IJITKM 0973-4414 Volume-9, Number-1 Jun-Dec 2015 pp. 103-110 Impact Factor 2.8 available online at www.csjournalss.com

# Data Mining Tools and Techniques used in Banking Sector

Divya Shree

**Abstract:** Banking systems collect huge amounts of data on day to day basis, be it customer information, transaction details, risk profiles, credit card details, limit and collateral details, compliance and Anti Money Laundering (AML) related information, trade finance data, SWIFT and telex messages. Organizations have been actively implementing data warehousing technology, which facilitates enormous enterprise wide databases. As a result, the amount of data that organizations possess is growing at a phenomenal rate. The next challenge for these organizations is how to interpret the data and how to transform it into useful information and knowledge. Data mining is one technology used for meeting this challenge. This article gives a comprehensive view of the technology's methods, support tools, and applications.
**Keywords:** Data mining, software, applications, tools, Banking industry, Customer Relationship Management.

## I. Introduction

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

Data mining refers to extracting knowledge from large amounts of data. The data may be spatial data, multimedia data, time series data, text data and web data. Data mining is the process of extraction of interesting, nontrivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amounts of data. It is the set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data. Using information contained within data warehouse, data mining can often provide answers to questions about an organization that a decision maker has previously not thought to ask . With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

## II. Data Mining Tools

Data mining is not all about the tools or database software that you are using. You can perform data mining with comparatively modest database systems and simple tools, including creating and writing your own, or using off the shelf software packages. Complex data mining benefits from the past experience and algorithms defined with existing software and packages, with certain tools gaining a greater affinity or reputation with different techniques. For example, IBM SPSS®, which has its roots in statistical and survey analysis, can build effective predictive models by looking at past trends and building accurate forecasts. IBM InfoSphere® Warehouse provides data sourcing, preprocessing, mining, and analysis information in a single package, which allows you to take information from the source database straight to the final report output.

It is recent that the very large data sets and the cluster and large-scale data processing are able to allow data mining to collate and report on groups and correlations of data that are more complicated. Now an entirely new range of tools and systems available including combined data storage and processing systems. You can mine data with a various different data sets, including, traditional SQL databases, raw text data, key/value stores, and document databases. Clustered databases, such as Hadoop, Cassandra, CouchDB, and Couchbase Server, store and provide access to data in such a way that it does not match the traditional table structure. In particular, the more flexible storage format of the document database causes a different focus and complexity in terms of processing the information. SQL databases impost strict structures and rigidity into the schema, which makes querying them and analyzing the data straightforward from the perspective that the format and structure of the information is known. Document databases that have a standard such as JSON enforcing structure, or files

that have some machine-readable structure are also easier to process, although they might add complexities because of the differing and variable structure. For example, with Hadoop's entirely raw data processing it can be complex to identify and extract the content before you start to process and correlate it.

It is rightfully said that data is money in today's world. Along with the transition to an app-based world comes the exponential growth of data. However, most of the data is unstructured and hence it takes a process and method to extract useful information from the data and transform it into understandable and usable form. This is where data mining comes into picture. Plenty of tools are available for data mining tasks using artificial intelligence, machine learning and other techniques to extract data. Here are six powerful open source data mining tools available:

**a. Rapidminer:**

Written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. A bonus: Users hardly have to write any code. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools. In addition to data mining, Rapid Miner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts. RapidMiner is distributed under the AGPL open source license and can be downloaded from Source Forge where it is rated the number one business analytics software.

**b. WEKA:**

The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. Its free under the GNU General Public License, which is a big plus compared to RapidMiner, because users can customize it however they please.

