Global Journals LaTeX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Application Areas of Data Mining in Indian Retail Banking Sector

Sudhakar M[1]

[1] SK University

Received: 8 December 2013 Accepted: 5 January 2014 Published: 15 January 2014

Abstract

Banking systems collect huge amounts of data on day to day basis, be it customer information, transaction details, risk profiles, credit card details, credit limit and collateral details, compliance and Anti Money Laundering (AML) related information, trade finance data, SWIFT and telex messages. Thousands of decisions are taken in a bank daily. These decisions include credit decisions, default decisions, relationship start up, investment decisions, AML and Illegal financing related. One needs to depend on various reports and drill down tools provided by the banking systems to arrive at these critical decisions. But this is a manual process and is error prone and time consuming due to large volume of transactional and historical data. Interesting patterns and knowledge can be mined from this huge volume of data that in turn can be used for this decision making process. This article explores and reviews various data mining techniques that can be applied in banking areas. It provides an overview of data mining techniques and procedures. It also provides an insight into how these techniques can be used in banking areas to make the decision making process easier and productive.

Index terms— data mining, banking, unstructured data, default detection, customer classification, AML.

1 Introduction

anking industry has hugely benefited from the advancements in digital technology (Sing and Tigga,2010). Concept of data stored at branches has given way to centralized databases. Number of channels to access bank accounts has multiplied. Banking systems have become technically strong and customer oriented with online transactions, electronic wire transfers, ATM and cash and cheque deposit machines (Bhambri, 2011). As number of channels has increased so is the number of transactions and the related data stored. So currently banks have huge electronic data repositories in their computing storage systems. Data has grown in terms of both dimensionality and size (Kaur and Sing, 2011). With advancements in data mining techniques and know how, this mountain of data is turning out to be the most valuable asset of the organization (Tiwari, 2010). Valuable knowledge and interesting patterns are hidden in this data. There are huge potential for banks to apply data mining in their decision making processes in areas like marketing, credit risk management, detection of money laundering, liquidity management, investment banking and detection of fraud transactions in time Failures in these areas can lead to unpleasant outcomes for the bank such as losing customers to competition, financial loss, reputational loss and hefty fines from the regulators.

Figure 1 shows decision making in conventional settings. They are mostly done by manual procedures. Users go through reports generated by banking information system and use it in their decision making process. They may also use drill down tools provided by the system for analyzing data to arrive at critical decisions. Manual analysis has limitations because volumes of data that can be manually analyzed are limited and hence the decisions may not be as accurate as intended ??Bhasin, 2006). For example, it could be possible that loan installments are being paid regularly though there is an alarming negative trend in the customers turnover and the account may be about to default. These associations are not easy to detect through manual processes. It is assumed that valuable information are hidden in this volume of operational and historic data that can be used for critical decision making process if they are discovered and put to use by capable tools (Kazi and Ahmed, 2012). For example, a decision support system based on data mining techniques can be employed to improve the quality of lending process in a bank (Ionita and Ionita, 2011). Figure 2 shows how data mining can improve decision making process.

2 II.

3 Data Mining and Knowledge Discovery Concepts

Data Mining and Knowledge Discovery is one of recent developments in line with data management technologies. It combines the fields of statistics, machine learning, database management, information science and visualization. It is an emerging field. Despite this, it is increasingly being used in the industry as a tool to study their customers and make smart decisions (Ramageri, 2010). Knowledge discovery from databases is defined as the process of identifying valid, novel, potentially useful and ultimately understandable patterns of data. One of the crucial steps in Knowledge discovery is Data Mining and often they are used as synonyms (Deshpan de and Thakare, 2010). Data Mining is the process of discovering valuable information from large data stores to answer critical business questions. It unveils implicit relationships, trends, patterns, exceptions and anomalies that were hidden to human analysis. In today's highly competitive market environment customers are spoilt by choices. Banks need to be proactive in analyzing customer preferences and profiles and tune their products and services accordingly to retain customer base (Bhambri, 2011). By segmenting customers into bad customers and good customers, bank can cut losses before it is too late. By analyzing patterns of transactions, bank can track fraud transactions before it affects its profitability where data mining could help. Data mining is the process of deriving knowledge hidden from large volumes of raw data. The knowledge must be new, not obvious, must be relevant and can be applied in the domain where this knowledge has been obtained. The logical process flow involved in data mining and knowledge discovery is shown in Figure ??. Data mining process can be broken down to the following iterative sequence of following steps. Data required for the analysis are identified and brought from the data source. This is the first step in data mining process. Data source can be from operational or historical database or from a data warehouse.

