Building Data Mining Application for Customer Relationship Management

Deeksha Bhardwaj
Department of Computer Engineering
G.H. Raisoni Institute of Engineering and Technology, Pune, Maharashtra, India

Anjali Kumari
Department of Computer Engineering
G.H. Raisoni Institute Of Engineering And Technology, Pune, Maharashtra, India

Navjot Kour
Department of Computer Engineering
G.H. Raisoni Institute Of Engineering And Technology, Pune, Maharashtra, India

Abstract- A customer service database usually stores two types of service information: (1) unstructured customer service reports record machine problems and its remedial actions and (2) structured data on sales, employees, and customers for day-to-day management operations. This paper investigates how to apply data mining techniques to extract knowledge from the database to support two kinds of customer service activities: decision support and machine fault diagnosis. A data mining process, based on the data mining tool DB Miner, was investigated to provide structured management data for decision support. Data warehousing implements the process to access heterogeneous data from data resources. Data mining is the process of analyzing data from different prospective and summarizing it into useful information. Data mining software is one of the numbers of analytical tools for analyzing data. Technically, data mining is the process of finding correlations or patterns among number of fields in large relational databases. Our paper focuses on need for information repositories and discovery of knowledge, working of data warehousing and data mining. The paper also focuses on how data mining can be used in conjunction with a data warehouse to help in certain types of decisions.

Keywords – Data mining, Knowledge discovery in databases, Customer service support, Decision support, Data Warehouse.

I. INTRODUCTION

The way in which companies interact with their customers has changed dramatically over the past few years. A customer's continuing business is no longer guaranteed. As a result, companies have found that they need to understand their customers better, and to quickly respond to their wants and needs. In addition, the time frame in which these responses need to be made has been shrinking. It is no longer possible to wait until the signs of customer dissatisfaction are obvious before action must be taken. To succeed, companies must be proactive and anticipate what a customer desires.

Successful companies need to react to each and every one of these demands in a timely fashion. The market will not wait for your response, and customers that you have today could vanish tomorrow. Interacting with your customers is also not as simple as it has been in the past. Customers and prospective customers want to interact on their terms, meaning that you need to look at multiple criteria when evaluating how to proceed. You will need to automate:

- The Right Offer
- To the Right Person
- At the Right Time
- Through the Right Channel

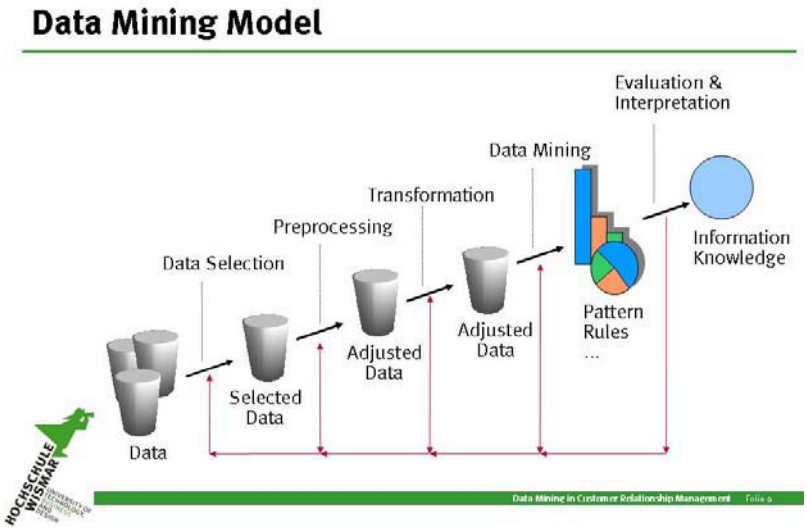


Figure1. Data Mining Model

II. PROPOSED ALGORITHM

Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

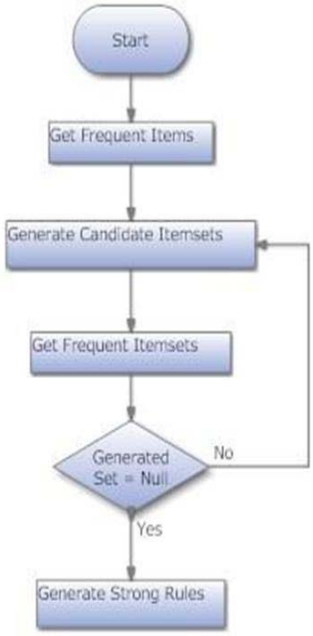
Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing). Each transaction is seen as a set of items (an *item set*). Given a threshold$C$, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database.

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length $k-1$. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent items among the candidates.

The pseudo code for the algorithm is given below for a transaction database T and a support threshold ϵ . Usual set theoretic notation is employed; though note that T is a multi set. C_k is the candidate set C for level K. Generate() algorithms is assumed to generate the candidate sets from the large item sets of the preceding level, heeding downward closure lemma. Count[c] accesses a field of a data structure that represents candidate set C, which is initially assumed to be zero. Many details are omitted below , usually the most important part of the implementation is the data structure used for storing candidate sets and counting their frequencies.

Assume that a large super markets tracks sales data by stock keeping unit(SKU) for each item , such as ”butter” or “bread”, is identified numerical SKU. The super market has a database of transactions where each transaction is a set of SKUs that were brought together.



Example:

Let the database of transactions consist of the sets {1,2,3,4}, {1,2}, {2,3,4}, {2,3}, {1,2,4}, {3,4}, and {2,4}. We will use Apriori to determine the frequent item sets of this database. To do so, we will say that an item set is frequent if it appears in at least 3 transactions of the database: the value 3 is the support threshold.