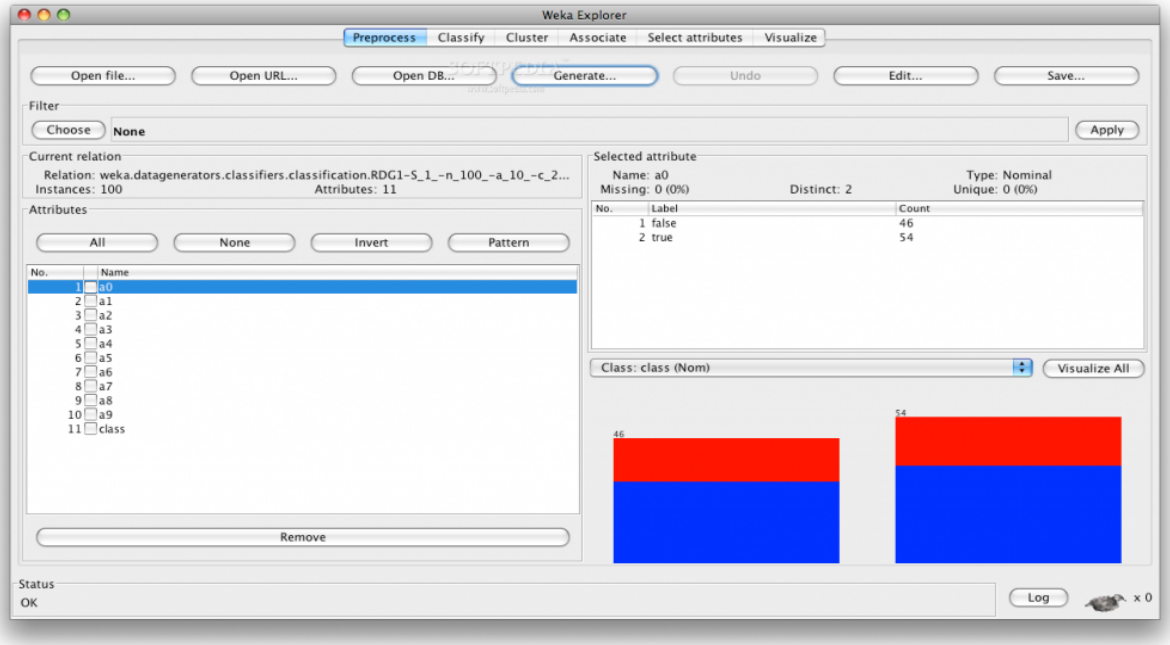


**Figure 1: WEKA Browser**

WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modeling, which currently is not included.

**c. R-Programming:**

What if I tell you that Project R, a GNU project, is written in R itself? It's primarily written in C and Fortran. And a lot of its modules are written in R itself. It's a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years. Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

