Customer Segmentation for Decision Support using Clustering and Association Rule based approaches

S.Balaji[1] ,Dr.S.K.Srivatsa[2]
[1]Research Scholoar,Vels University,Chennai Email: srisaibalaji@rediffmail.com
[2]Senior Professor, St.Joseph Eng.College Chennai

Abstract*-Key business areas that data mining techniques can be potentially applied to include business profitability, customer relationships, and business process efficieny. Customer Realtionship Management(CRM)has become a leading business strategy in highly competitive business environments.Clustering customers provides an in-depth understanding of their behavior. Clustering is one of the most important and useful technologies in data mining methods. Clustering is to group objects together, which is based on the difference of similarity on each object, and making highly homogeneity in the same cluster, or highly heterogeneity between each group. The scope of this paper to understand and predict the behavior of customers with behaviour segmentation methodology. The result of the study can support customer development by offering the right products to right customers and better targeting of product promotion campaigns. The policy holders claim dataset of health insurance company is taken for analysis. This behaviour segmentation methodology with clustering is applied in this chapter to predict distinct customer segments facilitating the development of customized new products and new offerings which better address the specific priorities and preferences of the customers. Apriori association rule performed on clusters of claim dataset gives the association among attributes in the claims dataset is derived from Clustering Based Association Rule Mining (CBARM) model. Association rule technique is applied on claim dataset to predict claim cost and the association among attributes that influences the claim cost of the policy holders.

Keywords:-CRM,Datamining,segmentation,clustering.

I.INTRODUCTION

In today’s competitive environment, understanding the customers better, especially, most profitable customer groups and the groups that have the biggest potential is the biggest challenge. By segmenting customers based on their behavior,we can better target their actions, such as launching tailored products, target one-to-one marketing and to meet the customer expectations. However, the problem often is that the data regarding customer behavior is available in several different sources and analyzing the large data set is exhaustive and time consuming.

II.BACKGROUND WORK

Clustering can be defined as the process of grouping a set of physical or abstract objects into classes of similar objects[2]. Clustering is also called unsupervised classification, because the classification is not dictated/ordered by given class labels. There are many clustering approaches, all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).This chapter identified extensive work on customer segmentation with data mining techniques and elaborates as follows:

Samira et al.(2007) applied segmentation of customers of Trade Promotion Organization of Iran using a proposed distance function which measures dissimilarities among export baskets of different countries based on association rules concepts. Later,in order to suggest the best strategy for promoting each segment,each cluster is analyzed using RFM model. Variables used for segmentation criteria are “the value of the groupcommodities”, “the type of group-commodities” and “the correlation between export group-commodities”.

Huang, Chang, and Wu (2009) applied K-means method, Fuzzy C-means clustering method and bagged clustering algorithm to analyze customer value for a hunting store in Taiwan and finally concluded that bagged clustering algorithm outperforms the other two methods.

Pramod et al.(2011) elaborates the use of clustering to segment customer profiles of a retail store.The study concluded that the K-Means clustering allows retailers to increase customer understanding and make knowledge-driven decisions in order to provide personalised and efficient customer service.

Hosseini et al. (2010) adopted K-means algorithm to classify the customer loyalty based on RFM values. Cheng and Chen (2009) used K-means and rough set theory to segment customer value based on RFM values. Chen et al. (2009) identified purchasing patterns based on sequential patterns.

Migueis.V.L et al.(2012) proposed a method for customers segmentation,given by the nature of the products purchased by customers.This method is based on clustering techniques, which enable segmenting customers according to their lifestyles. The author segmented customers of an European retailing company according to their lifestyle and proposed promotional policies tailored to customers from each segment, aiming to reinforce loyal relationships and increase sales.

Kanwal garg et. al(2008) applied clustering and decision tree techniques for identifying the trend of customer investment behavior in life insurance sector in India.This paper analyzed the prediction of customer buying preference over newly launched policies.

Our proposed work in this paper uses clustering technique to predict the customer behaviour for the claim dataset of health insurance company.In addition the association rule algorithm is implemented to predict the nature of health insurance claims and the health risks.

III.SEGMENTATION STRATEGY

The key goal of customer segmentation is identifying and achieving profitable sectors and provides products and services that are the customer's common need[1]. Sophisticated customer segmentation give the companies the ability to target profitable customers,understand customers' demands, allocation of resources and compete with rivals [14].The proposed segmentation framework is given in Figure 1.

