ISSN: 2321-7782 (Online)
Volume 3, Issue 8, August 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Customer Shopping in an Online with Data Mining Approach

D. Ranjith
Research Scholar,
Department of Computer Science,
H.H. The Rajah's College (Autonomous)

Abstract: ***Data mining techniques are expected to be a more effective tool for analyzing customer behaviors. Accurate prediction of shopping channel preference has become an important issue for retailers seeking to maximize customer loyalty. Data mining tools answer business questions that in the post were too time consuming to pursue. Yet, it is answer to these questions make customer relationship management possible. The objectives of this paper are to identify the high-profit, high-value, and low-risk customer by one of the data mining techniques-customer clustering. Association rule is employed to mine rule for trusted customer using sales data in a super market industry.***

Keywords: Data mining, Prediction, CRM, Customer clustering, Association rule.

I. INTRODUCTION

The study of the person who wants to buy certain product from marketing place, at the time of purchasing product, why he wants to buy it? The continuous development of the internet is challenging markets to analyze the heterogeneous transactions effects of consumer. Marketing research indicates that the online is changing the way in which customers use different information and shopping channels before purchasing a product[1]. Gathering information on homogeneous customer movement and predicting their decision processes is traditionally conducted in the domain of marketing modeling and more recently in data mining [2,3]. The Customer Relationship Management (CRM) techniques afford new opportunities for businesses to act on the concepts of relationship purchase. The previous model of "design-build-sell" is being replaced by "sell-build-redesign". In business-to-business (B2B) environments a tremendous amount of information is exchanged on a regular basis. This is particularly important in dynamic and ever-changing markets, where customers are driven by ever changing market competition and demands. While marketing modeling focuses on descriptive and normative modeling to provide explanations of the impact and relationship of independent variables upon class membership within a priori defined models, data mining seeks to identify relationships directly from the data to facilitate predictive modeling [4,5]. As a results, useful information is often overlooked, and the segmentations and the potential benefits of increased computational and data gathering capabilities are only partially realized. It is extract useful pattern and association from customer data [6]. Data mining techniques like clustering and associations can be used to meaningful patterns for future prediction [7,8]. Customer clustering and segmentation are two of the most important techniques used in marketing and customer relationship management. The major customer characteristics that are used to measure purchase behavior of customer include Recency, Frequency and Monetary values (RFM). Recency tell how long it has been since each customer made the last shopping. Frequency tells how many times each customer has purchased an item during certain intervals of time. Monetary tells how much each customer has spent in total. Monetary measures the total expenditure of the customer for a number of transactions over a period of time.

II. DATA MINING WITH FREQUENT PATTERN

Data mining evolved from a simple taking out of row data to an analytical process from large amount of data in order to collect knowledge. It can be done in six steps. There are,

1. Business understanding Business understanding includes determining business objectives, assessing the current situation, establishing data mining goals and developing a project plan.
2. Data Understanding Once business objectives and the project plan are established, data understanding considers data requirements. This step can include initial data collection, data description, data exploration and the verification of data quality. Data exploration such as viewing summary statistics can occur at the end of this phase.
3. Data Preparation Once the data resources available are identified, they need to be selected, cleaned, built into the form desired and formatted. Data cleaning and data transformation in preparation of data modeling needs to occur in this phase.
4. Modeling Data mining software tools such as visualization and cluster analysis are useful for initial analysis. Tools such as generalized rule induction can develop initial association rules.
5. Evaluation Model results should be evaluated in the context of the business objectives established in the first phase (business understanding). This will lead to the identification of other needs, frequently reverting to prior phase of CRISP-DM.
6. Deployment Data mining can be used to both verify previously held hypotheses, or knowledge discovery. The CRISP-DM process, sound models can be obtained that may than be applied to business operations for many purposes, including prediction or identification of key situations.

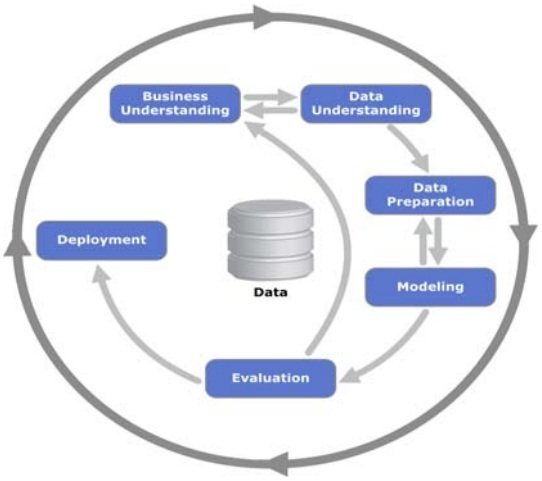

Figure 1. CRISP data mining process model.

III. TEMPORAL DATA MINING

A mining approach with respect to time is known as temporal data mining. It involves application of data mining technique on temporal uses data in order to discover temporal pattern and it also deal with temporal relationship of user or data. Temporal mining is a data mining that focus on time aspects and deal with the conclusion of temporal patterns from large data sets. The use of temporal data miming is continuous to raise as huge amounts of temporal data of everyday's behavior become present easily.