**d. Orange:**

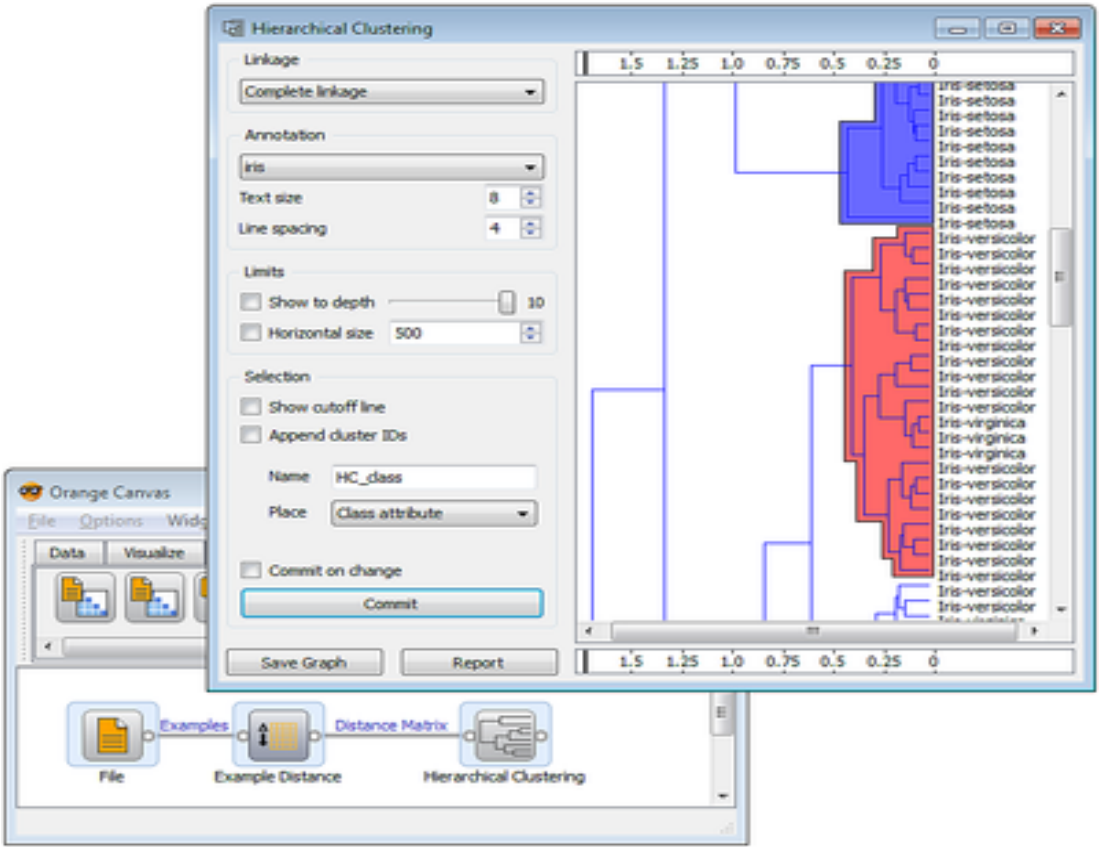


**Fig. 2: Orange Hierarchical Clustering**

Python is picking up in popularity because it's simple and easy to learn yet powerful. Hence, when it comes to looking for a tool for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and experts. You will fall in love with this tool's visual programming and Python scripting. It also has components for machine learning, add-ons for bioinformatics and text mining. It's packed with features for data analytics.

**e. KNIME:**

Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis. Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version.