The proposed framework has three phases:

1. Data preparation phase
2. Data clustering phase
3. Customer preference analysis

A part of the first phase includes collection of data from data store and the subsequent data cleansing.Second phase generates Behavioral segmentation based clusters and profile of the clusters.Third phase is concerned with identification of customers preferences over products and the risk levels for diseases diagnosed,treated and subsequent process of claim settlement.

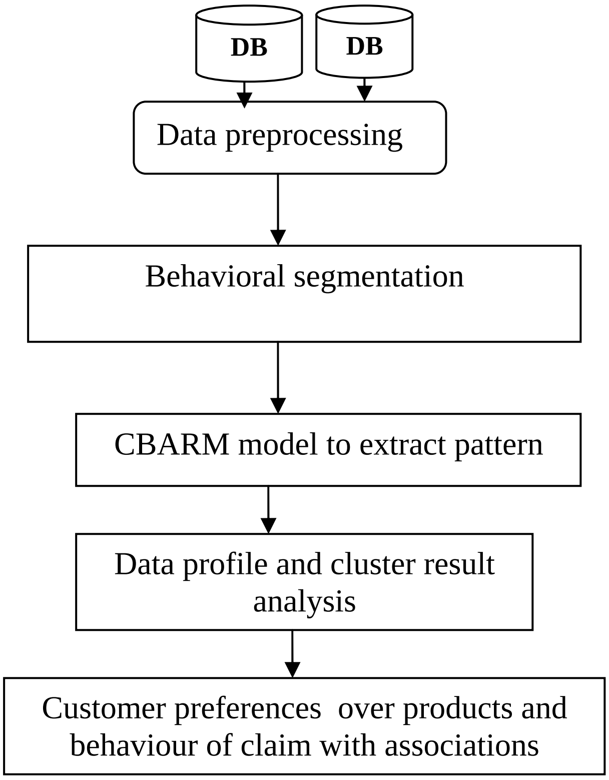
DB
DB
Data preprocessing
Behavioral segmentation
CBARM model to extract pattern
Data profile and cluster result analysis
Customer preferences over products and behaviour of claim with associations

Figure 1 Proposed framework for behavioral segmentation and clustering process

A. Data preparation

The source dataset was extracted from Insurance regulatory authority of India's (IRDA) of health insurance claim database.The dataset has 15000 customers claim information for the period of one year 2009-2010 .The initial dataset consists of 36 attributes.The attributes relevant are only taken for analysis. After preprocessing,only 21 attributes relevant for analysis are taken into consideration.Health insurance dataset has policy holders whose type of cover is one among the four types such as individual, individual floater , group and group floater.

B. Behavior analysis

In order to describe customer behavior,several attributes can be identified from the policy holders transactions.The policy holders can be characterized by kind of policies opted and with what kind of products they avail.Additionally usage based factors such as how many customers use different policies,during which occasions and how much they tend to avail claim for the health insurance policy opted.The risk associated with each policy holders decides the claim cost.The risk of the policy holders depends on healthy behaviour.The diseases diagnosised,Medical history,treatement undergone for recovery process influences the total claim of the policy.

C. Clustering Strategy

Customers are divided into distinct segments by using cluster analysis.The clustering fields, typically the component scores, are fed as inputs into a cluster model which assesses the similarities between the records/customers and suggests a way of grouping them.Customers claim information is based on health risk The cluster is formed in such a manner as to reveal significant areas of risk within an insurance portfolio.The framework of behavioral segmentation is used to predict the expected claim costs for different health risks diagnosed and treatment undergone. K-means is one of the unsupervised learning algorithm that solve the well known clustering problem. The procedure is to classify a given data set through a certain number of clusters (assuming k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because different locations cause different results.

D. Association rule based analysis

The analysis of claims data in our health insurance claim dataset is performed using apriori association algorithm.Association rule based analysis helps in identifying and establishing claiming patterns and relationships which represent appropriate utilization of services.

IV. EXPERIMENTAL METHODOLOGY AND RESULTS

Clustering an unsupervised learning technique is applied to form cluster of customers based on claim behavior of health insurance customers.The health risk of the customer has a significant impact on insurance claim. The WEKA ("Waikato Environment for Knowledge Analysis")machine learning work bench 3.7.5 has been considered for the purpose of analysis and test results.

