Data Mining Need In Profiling Internet Banking Customer

*Manoranjan Dash & *Sangram Kesari Nayak

*Institute of Business & Computer Studies, Siksha O Anusandhan University, Bhubaneswar, Odisha, India
*IGIT, CSE Department, Sarang, Odisha, India

Abstract - Customers are more demanding in terms of responsiveness, level of service, quality of products and costs. Banks only could reach these goals simultaneously if they invest in new technologies. A new technology that has received considerable attention from academics and practitioners, especially among banks, is the Internet banking because of its large scope of application, advantages and potentialities. Internet banking brings a drastic change in a common person’s life. It is not only the name of giving services to elite and privileged class of the society but is also to facilitate the common people, which normally represent a big proportion of the society. The objective of the paper is to use data mining techniques to analyse historical data of internet banking usages from a private sector bank in India. The technique K-Mean algorithm is used to segment customers into groups according to their demographic profiles and internet banking usages. Results of this study can be used to generate new financial service packages which are customised to each segment of internet banking users. We find that the strategy of banking groups to incorporate internet banks reflects some competitive edge that these banks have in their business models.

Keywords - K-Means, Apriori algorithm, Internet banking.

I. INTRODUCTION

Technology and innovation change have subject the revolution in the banking sector during the last decade. The adoption of technology and change has supported increased globalization of capital flows and financial organizations. Rapid technological advances have introduced significant changes in the global economic and business environment. The role of technology in banking sector can serve to reduce costs and often improve service reliability. Banks are adopting the internet platform for delivering the financial services and thereby adopting a multichannel strategy. Consumer adoption of internet financial services is an innovation in service delivery. Internet banking provides alternatives for faster delivery of banking services to a wider scope of customers. The Indian banking system has undergone a technological makeover over time. The banking system in India has created a niche for itself in the current competitive global arena where adoption of new and innovative technological development carries the key to expansion of business and its future development. The revolutionary advent of internet banking has empowered banks with new ways to cater to the needs of customers. Banks have attained more lee-way on the technology front than ever before. From just being a business enabler, IT is now a business driver for the banking sector. At the heart of the banking technology lies the Core Banking Solutions implemented by most banks. At present out of the total IT spending undertaken by banks, around 75% goes towards maintenance of existing systems and ensuring that the business of the bank goes through smoothly. Hence a major exercise in upgrading core banking architecture is something many banks may not be able to afford at present. To gain an edge over their competitors and address customer demands effectively, banks need to do a balancing act by replacing old systems with new platforms. The Indian banking system has come a long way from being a player restricted to domestic boundaries to a leading giant in the world of financial services across the globe. Credit for this goes to a mix of factors like, the introduction of various economic reforms which gave a boost to the banking sector, adoption of new technology by banks in order to stream line their line of businesses and increases profits by improving cost efficiency and offering doorstep banking convenience to their customers. Today almost all scheduled commercial banks are connected to all the branches on a real time basis extending the any where banking facility to the customers, which means that it is not necessary for a customer now to visit the branch personally in order to conduct a transaction. Now a customer has various options available with him i.e. internet banking, mobile banking, ATMs, phone banking etc. The future of Indian banking system lies in increased
investment in technology platforms and a greater focus on end-to-end solutions provided by IT majors like core banking products, vertical specific CRM and risk management software. As the new technology spreads across various sectors, businesses as well as academicians are tempted to assess the opportunities created by the new technology in a comparative manner with the existing operations. It is foreseen by many that competition will get tougher in the financial industry owing to the rapid technological change and globalization of financial markets. As increasing competition between the financial institutions has forced many of the players to offer similar prices on deposits and loans, the efforts for gaining competitive advantage were shifted towards non-price factors. In this context, new forms of distribution channels are invaluable outlets for financial institutions since they provide the opportunity for cutting costs without diminishing the existing service levels. In these days, business needs to satisfy customer’s demand to stay competitive. If products or services which customer currently buys, no longer meets his/her increasing needs, he/she will choose products or services of other producers. If business is unable to understand behaviours of its customers, it soon will lose revenue and customers. Providing products/services that customers need without customers’ request is one way to serve customers’ needs. To be able to do so, business has to understand behaviours of its customers. Information of behaviours of customers can come from customer data that business has collected. When business understands behaviours and attributes of customers, it will be able to develop products or services that satisfy customers’ demands. Banking is one of highly competitive business. Banks are attempting to create channels or products that will help distinguish themselves from competitors. To be different from competitors and be able to satisfy customer’s needs, e-banking or Internet banking is one channel that gives customers convenience and reduces customers’ costs of travelling. Internet banking has gained popularity and has been used widely in many countries.