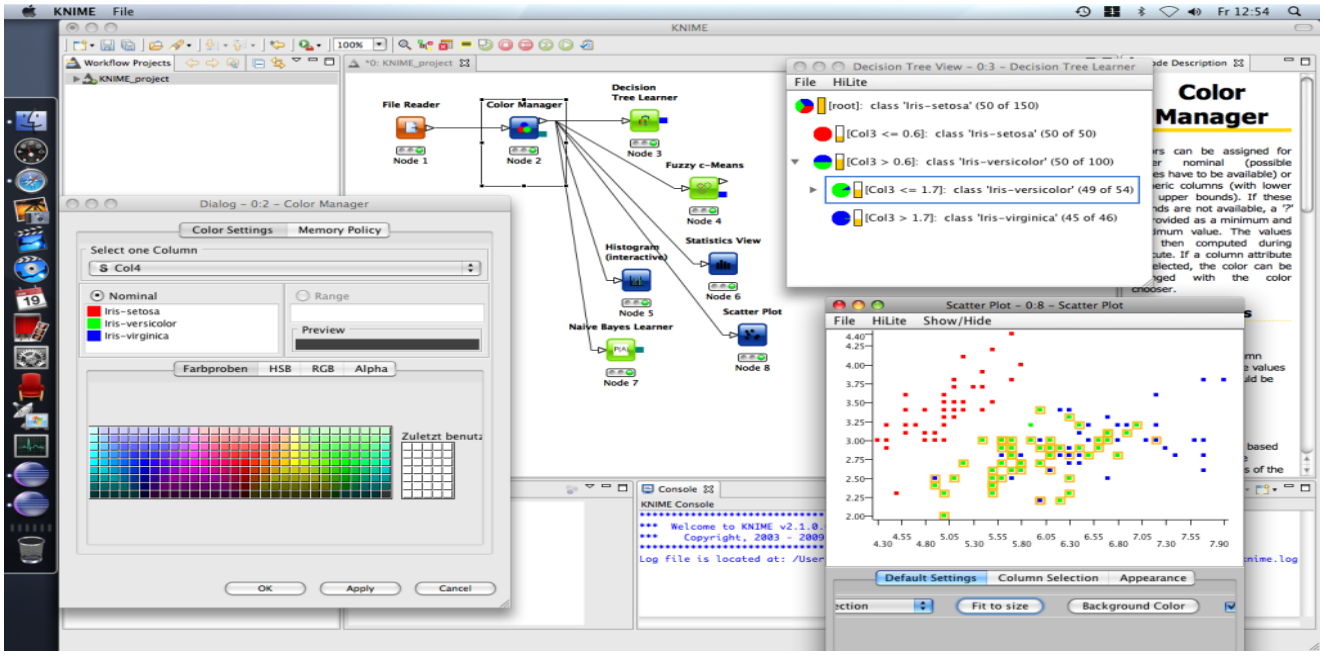


**Fig. 3: KNIME Browser**

**f. NLTK:**

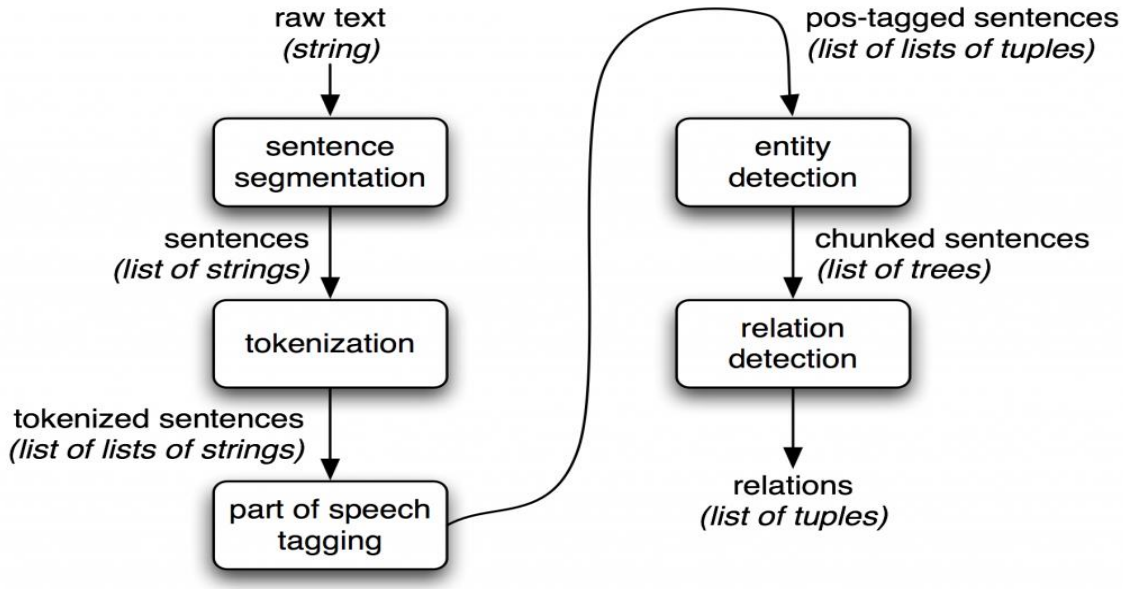


**Fig. 4: NLTK Tool**

When it comes to language processing tasks, nothing can beat NLTK. NLTK provides a pool of language processing tools including data mining, machine learning, data scraping, sentiment analysis and other various language processing tasks. All you need to do is install NLTK, pull a package for your favorite task and you are ready to go. Because it's written in Python, you can build applications on top if it, customizing it for small tasks.

## III. Data Mining Techniques

There are several major data mining techniques have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree. We will briefly examine those data mining techniques in the following sections.

### Association

Association is one of the best-known data mining techniques. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for customer and increase sales.

### Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

### Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.