A. Clusters for customer data analysis

The cluster formation and evaluation under classes to clusters mode in WEKA is implemented for policy type class atrribute. Each cluster shows us a type of behavior in customers, from which conclusions are drawn. Weka first ignores the class attribute and generates the clustering. Then, during the test phase, it assigns classes to the clusters, based on the majority value of the class attribute within each cluster. The total number of instances were not fairly distributed if the default seed size is 10 and number of clusters is set to four.It means the sum of the whole cluster was 99%. The distribution of instances were significantly improved, when the seed number was set to 100. When the seed size is 10, the number of iterations and sum of squared errors were 10 and 22193.62 respectively. However, if the seed size is set to 100, the number of iterations was 10 and sum of squared errors was 14864.65.

Cluster 0 –This group is mostly dominated by policy class of group policy holders who may be working in a company or members of a co-operative society and so on .The policy holders availed the high claim value compared to other groups.Female gender is dominated in the group and the claims are being mostly diagnosed with pregnancy, childbirth and the puerperium related. Nearly 30% of total claims comes from medicine charges.Surgical,consultation and investigation claims are highest among the cluster groups.

Cluster 1-This group is having policy holders with claims of higher value. Claims were dominated by eye and adnexa problems.This group is having policy holders of early forties and their claim values are dominated by miscellaneous expenses and also this group is dominated by individual health insurance policy holders.

Cluster 2-This group has low risk customers whose claim values are lowest among other groups.Most of the claims pertain to cases diagnosed with pain abdomen problems of medium age groups with surgical less treatments.The treatment value in all headers is low compared to other groups other than pre-hospitalization expenses.

Cluster 3-This group is dominated by group floater policies.Due to minimum sum assured,total claim is second lowest among the groups.Most of the claims are surgical in nature and the diganosed claims are of mostly normal pregancy related with moderate surgical expenses.The claim value is having least post and pre_hospitalization expenses.

The corresponding confusion matrix for the clustering model based on policy type attribute is given in Table I.

Table I CONFUSION MATRIX FOR THE CLUSTERING MODEL

```
=== Model and evaluation on training set ===
Clustered Instances
0       3699 ( 25%)
1       7109 ( 47%)
2        923 (  6%)
3       3269 ( 22%)
Class attribute: Txt_Type_of_Policy
Classes to Clusters:
0    1    2    3  <-- assigned to cluster
1359 2534   52  733 |    1
82  265   60   81 |    2
768 1289   62  433 |    3
1490 3021  749 2022 |    4

Cluster 0 <-- 3
Cluster 1 <-- 1
Cluster 2 <-- 2
Cluster 3 <-- 4
```

From the confusion matrix given in Table I ,it is evident that cluster 1 has the policy holders mostly dominated by individual policies,followed by cluster 2,cluster 0 and cluster 3 having policy type preferences of individual ,group and group floater policies. The tuples of cluster 1 with major contribution of 47% of instances followed by cluster 0,cluster 3 and cluster 2. By considering the above facts,each cluster is assigned ranks as cluster 1 having rank 1 followed by cluster 0,cluster 3 and cluster 2 having ranks 2,3, and 4 respectively.

B. Association rule based observation

Prediction of claim levels are based on distribution of diseses among the instances of the working set. Apriori association rule performed on claim dataset gives the association among attributes in the claims dataset.Different association rules express different regularities that underlie the dataset and predict different things. The values for number of rules considered, the decrease for minimum support (delta factor) and minimum confidence values are 3, 0.95 and 0.9 respectively.The number of itemsets and the corresponding rule mapping is given in Table II

TABLE II

APRIORI ASSOCIATION RULE SET ON CLAIM DATASET

	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Size of rule set	9	36	84	126	126	84	36	9	1

Best three rules generated with 14490,14986 and 14983 instances associated and its describtion as follows:

First rule specifies it is of single rule set category and if miscellaneous expenses of range (0-285712] has the impact on post hospitalization expenses of range(0-297650.66)with confidence level of 1. In addition,lift =1 indicates that the having miscellaneous expenses within the range increases the probability of producing the post_hospitalization expenses by a factor of 1. Leverage is the proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other.

In the second rule, which is of two set category,the attribute investigation_charges and miscellaneous charges have significant impact on post_hopitalization_expenses of the claim. The ranges of [0-153538.66] for investigation charges and [0-285712]for miscellaneous_charges have the impact on post_hospital_expenses within the range

[0-297650.66] with the confidence level of 1.