Temporal data mining is one step of knowledge discovery process from temporal data base. It calculates temporal patterns or models over the temporal data. The algorithm used in this process known as temporal mining algorithm.

Results of temporal association patterns for temporal data mining can be effectively applied in number of fields like trading, marketing, social networking, medicos, earth quick detection, robotics and assisted design. So that explorer's number of efficient algorithms for temporal data mining like symbolic time series, symbolic time sequences, symbolic interval series, numeric time series, item set sequences, etc. have proposed.

IV. TEMPORAL ASSOCIATE RULE

Association rules are often written as A→B meaning that whenever A appears B also tends to appear. A and B may be single items or sets of items. A is often referred to as the rule's antecedent and B as the consequent. A→B is a probabilistic relationship found empirically. The set of transactions, each transaction being a list of items. Items A and B appear together in only 10% of the transactions but whenever A appears there is an 80% chance that B also appears. The 10% presence of A and B together is called the support of the rule and the 80% chance is called the confidence of the rule.

Support(A) = (Number of times A appears)/N = P(A)

Support(AB) = (Number of times A and B appear together)/N = P(A∩B)

Confidence for A→B is defined as the ratio of the support for A and B together to the support for A.

Confidence of (A→B) = Support(AB)/Support(A) = P(A∩B)/P(A) = P(B | A)

P(A | B) is the probability of B once A has taken place, also called the conditional probability of B.

V. THE VQ BASED CLUSTERING ALGORITHM

It is an efficient algorithm designed by Linde, Buzo and Gray for the design of good block or vector quantizers with quite general distortion measurements is developed for use on either known probabilistic source descriptions of data [9]. It works well even when the distribution has discrete components, as is the case when a sample distribution obtained from a training sequence is used.

An N-level k-dimensional 'quantizer' is a mapping, q; that assigns to each input vector, x = (x_0, -, X_{k-1}), a reproduction vector, $x^{\square}$ = q(x), draw from a finite reproduction alphabet,{ A = bi; I = 1, -, N} [9]. The level N describes the number of times the division of the codebook occurs. The quantizer is completely described by the reproduction alphabet (or codebook). A together with the partition, S = {S_i; I = 1,....., N), of the input vector space into the sets s_i = { x: q(x) = v_i) of input vectors mapping into the reproduction vector [9], such quantizers are also called block quantizers, vector quantizers, and block source codes[9].

Here the input vectors x can be any kind of customer RFM values. Here it is assumed that the distortion caused by reproducing an input vector x by a reproduction vector i is given by a nonnegative distortion measure d(x, $x^{\square}$). Many such distortion measures have been proposed in the literature. The most common for reasons of mathematical convenience is the squared error distortion, which has been used in the implementation of the algorithm.

$$d(x, x^{\square}) = \sum_{i=0}^{k-1} |xi - xi|$$

An N-level quantizer will be said to be optimal if it minimizes the expected distortion, that is, is optimal if for all others quantizers 4 having N reproduction vectors D(q*) < D(q) [9]. A quantizer is said to be locally optimum if D(q) is only a local minimum, that is, slight changes in q cause an increase in distortion [9]. The goal of block quantizer design is to obtain an optimal quantizer if possible and, if not, to obtain a locally optimal and hopefully "good" quantizer [9]. Several such algorithms have been proposed in the literature for the computer-aided design of locally optimal quantizer [9]. A brief view of the steps of this algorithm is given below [9].

A) ***ALGORITHM VQ :***

1. Initialization: Given N = number of levels, a distortions threshold $\epsilon \geq 0$, and an intial N-level reproduction alphabet A_0, and a distribution F, Set m= 0 and $D_{-1} = \infty$.
2. Given Am = { y_i ; I = 1,…., N}, find its minimum distortion partition $P(A_m)$ = { S_i ; I = 1…N}: x ϵ S_i if d(x,y_i) for all j. compute the resulting average distortion, Dm = D({A_m,$P(A_m)$}) = E $\min_{EAm}$ d(X,Y).

3. If $(D_{m-1} - D_m)/D_m < \epsilon$, halt with A, and $P(A_m)$ describing final quantizer. Otherwise continue.

4. Find the optimal reproduction alphabet $d(P(A_m)) = \{ X^{\square}(S_i); i = 1,\ldots,N\}$ for $P(A_m)$. Set $Am \equiv X^{\square}(P(A))$. Replace m by m+1 and go to 1.

Earlier this algorithm was mainly used for image compression and other related works. But, I found it useful for my clustering approach.

