Using Data Mining Technique Predicting Customer Purchase in An Online Shopping Business

Khirendra Patel[1], Mr. Leelkanth Dewangan [2], Mr.Ghanshayam Sahu[3]
[1] B.C.E.T. Durg
[2, 3] Professor, B.C.E.T. Durg

Abstract- *Observing the segments of shoppers and their behavioural patterns over totally different time intervals, is a vital application for businesses, particularly just in case of the last tier of the net distributor that thinks about with "electronic Business-to-Customer relationship " . This can be notably vital in dynamic and changing markets, wherever customers square measure driven by ever ever-changing market competition and demands. Also, the availability of bespoke service to the shoppers is significant for a corporation to ascertain long lasting and pleasant relationship with customers. It's additionally been ascertained that keeping existing customers generates a lot of profit than attracting new ones. So, client retention may be a huge issue too. So, there's continuously a trade-off between client edges and group action prices, that must be optimized by the managers.The purpose of this thesis is to check, Associate in Nursingalyze numerous Data-mining tools and techniques so do an analysis of the sample / information to get a significant interpretation.*

Keywords- Data mining, , Clustering, Segmentation.

I. INTRODUCTION

1.1 Need for Customer Behavior prediction and data mining.

The emergence of the business-to-customer (B2C) markets has resulted in numerous studies on developing and up client retention and profit improvement. this is often in the main attributable to the retail business changing into more and more competitive with prices being driven down by new and existing competitors. In general, shopper markets have many characteristics like repeat shopping for over the relevant measure, an outsized variety of shoppers, and a wealth of data particularisation past client purchases. In those markets, the goal of CRM is to spot a client, perceive and predict the customer-buying pattern, determine associate acceptable supply, and deliver it during a customized format on to the client

1.2 Relevance of Data mining towards CRM.

ata mining techniques area unit the processes designed to spot and interpret knowledge for the aim of understanding and deducing unjust trends and planning methods supported those trends [3]. data processing techniques extract the data, so rework them to urge the reworked knowledge, so get pregnant patterns among the reworked knowledge. As businesses value their investments on promoting activities, they have a tendency to target their data processing techniques and capability. . the essential structure of CRM model life cycle is shown in fig.1. The model will have 2 initiating points. Firstly, the client will some purchase so the information is measured and evaluated. Afterwards, the corporate mines the evaluated knowledge so they will have associate degree understanding of the patterns that the client shows whereas getting. With the assistance of that knowledge, the organization will formulate its steps to maximise or optimize its business plans. Secondly, the organization takes some action for rising the customer's satisfaction by creating a decent informative provide, so studies the actions taken by the client. Then the actions of the client area unit once more evaluated associate degreed an understanding of the client is achieved.

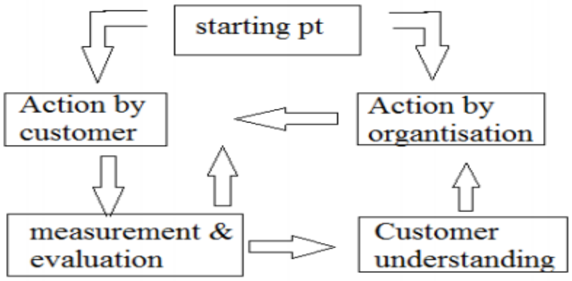

Fig1. The basic CRM cycle . [3]

II. LITERATURE REVIEW

2.1 The Customer Segmentation approach

2.1.1 An Introduction to clustering and customer segmentation.

Customer segmentation is one in every of the foremost vital space of knowledge-based selling. just in case of on-line retail stores, it's very a difficult task, as information bases ar massive and multidimensional . within the thesis we have a tendency to think about a clump rule, that primarily based|is predicated|relies} on a Vector quantisation based rule, and may be effectively wont to mechanically assign existing or new incoming customers into the several clusters .

2.1.2 A background of the Vector Quantization based clustering algorithm.

".It is Associate in Nursing economical algorithmic rule designed by Linde, Buzo and grey for the look of excellent block or vector quantizers with quite general distortion measurements is developed to be used on either legendary probabilistic supply descriptions or on an extended coaching sequence of knowledge, incorporated herein by reference, [1]. The algorithmic rule involves no differentiation; thence it works well even once the distribution has distinct parts, as is that the case once a statistical distribution obtained from a coaching sequence is employed. like the common variational techniques, the algorithmic rule produces a quantizer meeting necessary however not spare conditions for optimality. Usually, however, a minimum of native optimality is ensured in each approaches.

ALGORITHM:

1. Initialization: Given N = number of levels, a distortion threshold   0, and an initial N-level reproduction alphabet A0, and a distribution F. Set m = 0 and D-1 = ∞.
2. Given Am ={ yi ; i = 1, …, N}, find its minimum distortion partition P(Am) = {Si; i = 1... N}: x Si if d(x,yi) d(x, yj) for all j. Compute the resulting average distortion, Dm =D({Am, P(Am)}) = E minEAm d(X,y).
3. If (Dm-1 - Dm)/Dm < , halt with A, and P(Am) describing final quantizer. Otherwise continue.
4. Find the optimal reproduction alphabet d(P(Am)) = {x^(si); i = 1,…, N} for P(Am). set Am x^(P(Am)). 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.