4 b) Data Preprocessing

It involves Data Cleaning and Data Integration.

5 c) Data Cleaning

This is the stage where noise, irrelevant and inconsistent data are removed from the data selected.

6 d) Data Integration

In a production environment, there could be multiple databases storing same information. These heterogeneous data sources are combined in a common source.

7 e) Data Transformation and Data Reduction

Data are transformed or consolidated by performing summary or aggregation operations so that they are simpler to handle for the mining operations. Redundant or highly correlated data items can be dropped out so that data mining results would be more effective.

8 f) Data Mining

In this crucial step, intelligent data mining techniques are applied in order to extract data patterns. There could be many potentially useful patterns depending on the techniques used which need to be further analyzed for identifying the crucial ones.

9 g) Pattern Evaluation

In this stage, the patterns identified in the previous steps are evaluated for their relevance and usefulness in the applied domain. There are standard measures to find out if a pattern is interesting.

10 h) Knowledge Presentation

Here visualization and knowledge representation techniques are used to present mined knowledge to the user.

11 III.

12 Data Mining Techniques

Techniques applied for mining knowledge can be divided into various classes depending on the nature of knowledge that system is unearthing. We will now look into these important techniques.

13 a) Association

This technique is used to unearth unsuspected data dependencies. In other words, it tries to detect data items that are associated or connected or correlated with each other which are not obvious previously. For example, if customers who are enquiring about a banking product, more often enquire about another unrelated product, then this technique can find this pattern out and inform the marketing team. More formally, the task is to uncover hidden associations from a large database. The idea is to derive a set of strong association rules in the form of "A1?A2? ? Am? B1? B2? ? Bn" where Aj (for i?{1?m}) and Bj (for j?{1? n}) are set of attribute-values from the relevant data sets in a database. For example, data recorded by a point of sales system would indicate that if customers buy certain items, they are most likely to buy certain other items. Such information can be used as decisions for marketing activities promotional pricing or product placements (Tiwari, 2010). In addition to this, association rules are employed in application areas including web usage mining, intrusion detection and bioinformatics. Typically all association rules are not interesting. From a large data set, a very large and a high proportion of the rules mined will be usually of little value. An associative relationship is considered to be useful if it satisfies a predefined support and confidence values (Geng and Hamilton, 2006). Hence, a rule is discarded if it does not satisfy this minimum support threshold and minimum confidence threshold. All these discovered strong association rules may not be interesting enough to present. Additional analysis need to be performed to uncover interesting statistical correlations between associated attribute-value pairs (Geng and Hamilton, 2006). Various types of association include (Ramageri, 2010):? Multilevel association rule ? Multidimensional association rule ? Quantitative association rule ? Direct association rule ? Indirect association rule b)

14 Classification and Prediction

This is the most commonly applied data mining technique. It is employed when the classes of data in the population are known. For example, in the case of detecting fraudulent banking transactions from a bank's transactions database, there can only be two classes, namely fraudulent and non-fraudulent. It constructs a model from the sample data items with known class labels and use this model to predict the class of objects in the population whose classes are not known. Each tuple from the database contains one or more predicting attributes which determines the predicted class label of the tuple according to the constructed model. In the banking scene, classification technique is employed for Fraud detection (both corporate and credit fraud) These models are constructed usually using a decision tree model or a neural network model. A decision tree is a flow chart like recursive structure to express classification rules where each node specifies a test on an attribute value, each branch specifies a mutually exclusive outcome of the test together with a subsidiary decision tree for each outcome and tree leaves represent classes or class distributions. It can easily be converted to classification rules or can be used to compact description of data (Asghar and Iqbal, 2009). Fuzzy sets are applied to the classification techniques when parameters to consider are of fuzzy in nature. For example, the length of URL parameter for detecting phishing sites can range from low to highwith other values in between (Aburrous et al., 2010). Other commonly used classification technique involves application of neural networks. A neural network is essentially a network of processing nodes with weighted connections between the nodes where the weights are determined by a learning process using training data. Neural networks are computationally more expensive than their decision tree counterparts (Kumar et al. 2011).