The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately, by scanning the database a first time. We obtain the following result

Step 1:

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

All the item sets of size 1 have a support of at least 3, so they are all frequent.

Step 2: The next step is to generate a list of all pairs of the frequent items:

Item	Support
{1,2}	3
{1,3}	1
{1,4}	2

{2,3}	3
{2,4}	4
{3,4}	3

The pairs {1, 2}, {2,3}, {2,4}, and {3,4} all meet or exceed the minimum support of 3, so they are frequent. The pairs {1,3} and {1,4} are not. Now, because {1, 3} and {1,4} are not frequent, any larger set which contains {1,3} or {1,4} cannot be frequent. In this way, we can prune sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:
Step 3:

Item	Support
{2,3,4}	2

In the example, there are no frequent triplets -- {2, 3, 4} is below the minimal threshold, and the other triplets were excluded because they were super sets of pairs that were already below the threshold.
We have thus determined the frequent sets of items in the database, and illustrated how some items were not counted because one of their subsets was already known to be below the threshold.

III. Experiment and Result

For this study, the transaction of data of our organization retails smart store has been taken. Using these data, customers have been clustered using IBM Intelligent Miner tool. The first steps in the clustering process involve selecting the data set and the algorithm. There are two types of algorithms available in I-Miner Process.

i. Demographic Clustering process
ii. Neural Clustering process

In this exercise, the demographic Clustering process has been chosen, since it works best for the continuous data types.
The data set has all the data types are continuous. The next step in the process is to choose the basic run parameters for the process. The basic parameters available for demographic clustering process include are;

- Maximum number of clusters
- Maximum number of passes through the data
- Accuracy
- Similarity Threshold

For this assignment , maximum number of cluster is 4, maximum number of passes through the data is 3 and the accuracy is 0.5 has been chosen. The input parameters for the customer's clustering are;

1. Recency
2. Totally customer profit
3. Total customer revenue
4. Top revenue department

The top revenue department variable has been chosen as supplementary variable and rest of the variables have been chosen as active variables. The supplementary are used to profile the clusters and not to define them. The ability to add supplementary variables at the outset of clustering is a very useful feature of intelligent Miner that allows easy interpretation of clusters using data other than the input variables. The input and the output fields width are defined and the input data in mining is the production data of our organization retail smart store. The data is first extracted from the oracle databases and flat files and converted into flat files. Subsequently the I Miner process picks up the file and processed. The entire output dataset would have customer information appended to the end of each record. The clusters are ordered from top to bottom in order of size with a number of penalty made left indicate the size of a cluster as a percentage of the universe.

IV.CONCLUSION

Customer relationship management is a technology that manages the relationship with customer in order to improve the performance of business. In CRM, the customer identification is the important phase because it involves segmenting customers and analyzing their behavior for further customer attraction , retention and development. In this clustering technique data mining has used for customer segmentation and rule induction is used for describing customer behavior in each segment.Business people employ different benefits schemas for customer in different clusters and segments. So classifying a customer to the cluster plays an important role in CRM . For a good rule induction algorithm, the customers behavior in each cluster should be correctly characterized so that the new customer are predicted to the appropriate cluster.

REFERENCES

[1] Pieter Adriaans,Dolf Zantinge,(2011),Data Mining,Pearson Education Ltd,Sixth Impression,2011.
[2] Migueis,V.l.Camanho,A.S.Joao Falcao e Cunha (2012),Customer Data Mining for LifeStyle Segmentation,Expert Systems eith applications,Vol.39.3959-9366.
[3] Knowledge Management In CRM using Data Mining Tchniques,Sunil yadav,Aaditya Desai,Vandana Yadav,International Journal Of Scientific & Engineering Research, Volume 4,Issue &, July-2013.
[4] The Effect Of Clustering in Apriori Data Mining Algorithm : A Case Study, Nergis Yilmaz and Gulfem Isiklar Alptekin, Proceedings of the World Congress on Engineering 2013 Vol III,WCE 2013,July 3-5,2013,London,U.K.
[5] Data Mining: An Overview from a Database Perspective Ming-Syan Chen,Senior Member,IEEE,Jiawei Han,Senior Member,IEEE,andPhilip S. Yu,Fellow,IEEE, IEEE transaction On Knowledge And Data Engineering,Vol.8,No.8,No. 6,December 1996 .
[6] Mining Changes in customer behaviour in retail marketing,Mu-Chen Chen,Ai-Lun Chiu,Hsu-Hwa Chang,Expert System with Applications 28,(2005), 773-781.
[7] Data Mining By Heidi Kuzma And Sheila Vaidya.
[8] Margaret H. Dunham, "Data Mining Techniques And Algorithms,", Aug.23,2000.
[9] "Business Intelligence at Data Mining Techniques,"
[10] Richard Barker, "Managing Data Warehouse", Oct.2003.
[11] Berson, A., S. Smith, K. Thearling, "Building Data Mining Applications for CRM," McGraw- Hill, New York, 2000.
[12] Han, J. and M. Kamber, "Data Mining: Concepts and Techniques*,"* Morgan Kaufmann, San Francisco, 2000.
[13] M. S. Chen, J. Han, and P. S. Yu. "Data mining: An overview from a database Perspective," IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
[14] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining," AAAI/MIT Press, 1996.
[15] Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery," Third Edition (Potomac, MD: Two Crows Corporation, 1999), p. 4.
[16] George Cahlink, "Data Mining Taps the Trends," Government Executive Magazine, October 1, 2000, http://www.govexec.com/tech/articles/1000managetech.htm.