Third rule which is of two set category,that predicts post hospitalization expenses.It is observed from the third rule that if surgery charges of range(0 to 116666.67) and miscellaneous charges of range (0 to 285712) influences the post hospitalization expenses to fall within the range 0 to 297650.67.

The component lift = 1 indicates that having LHS predicates increases the probability of RHS predicates by a factor of 1 in the best 10 rule sets.

As a reference to third rule,the component post hospitalization of claim falls in the range (0-297650.67) is dependent on surgery charges of range (0-116666.67)and Miscellaneour charge of range(0-285712). Similarly,other two item sets rule 4, rule 5 and rule 9 justifies the dependent factors(LHS) on RHS attribute

Post_Hospitalisation_Expenses. Sixth rule is of three item set category specifies if the claim amount lies within the range 0 to 456470.67, investigation_charges between 0 to 0-153538.67 and Miscellaneous_Charges between 0 to 285712 has an significant impact on Post_Hospitalisation_Expenses of range 0 to 297650.67. This rule has 14979 instances justifies that.Similarly, rule numbers 7, 8 and 10 are of three item category justifies the dependent factors (LHS) on RHS attribute Post_Hospitalization_expenses.

V CONCLUSION

Clustering is one of unsupervised learning technique was applied on health insurance claim dataset to segment the customers. Clusters revealed the preferences of customers towards the products and factors that influence total claim.Association rules employed on working dataset to predict nature of the claim. The process of combining segmentation with data mining, provides marketers with high quality information on how their customers shop for and purchase their products or services. By combining standard market segmentation with data mining techniques can better predict and model the behavior of the segments. Segmentation with the help of data mining from various existing systems is a very important exercise and a must for effective business development. Making this intelligence available to the customer facing teams and can prove to be a great tool to increase cross selling and up selling capability of a company.

REFERENCES

[1] Blocker, C. P., & Flint, D. J.(2007). Customer segments as moving targets: Integrating customer value dynamism into segment instability logic. Industrial Marketing Management. Vol.36 ,pp. 810–822.

[2] Kamber.M, J. Han(2008), Data Mining : Concepts and Techniques, Morgan Kaufmann.

[3] Samira Malekmohammadi Golsefid, Mehdi Ghazanfari, Somayeh Alizadeh(2007),Customer Segmentation in Foreign Trade based on Clustering Algorithms, World Academy of Science, Engineering and Technology Vol.28 ,pp.405-411.

[4] Huang.S.C, E. C. Chang, & H. H. Wu,(2009), A case study of applying data mining techniques in an outfitter's customer value analysis, *Expert System with Applications*, Vol.36.Issue.6, pp.5909–5915.

[5] Hosseini, M., Anahita, M., Mohammad, R., G. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications,Vol. 37,pp.5259–5264

[6] Pramod Prasad,Latesh G. Malik.(2011),Generating customer profiles for retail
stores using clustering techniques,International Journal on Computer Science and Engineering,Vol.3,pp.2506-2510.

[7] Miguéis.V.L,Camanho.A.S,João Falcão e Cunha,(2012), Customer data mining for lifestyle segmentation,Expert Systems with Applications,Vol.39,pp.9359-9366.

[8] Kanwal garg,Dharminder kumar and M.C.Garg,(2008),Data mining techniques for identifying the customer behavior of investment in life insurance sector in India,International Journal of Information technology and knowledge management,Vol.1,Issue.1,pp.51-56.

[9] Larose,D.T.(2006) Data mining methods and model,Hoboken,New Jersey,John Wiley & sons,Inc.

[10] Mike McGuirk,.(2007) Customer segmentation and predictive modeling-http://www.iknowtion.com/downloads/Segmentation.pdf

[11] Pieter Adriaans,Dolf Zantinge,(2011),Data Mining,Pearson Eduction Ltd,Sixth Impression,2011.

[12] Roosevelt Mosley,(2005),The Use of Predictive Modeling in the Insurance Industry,Pinnacle actuarial resources.

[13] Two Crows Corporation (1998) Introduction to Data Mining and Knowledge Discovery,Second Edition Patomac, MD.

[14] Tsiptsis.K and A. Chorianopoulos. Data Mining Techniques in CRM: Inside Customer Segmentation (2009),John Wiley & Sons, Ltd,Second Edition.

[15] Westphal.C and T. Blaxton (2005), Introduction to Data Mining and Knowledge Discovery,Two cows Corporation,Third Edition.