Classification works with discrete and unordered data and helps to identify class labels of the members of the population. But prediction models works with continuous-valued functions. That is, it is used to predict missing or unavailable numerical data values from the sample attribute values. Commonly used technique for prediction is regression analysis. It is a statistical methodology that is used to forecast values from existing numerical values. In predictive models for data mining, we have a set of independent variables whose values are already known and a set of dependent or response variables whose values we want to predict. Regression helps us to express the relationship between these variables as a linear or non-linear function. In many real world problems related to banking such as stock price predictions, or credit scoring follow complex models with many independent variables and requires multidimensional regression analysis and logistic regression (Li and Liao, 2011).

15 c) Cluster Analysis and Concept Formation

Clustering is similar to classification. But subtle difference is that classes are not known before. Clustering is used to generate class labels. The objects are classified or grouped based on the principle of maximizing the similarity within a class based on the observed pattern. A regularly used and the simplest of clustering algorithms is K-means algorithm (Kaur and Kaur, 2013). Heuristics based on the domain information can be applied to cluster data where K-Means algorithm produces a large number of outliers (Shashidhar and Varadarajan, 2011). Self-Organizing Map is an important neural network based technique employed for clustering and has been applied for problems in banking domain (Shih, 2011). Concept formation is a closely related process to clustering and is used to learn summaries from data. This process integrates learning and classification tasks to identify summaries and organize learned summaries into a hierarchy. In banking area, clustering and concept formation can be employed for classifying customers with same kind of transactions or queries or profiles or subscribe to similar products or has similar risk aptitude. For example, in banking sector salaried customers tend to join investment plans with regular contributions. Knowledge about these classes will help banks to design products to each class of customers and can embark on targeted and more effective marketing campaigns.

16 IV. Application Areas of Data Mining in Banking

Banking information systems contains huge volumes of data both operational and historical. Data mining can assist critical decision making processes in a bank (Ionita and Ionita, 2011). Banks who apply data mining techniques in their decision making hugely benefit and hold an edge over others who don't. Some of these decisions are in the areas of marketing, risk management and default detection, fraud detection, customer relationship management and money (Khac and Kechadi, 2010; ??eepa and Dhanapal, 2009). These applications are described below.

17 a) Risk Management and Default Detection

Every lending decision a bank takes involve a certain amount of risk. Quantifying this risk can make the risk management process easier and limit the risk of financial loss to the bank. Knowing customers' capability to repay can greatly enhance a credit manager's decisions. Data mining can also help to identify which customer is going to delay or default a loan repayment (Kazi and Ahmed, 2012). This advanced knowledge can help the bank to take corrective measures to prevent losses. For such forecasting, parameters to consider are turnover trends,balance sheet figures, limit utilization, behavioral patterns and cheque return patterns. Historical default patterns can also help in predicting future defaults when same patterns are discovered (Costa et al., 2007). Data mining techniques are applied to enhance the accuracy of credit scores and predict default probabilities (Li and Liao, 2011). Credit score is a value representing a borrower's creditworthiness. Behavioral scores are obtained from probability models of customer behavior to forecast their future behaviors in various situations. Data mining can derive this score using the past behaviors of the borrower related to debt repayments by analyzing available credit history (He et al., 2010).

