Market basket analysis in insurance industry

Marzieh Vahidi Roodpishi[a*] and Reza Aghajan Nashtaei[b]

[a]*M.A. of Business Management, Rasht Branch, Islamic Azad University, Rasht, Iran*
[b]*Department of Business Management, Rasht Branch, Islamic Azad University, Rasht, Iran*

CHRONICLE	ABSTRACT
	Nowadays, many organizations focus on discovering their customers' hidden patterns to maintain their competitive position through customer analysis. In fact, more and more organizations are realizing that customers are their most valuable resources. This paper performs a research using data associated with 300 clients of an insurance company in city of Anzali, Iran and they are analyzed using K-Means clustering method. Using demographic variables including gender, age, occupation, education level, marital status, place of residence and clients' incomes, the study determines the optimal numbers of clusters in order to achieve necessary data for grouping customers. Next, the study uses the method of association rules to find hidden patterns for the insurance industry.

1. Introduction

With the advent of new technology and competition facilities, the market environment of the insurance industry has become highly competitive. Insurance industry and the entry of private insurance have led to intense competition among new firms and older firms pay more attention to ways of analyzing customer data. Therefore, considering the market basket helps insurance firms reach better understanding about clients' demands. Decision-making and understanding customer behavior are critical and challenging issues for organizations to maintain their position in the competitive market. Technological innovation also helps firms achieve better processing of customers' needs. Data mining tools as the most reliable tool for analyzing large amounts of data helps achieve success in decision-making. Data mining analysis can be effectively used to determine the patterns and to optimize the dynamic behavior of transactions made by consumers for the purchase of specific products. Mansur and Kuncoro (2012), in a survey, tried to find out more about consumers' behavior in buying the products so it could be used to forecast the purchasing for the next period. Later, they used the prediction as a decision support in detecting the suitable amount of inventory for each product at

*Corresponding author.
E-mail addresses: kvahidi1989@yahoo.com (M. Vahidi Roodpishi)

Karomah Brass, which was a small and medium enterprise. The techniques used in this study were based on the Market Basket Analysis (MBA) and Artificial Neural Network (ANN) Back propagation. They used MBA to investigate the buying behavior of customer while ANN Back propagation was applied to predict product inventory's requirements/needs for each product. They reported that the customers frequently buy products that serve as a type of antique closet accessories and if customer purchased that certain product, then they would also buy similar products in accordance with 21 rules obtained from the mining of transaction data.

Trnka (2010) explained the way of MBA implementation to Six Sigma methodology. Data mining techniques provide significant amount of opportunities in the market sector. Data mining applications are becoming increasingly popular for many applications across a set of very divergent fields. Six Sigma methodology implements several statistical techniques (Pande & Abdel-Aty, 2009). With implementation of MBA as a part of Data Mining (Han et al., 2006) to Six Sigma Trnka improved the results and changed the Sigma performance level of the process. In his survey, General Rule Induction algorithm to produce association rules between products in the market basket was implemented. Cheng and Chen (2009) presented a new procedure, joining quantitative value of RFM attributes and K-means algorithm into rough set theory (RS theory), to obtain meaning rules and to find out the characteristic of customer in order to strengthen CRM.

According to Russell and Petersen (2000), Market basket choice is a decision process in which a consumer chooses different items from a number of product categories on the same shopping trip. The primary feature of market basket choice is the interdependence in demand relationships across the things in the target basket. Russell and Petersen (2000) developed a new method to the specification of market basket models, which helps a choice model for a basket of products to be built using a set of "local" conditional choice models corresponding to each item in the basket. The method provides a parsimonious market basket model, which permits for any kind of demand relationship across product categories and can be forecasted based on simple modifications of standard multinomial logit software.

Tang et al. (2008) presented a method to perform MBA in a multiple-store and multiple-period environment. They first defined a time concept hierarchy and a place (location) hierarchy, based on the application and requirements. A set of contexts was systematically extracted from the two hierarchies by integrating the concept levels of the two hierarchies. They developed an efficient approach for extracting the association rules, which could meet the support and confidence needs for all the contexts. Using the approach, a decision maker is able to analyze purchasing patterns at very detailed concept levels of time and place and combinations of detailed levels of one with general level of the other. The association rules appeared to be well organized, because they were generated based on the contexts extracted from the time and place hierarchies.