### Prediction

The prediction, as its name implied, is one of a data mining techniques that discover the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

### Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

### Decision trees

A decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer

then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. For example, We use the following decision tree to determine whether or not to play tennis:

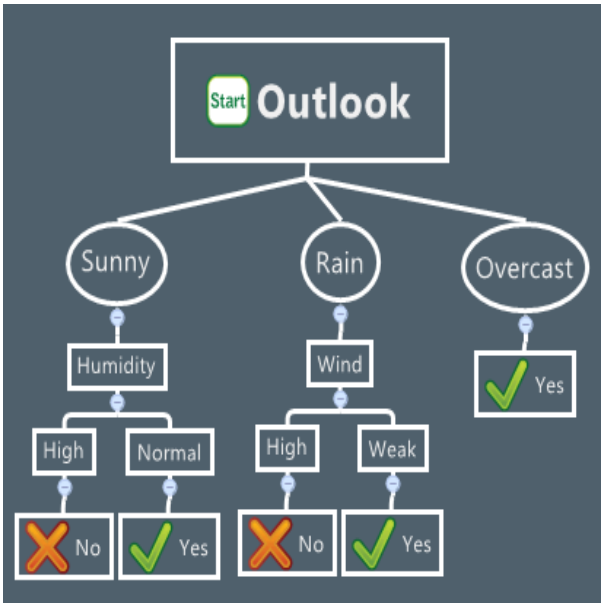


**Fig. 4: NLTK Tool**

Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is the week. And if it is sunny then we should play tennis in case the humidity is normal. We often combine two or more of those data mining techniques together to form an appropriate process that meets the business needs.

## IV. Data Mining in Banking Sector

By using data mining to analyze patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers will be likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable. The banking industry is widely recognizing the importance of the information it has about t its customers. Undoubtedly, it has among the richest and largest pool of customer information, covering customer demographics, transactional data, credit cards usage pattern, and so on. As banking is in the service industry, the task of maintaining a strong and effective CRM is a critical issue.

To do this, banks need to invest their resources to better understand their existing and prospective customers. By using suitable data mining tools, banks can subsequently offer „tailor-made“ products and services to those customers. There are numerous areas in which data mining can be used in the banking industry, which include customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations, optimizing stock portfolios, and ranking investments. In addition, banks may use data mining to identify their most profitable credit card customers or high-risk loan applicants. There is, therefore, a need to build an analytical capability to address the above-stated issues and data mining attempts to provide the answer. Following are some examples of how the banking industry has been effectively utilizing data mining in these areas.

### Marketing

Bank analysts can also analyze the past trends, determine the present demand and forecast the customer behavior of various products and services in order to grab more business opportunities and anticipate behavior patterns. Data mining technique also helps to identify profitable customers from non-profitable ones. ‘Cross-selling’ is another marketing area where data mining can be extensively used. Here, a service provider makes it attractive for a customer to buy additional products or services with the same business. The more products and services a bank can provide for customers, the more likely the bank is to retain those customers.

**Risk Management**

Data mining is widely used for risk management in the banking industry. Bank executives need to know whether the customers they are dealing with are reliable or not. Offering new customers credit cards, extending existing customers lines of credit, and approving loans can be risky decisions for banks if they do not know anything about their customers. Data mining, however, can be used to reduce the risk of banks that issue credit cards by determining those customers who are likely to default on their accounts. An example was reported in the press of a bank discovering that cardholders who withdrew money at casinos had higher rates of delinquency and bankruptcy. It is a common practice on the part of banks to analyze customers‟ transaction behaviors in their deposit accounts to determine their probability of default in their loan accounts. Credit scoring, in fact, was one of the earliest financial risk management tools developed. Credit scoring can be valuable to lenders in the banking industry when making lending decisions. Lenders would not have expanded the number of loans they give out without having an accurate, objective, and controllable risk assessment tool. Data mining can also derive the credit behavior of individual borrowers with installment, mortgage and credit card loans, using characteristics such as credit history, length of employment and length of residency. A score is thus produced that allows a lender to evaluate the customer and decide whether the person is a good candidate for a loan, or if there is a high risk of default.