18 b) Marketing

Marketing is one of the mostly used application area for Data Mining by the industry in general (Zhang et al., 2008). Banking is not an exception. Retaining customers and finding new customers are getting increasingly difficult because of cut throat competition prevailing in the market these days. Only way to retain a customer or win a new customer is to be proactive and know beforehand what the customer expects and offer him what he wants. This is where data mining can help a great deal (Chopra et al., 2011). Data mining applied to customer relationship management systems can analyze customer data and can discover key indicators to help the bank to be equipped with the knowledge of factors that affected customer's demands in the past and their needs in the future (Ngai et al., 2009). This enables the bank to targeted marketing. Sequential patterns can be analyzed to investigate changing customer preferences and can approach customers pro-actively (Sundari and Thangadurai, 2010). Data mining techniques can help in classifying customers according to the customer's attributes, behavior, needs, preferences, value and other factors (Ren et al., 2010). Mainly two scoring models are used for this classification purposes, namely credit scoring model and behavioral scoring model. This classification is valuable information for making customer oriented marketing strategies tailor made for the target category and provide different services for each customer category (Ping and Liang, 2010). For example it can determine how customers will react to a change in interest rates, which customers will be likely to accept new product offers, what collateral would require from a specific customer segment for reducing loan losses. Different levels of analysis like RFM (Recency, Frequency and Monitory) analysis, LTV (Life Time Value) of customers coupled with K-Means clustering can be employed to develop an effective customer segmentation thereby increasing targeted marketing ??Varun et al., 2012). Data mining can also reveal possibility of cross selling such as selling home loans to credit card customers by analyzing associations from the past data (Qiu et al., 2009). It can also develop a model of existing home loan customers to analyze their profiles to explore similar customers in other portfolios (like demand deposits or customers with Sreekumar Pulakkazhy and R.V.S. Balan / Journal of Computer Science 9 (10): 1252-1259, 2013insurance products) to find out potential customers for home loans (Shinde et al., 2012).

19 c) Fraud Detection

Banks lose millions of dollars annually to various frauds. Detecting fraudulent transactions can help the banks to act early and limit the damages. Fraud detection is the process of identifying fraudulent transactions from genuine transactions or in other words this process segregates a list of transactions into two classes namely fraudulent and legitimate ??Ogwueleka, 2011). Most important area where fraud detection can help is the credit card products. Clustering methods can be used to classify transactions and outliers can be analyzed for frauds (Dheepa and Dhanapal, 2009). Probability density of credit card user's past behavior can be modeled and the probability of current behavior can be calculated to detect frauds (Dheepa and Dhanapal, 2009). Patterns of customer's transactions can be discovered and alerts can be generated if any measurable deviations are found. Financial statement fraud detection is another area that can employ data mining principles to effective use.

Banks make credit decisions based on financial statements produced by customers. These statements may contain overstated assets, sales and profits or it may understate losses and liabilities. Even though these statements may have been audited, these kinds of frauds are hard to detect using normal auditing procedures. Classification techniques based on neural network, regression and decision tree are used for classifying fraudulent ratios in the statements from the nonfraudulent data (Sharma and Panigrahi, 2012).

20 d) Money Laundering Detection

Money Laundering is the process of hiding the illegal origin of "black" money so as to legitimize it (Khac and Kechadi, 2010). Banks are commonly used as Year 2014 c channels to launder money. Therefore governments and financial regulators require banks to implement processes, systems and procedures to detect and prevent money laundering transactions. Failure to detect and prevent such illegal transactions can invite hefty fines both monetarily and operationally which can prove very costly for the bank and even can make its survival difficult. Conventional rule-based transaction analysis based on reports and tools will not be sufficient to detect more complicated transaction patterns like smurfing and networked transactions (Khac et al., 2011). Here data miningtechniques can be applied to dig out transaction patterns that can lead to money laundering. Typically such systems take client risk assessment data, transaction risk measurement data and patterns and behavior patterns into consideration for detecting money laundering patterns. Transactions are then grouped into clusters based on their similarities found in these chosen attributes (Khac et al., 2011). In a large database of banking transactions, it is possible that a huge number of patterns emerge and will be classified as money laundering transactions thereby increasing false positives. Statistical false reduction methods based on decision tree classification are employed to limit the number of false patterns detected (Anuar et al., 2008).

21 e) Investment Banking

Investment is an action of investing money into an asset or item for profit/income. Banks often offer investment services to their customers. There are a vast number of financial instruments in the market.