Cavique (2007) presented a method to discover large itemset patterns for the MBA, where the condensed data were used and was obtained by transforming the market basket problem into a maximum-weighted clique problem. Wick and Wagner (2006) applied MBA to integrate and motivate topics in discrete structures. Dhanabhakyam and Punithavalli (2011) examined customer buying patterns by detecting associations among different items that customers place in their shopping baskets. Raorane et al. (2012) investigated the huge amount of data thereby exploiting the consumer behavior and making the correct decision leading to competitive edge over rivals. Yun et al. (2006) explored a clustering of market-basket data, which was different from the traditional data.

2. The proposed model

This paper performs a research using data of 300 clients of insurance company in city of Anzali, Iran and they are analyzed using K-Means clustering method. Using demographic variables including gender, age, occupation, education level, marital status, place of residence, income clients the study determines the optimal number of clusters in order to achieve necessary data for grouping customers.

Next, the study uses the method of association rules and practice of insurance policies to find hidden patterns in the insurance industry. Fig. 1 demonstrates the summary of the proposed study.
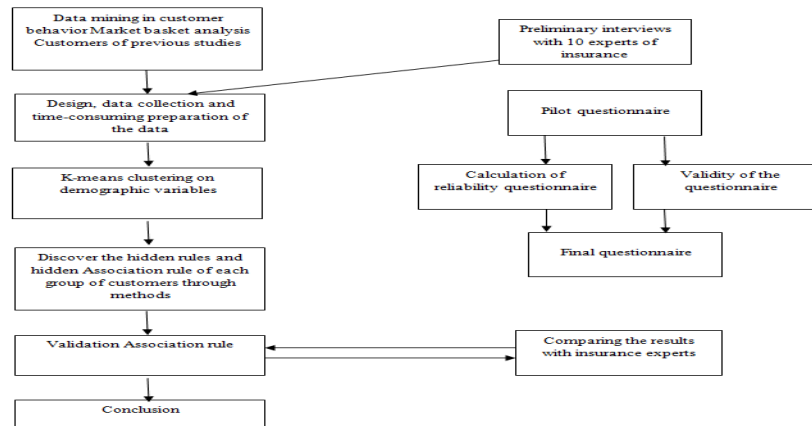

Fig. 1. The structure of the proposed study

In Fig. 1, the structure proposed study includes the following:

• The first step is to review the literature and identify relevant variables for modeling the behavior of the customers and design the questionnaire to collect the data from the insurance companies.

• The second step of the K-Means is associated with clustering and using demographic variables including gender, age, occupation, education level, marital status, place of residence, income clients to determine the optimal number of clusters, on the banks intelligence.

• The third step uses the method of association rules to determine hidden patterns in cart insurance industry clients.

• The fourth step is to validate the results obtained by our data on the association rules.

The study has accomplished among customers of third party car insurance in city of Anzali located in province of Gilan, Iran in 2014. All questionnaires were distributed in 6 Insurance Agents in Iran. This paper performs a research using data of 300 clients of insurance company in city of Anzali, Iran and they are analyzed using K-Means clustering method.

2.1 K-means technique

The K-means technique has become popular method in generating appropriate clustering results for many real-world case studies. K-means clustering is a well-known data mining clustering model, which tries to partition N observations into K clusters where each observation is assigned to the cluster with the closest mean. Normal evaluation of an appropriate K is performed by minimizing the inner-cluster variation and maximizing the among-cluster variation, concurrently. K-means clustering is normally sensitive to outliers, therefore, outliers have to be eliminated before accomplishing clustering. Edwards (2003) and Kantardzic (2011) described the K-means method with the following steps,

1. Select a primary part of K categories including samples randomly chosen and compute the mean of each pair,
2. Build a new section of each part by computing the nearest center core,
3. Compute the new batches as the main centers,
4. Repeat step 2 and step 3 until the algorithm reaches termination criteria.