II. DATA MINING

1. Data Warehouse

With the reduction in cost of computing power, companies are collecting all kinds of data about their customers. The repository for this large amount of data has become known as a data warehouse (Greengard 1999; Adriana and Zantinge 1996). A data warehouse is designed for decision makers strategic design support and is made up largely from parts of a operational database. This data warehouse can contain billions of records.

2. Data Mart

A specialized repository of data used by specific departments such as finance or sales and fed from an enterprise-wide data warehouse is called a data mart (Greengard 1999; Peacock 1998a). Average cost to build a data mart is usually between $10,000 and $1 million and can be up-and-running in less than six months.

3. Knowledge Discovery in Databases (KDD)

In its broadest scope data-mining is referred to as KDD (Peacock 1998a). However, data mining is generally thought of as a particular activity of KDD that applies a specific algorithm to extract patterns that help convert data into knowledge (Yoon 1999). KDD has been performed in some form since the first business enterprise, but usually on an ad-hoc, catch-as-catch-can role that supported decision makers. The difference in this past role and today is that the process is being continuously operated and is becoming central to the core of business operations (Peacock 1998a).

4. Data Mining

Data mining has been defined as the process of sifting through large amounts of data to spot patterns and trends that can be used to improve business functions. Simply put, it is prospecting for profits in the depths of the company's database or "like looking for gold in your computer" (Cohen 1999). It combines techniques from statistics, databases, machine learning and pattern recognition to extract (mine) concepts, concept interrelations and interesting patterns automatically from large business databases (Yoon 1999). The difference between data mining and other analytical methods is the approach they use in exploring the data. Most analytical tools use the verification based method - the user hypothesizes about specific relationships and tries to prove or refute the presumptions. Data mining uses what is called discovery-based approaches in which pattern matching and other algorithms are employed to determine the key relationships in the data (Weir 1998). Actually it is nothing more than the analysis of existing data to extract new or previously unknown or unrealized information. This analysis of existing data benefits both businesses and consumers as the growing capabilities of the new technique are realized (Chrys 1999). Data mining is often referred to as having two scopes. The narrow scope is defined as the automated discovery of "interesting" non-obvious patterns hidden in a database that have a potential for contributing to the overall profit of the firm. This narrow definition encompasses computer-based or "machine learning" methods such as neural networks, genetic algorithms and decision trees to extract patterns of information from data while
III. CLUSTERING

Clustering is a division of data into groups of similar objects. Each group, called a cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. Cluster analysis classifies a set of observations into two or more mutually exclusive unknown groups based on combinations of interval variables. The purpose of cluster analysis is to discover a system of organizing observations, usually people, into group, where members of the groups share properties in common. It is cognitively easier for people to predict behavior or properties of people or objects based on group membership, all of whom share similar properties. It is generally cognitively difficult to deal with individuals and predict behavior or properties based on observations of other behaviors or properties. Cluster analysis classifies unknown groups while discriminant function analysis classifies known groups. The procedure for doing a discriminant function analysis is well established. There are few options, other than type of output, that need to be specified when doing a discriminant function analysis. Cluster analysis, on the other hand, allows many choices about the nature of the algorithm for combining groups. Each choice may result in a different grouping structure.

IV. K-MEANS CLUSTER ANALYSIS

This procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that can handle large numbers of cases. However, the algorithm requires you to specify the number of clusters. You can specify initial cluster centres if you know this information. You can select one of two methods for classifying cases, either updating cluster centres iteratively or classifying only. You can save cluster membership, distance information, and final cluster centres. Optionally, you can specify a variable whose values are used to label case-wise output. You can also request analysis of variance F statistics. While these statistics are opportunistic (the procedure tries to form groups that do differ), the relative size of the statistics provides information about each variable's contribution to the

The k-means Algorithm

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation’s proximity to the mean of the cluster. The cluster’s mean is then recomputed and the process begins again. Here’s how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centres ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

Choosing the Number of Clusters

One of the main disadvantages to k-means is the fact that you must specify the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance. For example, if you had a group of people that were easily clustered based upon gender, calling the k-means algorithm with k=3 would force the people into three clusters, when k=2 would provide a more natural fit. Similarly, if a group of individuals were easily clustered based upon home state and you called the k-
K-means algorithm with k=20, the results might be too
generalized to be effective.

V. K-MEANS ALGORITHM

K-means algorithm is the simplest clustering
algorithm and widely used. K-means requires an input
which is a predefined number of clusters. This input is
named k. The steps of the K-means algorithm are given
below.

1. Select randomly k points to be seeds for the
centroids of k clusters.
2. Assign each point to the centroids closest to the
point.
3. After all points have been assigned, recalculate new
centroids of each cluster.
4. Repeat step 2 and step 3 until the centroids no
longer move.