Data mining like K-means clustering can be applied to choose the best investments based on customer's profile (Ingle and Meshram, 2012). Capability to predict asset prices (for example stock prices) from historic prices can increase returns from investment tremendously. Data mining techniques for prediction like neural networks and linear regression can be employed for prediction of prices for stocks (Naeini et al., 2010). Data mining can also be applied in time series analysis for financial applications (Tak-chung, 2011).

V.

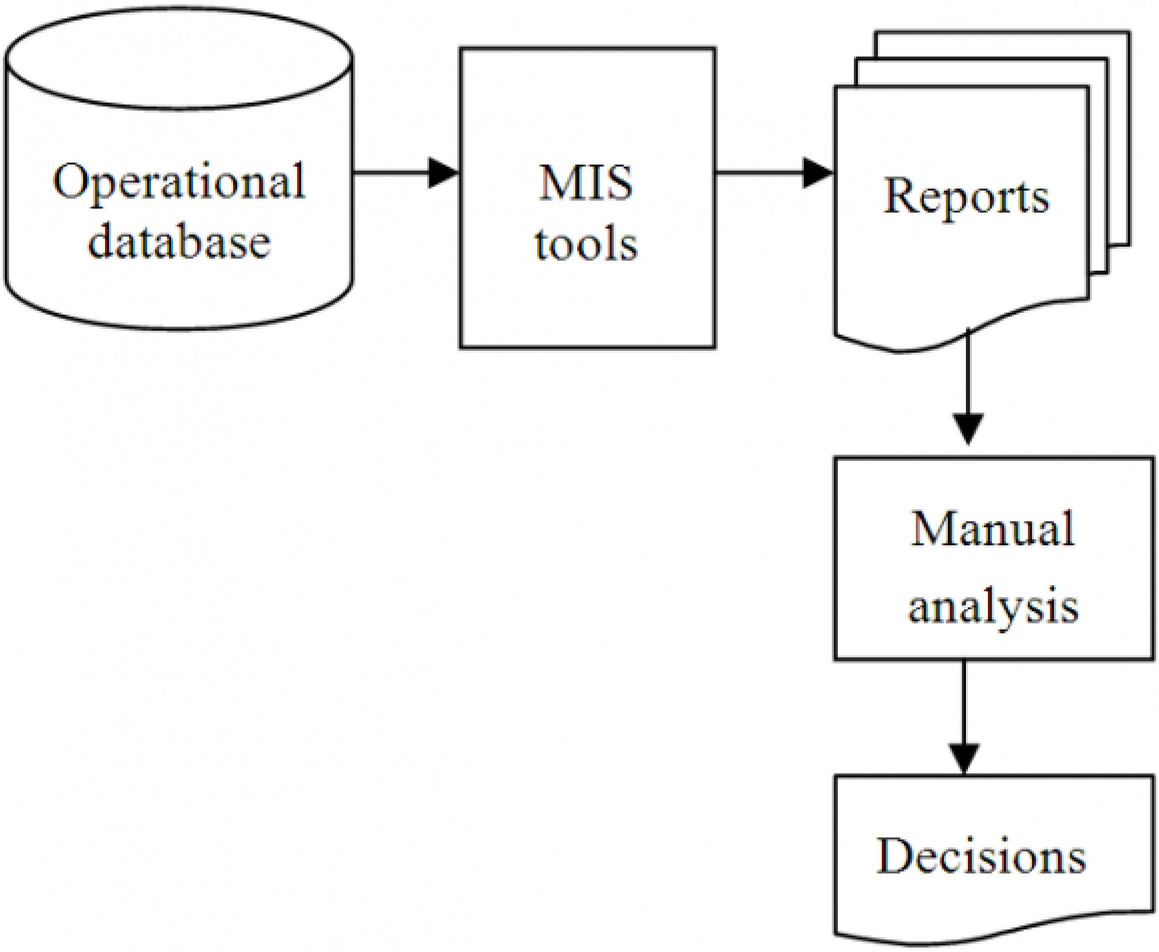
22 Conclusion

Data mining is a process to extract knowledge from existing data. It is used as a tool in banking and finance in general to discover useful information from the operational and historical data to enable better decision-making. It is an interdisciplinary field, confluence of Statistics, Database technology, Information science, Machine learning and Visualization. It involves steps that include data selection, data integration, data transformation, data mining, pattern evaluation, knowledge presentation. Banks use data mining in various application areas like marketing, fraud detection, risk management, money laundering detection and investment banking. The patterns detected help the bank to forecast future events that can help in its decision-making processes. More and more banks are investing in data mining technologies to be more competitive. [1]

