Chuang, "A Study on the Applications of Data Mining Techniques to Enhance Customer Lifetime Value", WSEAS Transaction on Information Science and Applications, Issue 2, Vol. 6, February 2009.

[16] Pradip Kumar Bala, "Exploring Various Forms of Purchase Dependency in Retail Sale", Proceedings of the World Congress on Engineering and Computer Science 2008, San Francisco, USA.

[17] Chinho Lin and Chienwen Hong, "Using customer knowledge in designing electronic catalog", Elsevier Expert Systems with Applications, Vol. 34, 2008.