**Fraud Detection**

Another popular area where data mining can be used in the banking industry is in fraud detection. Being able to detect fraudulent actions is an increasing concern for many businesses; and with the help of data mining more fraudulent actions are being detected and reported. Two different approaches have been developed by financial institutions to detect fraud patterns. In the first approach, a bank taps the data warehouse of a third party (potentially containing transaction information from many companies) and uses data mining programs to identify fraud patterns. The bank can then cross-reference those patterns with its own database for signs of internal trouble. In the second approach, fraud pattern identification is based strictly on the bank‟s own internal information. Most of the banks are using a hybrid approach. One system that has been successful in detecting fraud is Falcon's fraud assessment system. It is used by nine of the top ten credit card issuing banks, where it examines the transactions of 80 per cent of cards held in the US. Mellon Bank also uses data mining for fraud detection and is able to better protect itself and its customers funds from potential credit card fraud.

**Customer Acquisition and Retention**

Not only can data mining help the banking industry to gain new customers, it can also help retain existing customers. Customer acquisition and retention are very important concerns for any industry, especially the banking industry. Today, customers have so many opinions with regard to where they can choose to do their business. Executives in the banking industry, therefore, must be aware that if they are not giving each customer their full attention, the customer can simply find another bank that will. Data mining can also help in targeting "new" customers for products and services and in discovering a customer‟s previous purchasing patterns so that the bank will be able to retain existing customers by offering incentives that are individually tailored to each customer‟s needs.

**Conclusion**

Data Mining techniques can be of immense help to the banks and financial institutions in this arena for better targeting and acquiring new customers, fraud detection in real time, providing segment based products for better targeting the customers, analysis of the customers' purchase patterns over time for better retention and relationship, detection of emerging trends to take proactive approach in a highly competitive market adding a lot more value to existing products and services and launching of new product and service bundles. Data mining is more than running some complex queries on the data you stored in your database. You must work with your data, reformat it, or restructure it, regardless of whether you are using SQL, document-based databases such as Hadoop, or simple flat files. Identifying the format of the information that you need is based upon the technique and the analysis that you want to do. After you have the information in the format you need, you can apply the different techniques (individually or together) regardless of the required underlying data structure or data set. A majority of the banks in developing countries (particularly in the public sector) are not usually known to exploit their information asset for deriving business value through data mining and gain competitive advantage. But with progressive liberalization of rules on entry for private and foreign multinational banks, under the GATS framework of WTO, competitive pressure on domestic banks is increasing. Thus, customer retention and acquisition will be an important determinant of the banks bottom lines.

## REFERENCES


[1] Bhambri, V., 2011. Application of data mining in banking sector. IJCST, 2: 199-202.

[2] Bhattacharya, S., S. Jha, K. Tharakunnel and J.C. Westland, 2011. Data mining for credit card fraud: A comparative study. Decision Support Syst., 50: 602- 613. DOI: 10.1016/j.dss.2010.08.008.

[3] Babcock C. (1994) Parallel Processing Mines Retail Data. Computer World.

[4] Berry M.J.A. & Linoff G. (1999) Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley and Sons, Inc.

[5] Davenport T.H. & Prusak L. (2000). Working Knowledge: How organizations manage what they know. Boston, Massachusetts; Harvard Business School Press.

[6] B. Desai and Anita Desai, "The Role of Data mining in Banking Sector", IBA Bulletin, 2004.

[7] S.S.Kaptan, “New Concepts in Banking”, Sarup and Sons, Edition, 2002

[8] S. S. Kaptan, N S Chobey, "Indian Banking in Electronic Era", Sarup and Sons, Edition 2002.

[9] Rajanish Dass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad.

[10] Moradi, M., M. Salehi, M.E. Ghorgani and H.S. Yazdi, 2013. Financial distress prediction of Iranian companies by using data mining techniques. Organizacija, 46: 20-27.

[11] Petry, F.E. and L. Zhao, 2009. Data mining by attribute generalization with fuzzy hierarchies in fuzzy databases. Fuzzy Sets Syst., 160: 2206-2223. DOI:10.1016/j.fss.2009.02.014.]

[12] Shinde, P., 2012. Data mining using artificial neural network tree. IOSR J. Eng. Tremblay, M.C., K. Dutta and D. Vandermeer, 2010.

[13] Using data mining techniques to discover bias patterns in missing data. J. Data Inform. Q., 2: 1-19.