[1] © 2014 Global Journals Inc. (US)

1

Figure 1: Figure 1 :



2

Figure 2: Figure 2 :

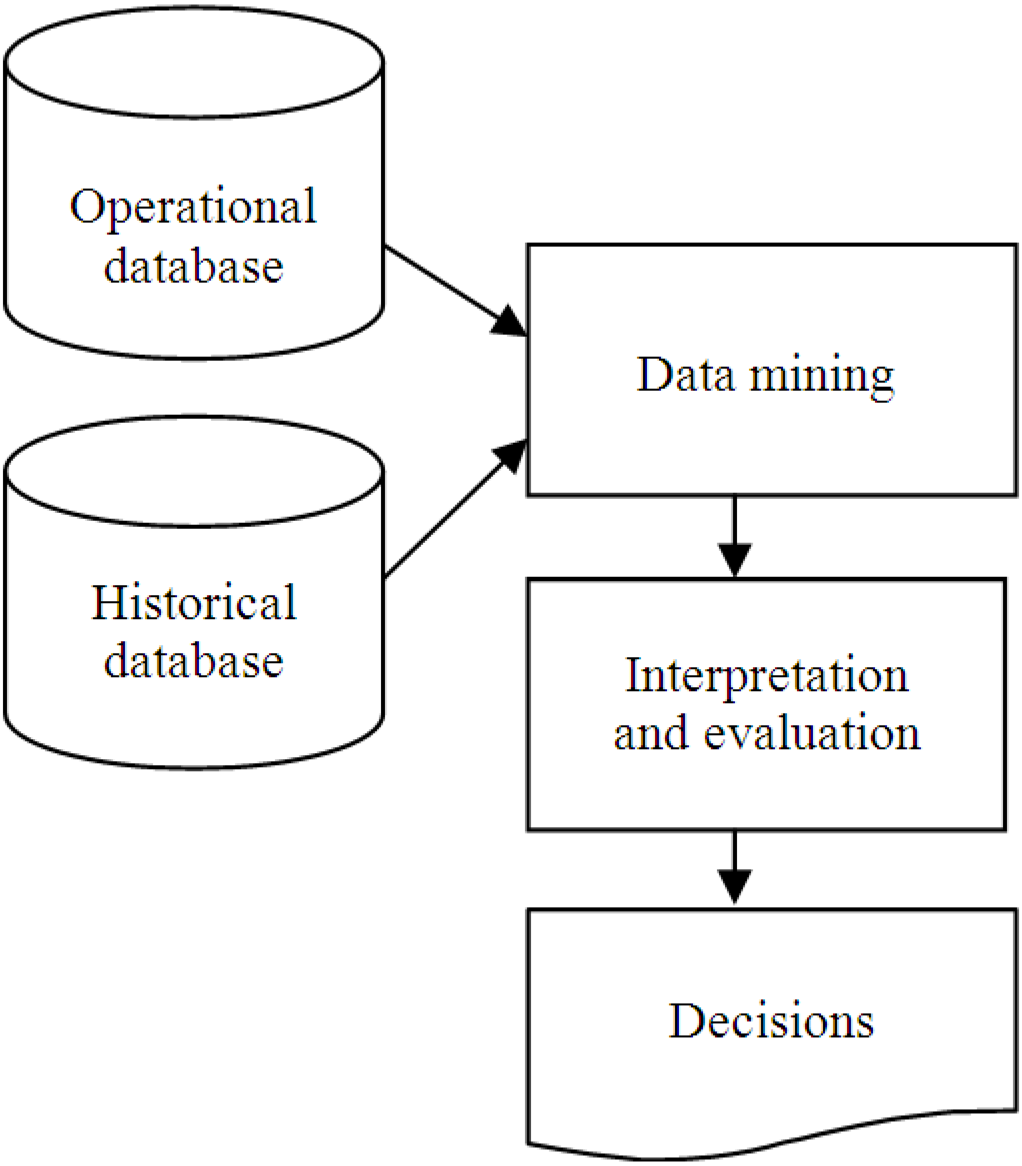


Figure 3: Application

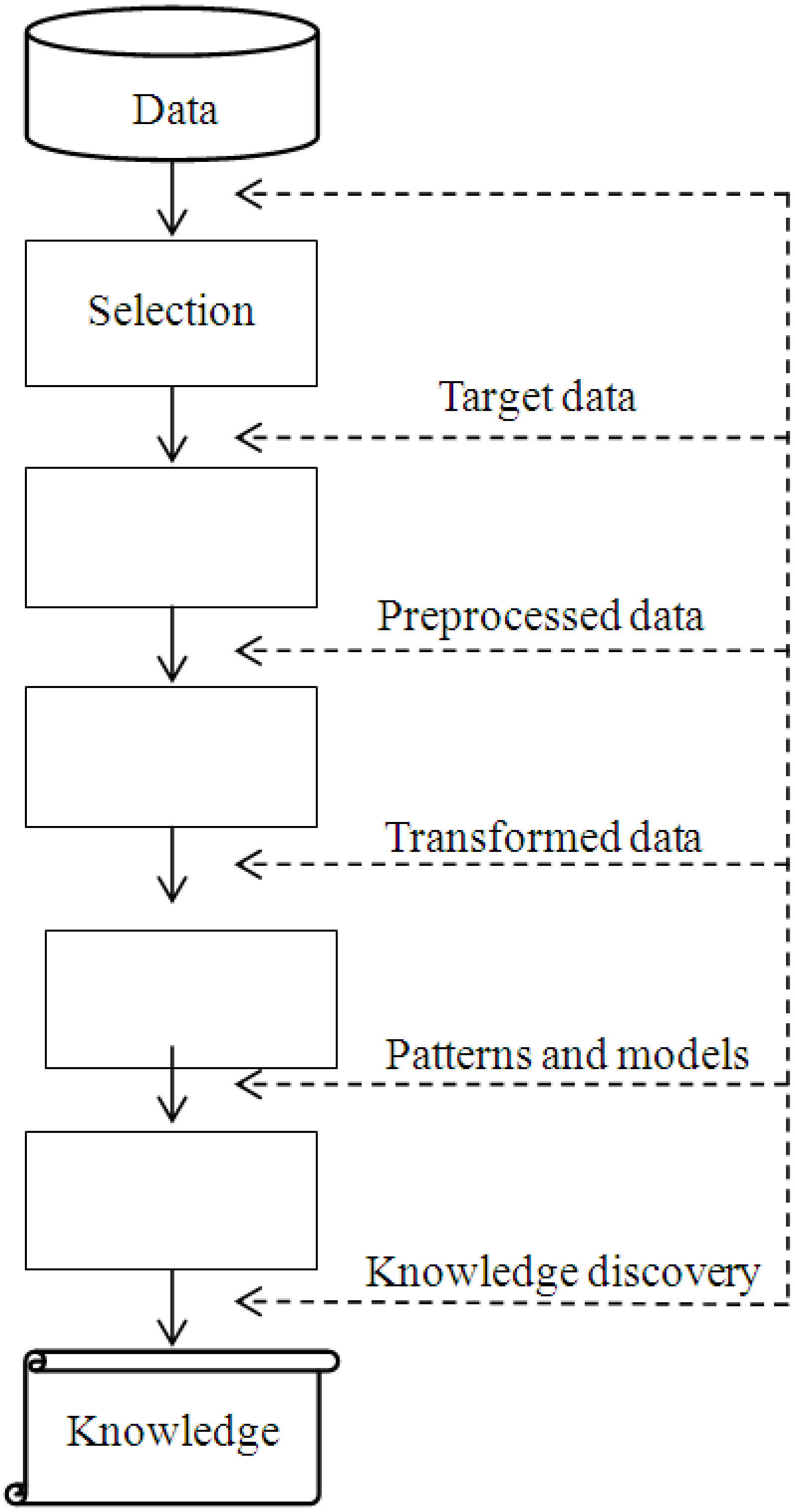


Figure 4: Application

This page is intentionally left blank

[Pulakkazhy and Balan ()] , Sreekumar Pulakkazhy , R V S Balan . *Journal of Computer Science* 2013. 9 (10) p. .