VI. EVALUATION OF CLUSTERING

Three methods used for evaluating the efficiency of
clustering are as follows.

1. Standard Deviation (SD)
   The standard deviation is the most commonly used
   for measuring the variation of values in the defined
dataset. The lower SD value means the better clustering.

2. Root Mean Square Standard Deviation (RMSSTD)
   The RMSSTD is the variance of the clusters; RMSSTD
   measures the homogeneity of the clusters to
   identify homogenous groups, the lower RMSSTD value
   means the better clustering.

3. R Squared (RS)
   RS is used to measure the dissimilarity of clusters.
   Formally it measures the degree of homogeneity degree
   between groups. The values of RS range for 0 to 1
   where 0 means there is no difference among the clusters
   and 1 indicates that there are significant difference
   among the clusters.

VIII. DATA ANALYSIS AND INTERPRETATION

A systematic method was used to collect internet
banking transactions of private bank in Odisha. The
target data were internet banking transactions from web
log file of that bank. There were 5,195 transactions
collected from 1st to 31st of July 2009. Each session
gave details of web usage including user accounts of
those who accessed the web sites, requested web pages
and their order, and the period of time pages were
viewed. This data were used as the basis for analysis in
this study. Four factors used in data segmentation
included: (1) Date, (2) Time, (3) Status of Transaction, and
(4) Type of Transaction.

1. Date was divided in 2 groups:
   Date 1 - 1st - 15th of month
   Date 2 - 16th to the last day of the month
2. Time was divided in 4 groups:
   Interval 1 - 00.00 hrs. To 5.59 hrs.
   Interval 2 - 06.00 hrs. To 11.59 hrs.
   Interval 3 - 12.00 hrs. To 17.59 hrs.
   Interval 4 - 18.00 hrs. To 23.59 hrs.
3. Status of Transaction was divided in 2 groups:
   Status1 - Real-Time Transaction.
   Status2 - Schedule Transaction.
4. Type of Transactions was divided in 4 groups:
   Type1 - Balance Inquiry.
   Type2 - Report Financial Transactions such as
   Statement and Bill Payment Report.
   Type3 - Money Transfer.
   Type4 - Payment Transaction - Bill Payment, Payroll
   and Paid for supplier etc.

The data were analysed using SPSS 20.0 software using
the K-means techniques.
Initial Cluster Centers

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TIME</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>STATUS</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TYPE</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Iteration History (a)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Change in Cluster Centers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.807  .969  .658  .069</td>
</tr>
<tr>
<td>2</td>
<td>.000  .000  .000  .000</td>
</tr>
</tbody>
</table>

a Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 1.732.

Distances between Final Cluster Centers

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.436</td>
<td>2.941</td>
<td>2.151</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.436</td>
<td>2.237</td>
<td>2.406</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.941</td>
<td>2.237</td>
<td>1.386</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.151</td>
<td>2.406</td>
<td>1.386</td>
<td></td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Error</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATE</td>
<td>Mean Square</td>
<td>df</td>
<td>Mean Square</td>
</tr>
<tr>
<td>TIME</td>
<td>30.759</td>
<td>3</td>
<td>.026</td>
</tr>
<tr>
<td>STATUS</td>
<td>2.611</td>
<td>3</td>
<td>.182</td>
</tr>
<tr>
<td>TYPE</td>
<td>84.162</td>
<td>3</td>
<td>.082</td>
</tr>
</tbody>
</table>

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Cluster 1 has the maximum members. Customer uses the system complete a few transactions. Most usage occurs between 12.00 - 17.59pm. Most transactions are account inquiry and reports. Customers infrequently use the system for financial transactions and thus low volume and low profit.

Cluster 2 has less members. Customers use the system often but complete a few transactions. Day time usage includes account inquiry and reports. Customers use the system regularly for financial transactions. Average volume and average profit.

Cluster 3 Customers use the system regularly to complete a lot of transactions between 12.00 - 17.59pm. Transactions are normally set in advance, most are balance transfer. Use the system regularly for financial transactions, high volume and high profit.

Cluster 4 has a few members. Customers use the system occasionally between 12.00 - 17.59pm. Most transactions are account inquiry and payment transaction. Use the system occasionally for financial transactions. High volume and average profit.

IX. CONCLUSION

This research focuses on clustering internet banking customer to analyze customer characteristics and behaviors with appropriated criteria: access time, transaction access. The benefits are valuable for the bank to improve services. The research shows distinct clustering results as follows:

REFERENCES


13. http://www.ucl.ac.uk/oncology/MicroCore/HTML