2.2. Association rules

Non-regulatory mining association rules in data mining to establish the relationship between the items are created. An association rule of the form X → Y is a term meaning that X and Y are discrete sets of items. (X ∩ Y = Y) means that Power of Community law can be given support and confidence measures (Ngai et al., 2009).

3. Implementation

For the proposed study of this paper, the implementation of the data modeling, data mining software are executed using SPSS Clementine data mining software, which is widely used for modeling. We consider the following three steps:

Step One: Use the clustering method to get the direction of grouping customers,
Step Two: Use the method of association rules to find hidden patterns in the insurance industry customer MBA,
Step Three: Review the insurance industry clients Basket validation by association rules.

3.1. Clustering

Clustering the data is accomplished based on the demographic variables including gender, age, occupation, education level, marital status, place of residence and clients' income clients. These are more descriptive data and the proposed study uses a K-Means clustering. For the implementation of K-Means clustering, the number of clusters is important. Therefore, we optimized it using SSE criterion for assessing the quality of clustering to determine the number of clusters. Due to the high volume of data to compare, the number of clusters starts from 2 clusters and Table 1 demonstrates the results of SSE for different numbers of clusters.

Table 1
The number of parameters career clusters

Number of clusters	2	3	4	5	6
SSE	1.548341	1.547119	1.600229	1.593326	1.629742

Characteristics of each cluster are as follows :

Cluster 1: Customers or employees, mostly engineers, earn a monthly income of between 5×10^6-10×10^6 Rials.

Cluster 2: Customers, employees or farms with income between 2.5×10^6-5×10^6 Rials and are mostly high school graduates.

Cluster 3: Customers aged 35-50 with 12-14 years of education and a monthly income of 5×10^6-20×10^6 Rials.

Cluster 4: Customers aged 35-50 with Bachelor degree of science and a monthly income of 10×10^6-20×10^6 Rials.

Cluster 5: Customers aged, at least, 50 with 12 years of education and a monthly income of at least 20×10^6 Rials.

3.2. Association rules and hidden patterns

Once the data for insurance customers have been collected and clustering has been accomplished, we examined the association rules hidden in each cluster using causal extraction algorithm. The following items have been investigated for validation of the data,

• Which insurance services do customers use (auto insurance, life insurance, health insurance, accident insurance, liability insurance, engineering insurance),
• Term life insurance,
• Annual fee for insurance,
• Specifications of the month of insurance,
• Location of the insurance firm,
• Validity of the leading insurance company in the insurance services,
• Use the services of the insurance company upon the recommendation of friends and family,
• The advantage of insurance company's offer to friends and family.

Tables 2-6 present some examples of applying Association rule derived based on on clusters of 1 to 5.

Table 2
Examples of association rules in cluster

Lift	Confidence%	Support%	Antecedent	Consequence
1.106	80.0	26.596	Time = 2-5 and Month = 2nd quarter and Credit = yes and Life= yes and Income = 5×10^6-10×10^6 Rials	Car Insurance = Yes
1.016	80.0	15.957	Time = 2-5 and Month = 2nd quarter and Credit = yes and Life= yes and Income = 5×10^6-10×10^6 Rials and Engineering = Yes	Life Insurance= Yes
1.117	80.769	27.66	Time = 2-5 and Month = 2nd quarter and Credit = yes and Accident= yes and Income = 5×10^6-10×10^6 Rials and Engineering = Yes	Car Insurance = Yes
1.152	83.333	31.914	Time = 2-5 and Month = 2nd quarter and Credit = yes and Life= yes and Income = 5×10^6-10×10^6 Rials and Engineering = Yes	Car Insurance = Yes
1.144	85.185	28.723	Time = 2-5 and Month = 2nd quarter and Credit = yes and Life= yes and Income = 5×10^6-10×10^6 Rials and Engineering = Yes	Accident= yes
1.163	86.667	24.468	Time = 1-2 and Month = 2nd quarter and Credit = yes and Life= yes and Income = 5×10^6-10×10^6 Rials and Engineering = Yes	Accident= yes
1.0217	100.0	24.468	Time = 2-5 and Month = 2nd quarter and Accident = yes and Credit = Yes and Income = 5×10^6-10×10^6 Rials	Engineering = Yes
1.0217	100.0	26.596	Time = 2-5 and Car ownership = yes and Credit = Yes and Income = 5×10^6-10×10^6 Rials	Engineering = Yes
1.144	85.185	28.723	Time = 2-5 and Month = 2nd quarter and Credit = Yes and Income = 5×10^6-10×10^6 Rials and Engineering = Yes	Accident= yes