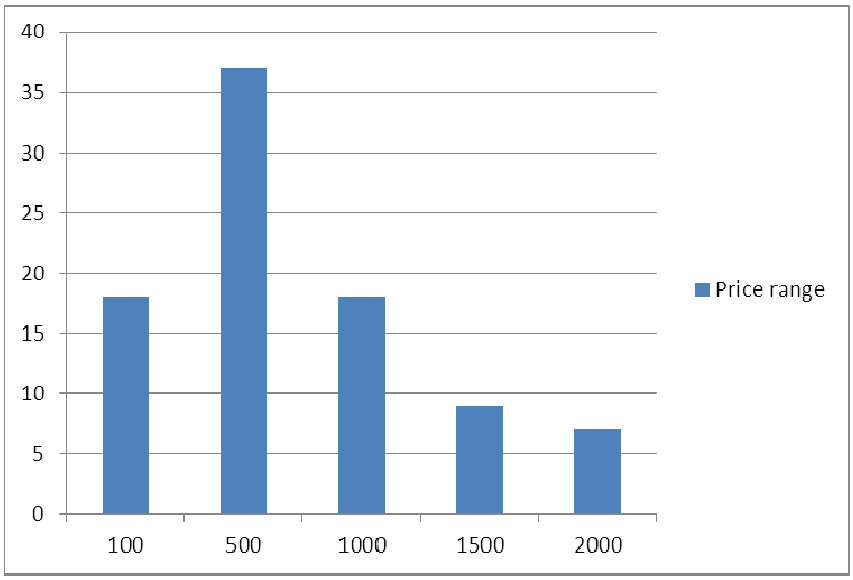

Fig 2. The output shown in a graphical form. X-axis represents the price range and Y-axis represents the number of customer

VI. FUTURE WORK

The Association rule algorithm implemented here takes up a large of execution time. So, the optimization of the algorithm can be done to ensure a better performing algorithm. The VQ algorithm implemented here doesn't take into account the other aspects like Recency, Frequency and Monetary values. Either, these values can be taken into account one at a time. It here a simple database was used, so there isn't much space for taking the frequency values into account. But it can be extended to a larger and more comprehensive database to analyze the aforementioned values.

VII. CONCLUSION

VQ approach can be basically used to segment customers, according to any of the RFM values, or all of them together. It needs the initial vectors as its input for it to start creating the clusters. The change of the level N of the algorithm results in more number of segmentations, and is to be adjusted according to the requirements of the clustering approach. The threshold value 'ϵ' strictly keeps a check on the difference between the customer's monetary value (or for that matter, any other value required) and its respective quantizing value found by the algorithm. In terms of predictions, the clusters obtained can show the different segments of customers and the more populated segments can be targeted specifically.

References

1. "A perpect market. Survey: E-Commerce," in Economist. London, 2004.
2. M.H. Dunham, Data mining introductory and advanced topics. Upper Saddle River, N.J.: Prentice Hall, 2003.
3. I.H. Written and E. Frank, Data mining: practical machine learning tools and techniques, 2nd ed. Boston, MA, 2005.
4. P.S.H. Leeflang and D. R. Wittink, "Building models for marketing decisions: Past, present and future, "Int. Journal of Research in Marketing, vol. 17, pp. 105, 2000.
5. J.-B.E.M. Steenkamp, "Introduction: Marketing Modeling on the Threshold of the 21st Century," International Journal of Research in Marketing, vol.17, pp. 99, 2000.
6. Association Analysis of Customer Services from the Enterprise Customer Management System-ICDM-2006.
7. Terry Harris, (2008) "Optimization creates lean green supply chains", Data Mining Book.

8. Matt Hartely (2005) " Using Data Mining to predict inventory levels" Data Mining Book.

9. Yoseph Linde, Andres Buzo, Robert M. Gray : An Algorithm for Vector Quantizer Design, IEEE Transactions on communications, vol. com-28, no. 1, (january 1980), pp. 84-86.

10. Danuta Zakrzewska, Jan Murlewski : Clustering Algorithms for Bank Customer Segmentation, 5th International Conference on Intelligent Systems Design and Applications, (2005),pp 1-2.

11. Abdullah Al-Mudimigh, Farrukh Saleem, Zahid Ullah Department of Information System: Efficient implementation of data mining: improve customer's behavior, 2009 IEEE, (2009),pp.7-10.

12. Sung Ho Ha , Sang Chan Park, Sung Min Bae : Customer's time-variant purchase behavior and corresponding marketing strategies: an online retailer's case, Computers & Industrial Engineering 43 (2002) 801– 820, (2002),pp.801-806.

13. Euiho Suh, Seungjae Lim, Hyunseok Hwang, Suyeon Kim : A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study, Expert Systems with Applications 27 ,(2004), pp. 245-250.

14. Mu-Chen Chen , Hsu-Hwa Chang, Ai-Lun Chiu : Mining changes in customer behavior in retail marketing, Expert Systems with Applications 28 ,(2005), pp. 773-776.

15. Sriram Thirumalai, Kingshuk K. Sinha : Customer satisfaction with order fulfillment in retail supply chains: implications of product type in electronic B2C transactions, Journal of Operations Management 23 ,(2005), pp. 291-296.