2.1.3 The explanation of the VQ algorithm.

This algorithm for the study of customer segregation consisted of the following components. Firstly, I two double arrays ‘c’ and ‘q’ were taken to store the code-vectors and the quantizers values. The an initial ‘cref’ value is taken as the initial reference value for the codebook, which is the average of all the input vectors x. The input vectors are the data obtained from the sample database, which can be any of the RFM values or all of them. An initial ‘dref’ value is taken as the mean of all the mean squared distortion values of the input vectors and the quantizers. The order of the Vector quantization algorithm has been denoted by ‘n’. The threshold for distortion value is denoted by ‘e’.

Initially, the reference code-vector is split into 2 new code vectors by the following process, shown by a pseu code below :

```
 c[p-1] = (1.0+e)*cref[p-1];
 c[n-p+1] = (1.0-e)*cref[p-1];
```

Assigning of quantizers to the individual vectors is done by the following method :

```
   if((sum[s]-c[p])*(sum[s]-c[p])<min)
          min=(sum[s]-c[p])*(sum[s]-c[p]);
          q[s]=c[p];
```

updation of code vectors for the next iteration is done as follows :

```
 if (q[s]==c[p])
 sum3=sum3 + sum[s];
 cnte=cnte +1.0;
 if(cnte==0.0)
         c[p]=sum3;
 else
                 c[p]=sum3/cnte;
```

2.2 The Association rules primarily based approach for client purchase predictions.

2.2.1 A background of the Association rules primarily based data processing approach.

Association rules measure like classification criteria. There square measure usually the left-hand facet of the rule, referred to as the antecedent and also the hand facet of the rule, referred to as the logical thinking half.Association rules were at first applied to research the relationships of product things purchased by customers at retail stores. In data processing, association rules square measure descriptive patterns of the shape X=>Y, wherever X associated Y square measure statements concerning the values of attributes of an instance during a info. X is termed the left-hand-side (LHS), associated is that the conditional a part of an association rule.Meanwhile, Y is named the right-hand-side (RHS),and is that the resultant half. the foremost typical application of association rules is market basket analysis, within which the market basket contains the set of things (namely itemset) purchased by a client throughout one store visit. it's additionally referred to as the pushcart analysis of the client purchases,The process of Association rule mining approach typically finds out an enormous variety of rules. These rules square measure then cropped down on the premise of their “Coverage” worth, that is outlined because the variety of instances from the entire set, wherever the rule predicts properly, and their accuracy (the same variety expressed because the proportion of instances to that the rule properly applies). Nowadays, what we tend to decision coverage is usually referred to as as “support” and what we tend to decision accuracy is additionally referred to as “confidence”. we tend to square measure solely curious about association rules with high coverage values.

The distinction between LHS and RHS of the association rules obtain combos of attribute – worth pairs that have predefines minimum coverage. These square measure referred to as item-sets. associate attribute worth combine is named associate item. It generally comes from the method of “Market basket analysis” wherever the shop manager analyzes the various things purchased by the client during a single purchase, and tries to seek out out association rules among them.

2.2.2 A Description of the Association rules mining model projected here.

This study involves the Apriori rule, accustomed find rules and prune them per their coverage. ‘Apriori’ is that the most simple rule for learning association rules.Apriori is intended to control on databases containing numerous sorts of transactions (for example,collections of things bought by customers, or details of an internet site surfing). As is common in association rule mining, given a group of item-sets (for instance, sets of retail transactions, every listing individual things purchased, during this study), the rule tries to search out subsets that ar common to a minimum of a minimum range K of the itemsets [8]. Apriori uses a "bottom up" approach for its execution, wherever the desired subsets ar extended one item at a time (a step called candidate generation), and such candidates ar tested against the info. The rule terminates once no any prosperous things may be additional to the prevailing itemsets. ‘Apriori’,while terribly basic and traditionally necessary, suffers from variety of shortcomings or trade-offs. Candidate generation step generates massive numbers of subsets (the rule tries to fill the candidate set with as several as doable before every scan of the database). And with increase within the range of itemsets, the procedure over head additionally will increase.It has been thought of here, the appliance of the Apriori rule for progressive stages, first for a 1-itemset, then 2-itemset and at last for a 3-itemset, every of whom have shown varied results (discussed within the later chapter). The one itemset approach apllies the formulation of rules supported the standards on “if item X purchased”. that's it tries to search out out the prevalence of individual things from variety of orders. Then the results ar keep , and a pool of rules is obtained that ar denoted by “occurrence” here.

III. METHODOLOGY

The implementation of the client segmentation used a distinct table for the VQ approach. It used the client column and also the freight column of the ‘Orders’ knowledge table and so used that knowledge to cluster the shoppers supported the various levels or clusters supported the worth of freight.

This study involves the Apriori rule, accustomed observe rules and prune them per their coverage. ‘Apriori’ is that the most simple rule for learning association rules.Apriori is intended to work on databases containing numerous styles of transactions (for example,collections of things bought by customers, or details of a web site surfing).

IV. CONCLUSION

VQ approach is essentially accustomed section customers, in step with any of the RFM values, or all of them along. It desires the initial vectors as its input for it to start out making the clusters. The client purchase patterns approach, victimization the association rules mining technique, is a good method of extracting the foundations from the data and inferring the shopping for patterns among them.The implementation shows that increasing the “coverage” values leads to higher pruning of rules, and a a lot of trustworthy rule set. From the association rules with enough coverage, we will predict that merchandise the client tends to shop for in conjunction with the acquisition of specific merchandise.

REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Ian H. Witten & Eibe Frank : Data Mining : Practical machime learning tools and techniques, San Francisco, Morgan Kaufmann publishers, (2005), pp. 112-118,136-139.