The results of Table 2 indicate that the cluster is ruled by one of the following:

1) People who use their car insurance usually have maintained between 1 to 5 years of insurance services. Usually, this is the second quarter to extend the validity of the insurance care, Life insurance and engineering services and income between 5×10^6-10×10^6.

2) People who use life insurance usually hold between 1 to 2 years insurance. Usually, this is the second quarter to extend insurance and engineering insurance services are between 500 thousand to 1 million dollars.

3) People who usually have between 1 to 5 years of car insurance. This is usually the second quarter to extend insurance and engineering insurance and life insurance services.

4) People who use their car insurance and it has been between 1 to 2 years of using their insurance. This is the second quarter to extend their insurance and engineering insurance and have incomes between 5×10^6-10×10^6 Rials.

5) People who use their engineering insurance and they have maintained between 2 to 5 years of using their insurance. Usually this is the second quarter of credit insurance services, life insurance is important for them to use and their incomes are between 5×10^6-10×10^6 Rials.

Table 3

Examples of association rules in the second cluster

Lift	Confidence%	Support%	Antecedent	Consequence
0.962	87.5	18.181	Time = 2-5 and Month = 3rd quarter and Place = yes and Income < 25×10^6 Rials	Car Insurance = Yes
1.1	100.0	13.636	Time = 5-10 and Month = 3rd quarter and Place = yes and Income < 25×10^6 Rials	Car Insurance = Yes

Based on the results of Table 3, the following can be concluded,

1) People who use their car insurance and maintain between 5 and 10 years of insurance. Place of issuing insurance is important.

2) People who use their car insurance and maintain between 5 and 10 years of insurance. Place of issuing insurance is important.

Table 4

Examples of association rules in cluster 3

Lift	Confidence%	Support%	Antecedent	Consequence
0.983	80.0	819.60	Time = 2-5 and Month = 3rd quarter and Place = yes Income =5×10^6-10×10^6 Rials and Car Insurance = yes	Life insurance = yes
0.986	80.246	279.41	Place = yes and Responsibility insurance = yes and Accident insurance = yes and Car Insurance = yes	Life insurance = yes
1.012	82.352	716.66	Time = 1-2 and Month = 3rd quarter and Place = yes Income = 5×10^6-10×10^6 Rials and Responsibility Insurance = yes	Life insurance = yes
0.913	83.33	11.764	Time < 1 and Life insurance = yes and Month = 3rd quarter and Credit = yes and Place = yes	Accident insurance = yes
0.877	80.0	614.70	Time < 1 and Car insurance = yes and Month = 3rd quarter and Credit = yes and Place = yes	Accident insurance = yes
1.009	100.0	12.745	Time < 1and Month = 3rd quarter and Credit = yes Income =5×10^6-10×10^6 Rials	Car Insurance = yes

According to Table 3, the following can be concluded,

1) People use their life insurance usually between 2 and 5 years. They renew their insurance in the third quarter of the year and place of insurance firm is important for them.

2) People who use their life insurance usually between 1 to 2 years of insurance plan and renew their life insurance in the third quarter of year and provide liability and accident insurance.

3) People who use life insurance where the insurance company is important in choosing their insurer. And liability insurance and auto accidents are often used.

4) People who have insurance are usually under one year of using insurance. They renew their insurance in the third quarter of year.

Table 5

Examples of association rules in cluster 4

Lift	Confidence%	Support%	Antecedent	Consequence
0.925	80.0	27.027	Time = 2-5 and Credit = yes Income =10×10^6-20×10^6 Rials and Engineering Insurance = yes and Life Insurance = yes	Accidence Insurance = yes
0.961	90.909	329.7	Time = 2-5 and Month = 2nd quarter and Accident = yes and Income =10×10^6-20×10^6 Rials and Credit = yes	Life insurance = yes
0.961	90.909	329.7	Time = 2-5 and Month = 2nd quarter and Accident = yes and Income =10×10^6-20×10^6 Rials and Engineering insurance = yes	Life insurance = yes
1.0571	100.0	62.162	Time = 2-5 and Month = 2nd quarter and Accident = yes and Income =10×10^6-20×10^6 Rials and Engineering insurance = yes	Engineering Insurance = yes
1.0571	100.0	27.027	Time = 2-5 and Month = 2nd quarter and Life insurance = yes and Income =10×10^6-20×10^6 Rials and Credit = yes	Engineering Insurance = yes
1.073	92.857	37.837	Time = 2-5 and Life insurance = yes and Income =10×10^6-20×10^6 Rials and Credit = yes and Engineering insurance = yes	Accidence Insurance = yes
1.108	95.833	64.864	Month = 2nd quarter and Life insurance = yes and Income =10×10^6-20×10^6 Rials and Credit = yes	Accidence Insurance = yes

According to Table 5, the following can be concluded,

1) People who use their life insurance have between 2 and 5 years of insurance and extend their insurance and engineering insurance services during the second quarter of the season. Credit insurance is also important for them.

2) People who use insurance have 1 to 2 years of insurance, extend their insurances including life insurance services and engineering in the second quarter of year.

4. Discussion and conclusion

The present study was designed to analyze the insurance company's customers based on Market basket analysis. The paper has performed a research using data associated with 300 clients of insurance company in city of Anzali, Iran and they have been analyzed using K-Means clustering method. Using demographic variables including gender, age, occupation, education level, marital status, place of residence, clients' income, the study has determined the optimal numbers of clusters in order to achieve necessary data for grouping customers. Next, the study used the method of association rules and practice of insurance policies to find hidden patterns in the insurance industry. The results of this survey could be used for targeting appropriate customers in cities located in north regions of Iran. The study can be also extended for other regions of the country and we leave it for interested researchers as future studies.

Acknowledgement

References

Cavique, L. (2007). A scalable algorithm for the market basket analysis. *Journal of Retailing and Consumer Services*, *14*(6), 400-407.

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3), 4176-4184.

Dhanabhakyam, M., & Punithavalli, M. (2011). A survey on data mining algorithm for market basket analysis. *Global Journal of Computer Science and Technology*, *11*(11).

Edwards, D. (2003). Data mining: Concepts, models, methods, and algorithms. *Journal of Proteome Research*, *2*(3), 334-334.

Han, J., Kamber, M., & Pei, J. (2006). *Data mining, Southeast Asia edition: Concepts and techniques*. Morgan Kaufmann.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Mansur, A., & Kuncoro, T. (2012). Product inventory predictions at small medium enterprise using market basket analysis approach-neural networks. *Procedia Economics and Finance*, *4*, 312-320.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

Pande, A., & Abdel-Aty, M. (2009). Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool. *Safety science*, *47*(1), 145-154.

Raorane, A. A., Kulkarni, R. V., & Jitkar, B. D. (2012). Association rule–extracting knowledge using market basket analysis. *Research Journal of Recent Sciences*, 1(2), 19-27.

Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing*, *76*(3), 367-392.

Tang, K., Chen, Y. L., & Hu, H. W. (2008). Context-based market basket analysis in a multiple-store environment. *Decision Support Systems*, *45*(1), 150-163.

Trnka, A. (2010, June). Market basket analysis with data mining methods. *Networking and Information Technology (ICNIT), 2010 International Conference on* (pp. 446-450). IEEE.

Wick, M. R., & Wagner, P. J. (2006, March). Using market basket analysis to integrate and motivate topics in discrete structures. In *ACM SIGCSE Bulletin*(Vol. 38, No. 1, pp. 323-327). ACM.

Yun, C. H., Chuang, K. T., & Chen, M. S. (2006). Adherence clustering: an efficient method for mining market-basket clusters. *Information Systems*, *31*(3), 170-186.