[Qiu et al. (2009)] 'A model for a bank to identify cross-selling opportunities'. D H Qiu , Y Wang , Q F Zhang . 10.1109/CISE.2009.5362870. *Proccedigns of the International Conference on Computational Intelligence and Software Engineering*, (cedigns of the International Conference on Computational Intelligence and Software EngineeringWuhan) 2009. Dec. 11-13. IEEE Xplore Press. p. .

[Sharma and Panigrahi ()] 'A review of financial accounting fraud detection based on data mining techniques'. A Sharma , P K Panigrahi . 10.5120/4787-7016. *Int. J. Comput. Appli* 2012. 39 p. .

[Kaur and Kaur ()] 'A survey on various clustering techniques with k-means clustering algorithm in detail'. S Kaur , U Kaur . *Int. J. Comput. Sci. Mob. Comput* 2013. 2 p. .

[Li and Liao (2004)] 'An empirical study on creditscoring model for credit card by using data mining technology'. W Li , J Liao . 10.1109/CIS.2011.283. *Proceedigngs of the 7th International Conference on Computational Intelligence and Security*, (eedigngs of the 7th International Conference on Computational Intelligence and SecurityHainan) 2011. Dec. 3-4. IEEE Xplor Press. p. .

[Khac et al. ()] 'An investigation into data mining approaches for anti money laundering'. N A L S Khac , M Markos , A O'neill , M Brabazon , Kechadi . *Proceedings of the International Conference on Computer Engineering Applications, (EA' 11)*, (the International Conference on Computer Engineering Applications, (EA' 11)Singapor) 2011. Lacsit Press. p. .

[Dheepa ()] 'Analysis of credit card fraud detection methods'. V Dheepa , R . *Int. J. Recent Trends Eng* 2009. 2 p. .

[Khac and Kechadi (2010)] 'Application of data mining for anti-money laundering detection: A case study'. N A L Khac , M Kechadi . 10.1109/ICDMW.2010.66. *Proccedigs of the International Conference on Data Mining Workshop*, (cedigs of the International Conference on Data Mining WorkshopSydney, NSW) 2010. Dec. 13-13. IEEE Xplore Press. p. .

[Bhambri ()] 'Application of data mining in banking sector'. V Bhambri . *Internat. J. Comput. Sci. Technol* 2011. 2 p. .

[Ngai et al. ()] 'Application of data mining techniques in customer relationship management: A literature review and classification'. E W T Ngai , L Xiu , D C K Chau . 10.1016/j.eswa.2008.02.021. *J. Eng. Sci. Technol* Ogwueleka, F.N. (ed.) 2009. 2011. 36 p. . (Expert Syst. Appli.)

[Asghar (2005)] 'Automated data mining techniques: A critical literature review'. S Asghar , K . 10.1109/ICIME.2009.98. *IEEE Proccedings of the International Conference on Information Management and Engineering*, (Kuala Lumpur) 2009. Apr. 3-5. IEEE Xplore Press. p. .

[Ren et al. (2010)] 'Customer segmentation of bank based on data warehouse and data mining'. S Ren , Q Sun , Y Shi . 10.1109/ICIME.2010.5477693. *Proceedings of the 2nd IEEE International Conference on Information Management and Enginerring*, (the 2nd IEEE International Conference on Information Management and EnginerringChengdu) 2010. Apr. 16-18. IEEEXplore Press. p. .

[Ping and Liang (2010)] 'Data mining application in banking-customer relationship management'. Z L Ping , S Q Liang . 10.1109/ICCASM.2010.5619002. *Proccedigns of the International Conference on Computer Application and System Modeling*, (cedigns of the International Conference on Computer Application and System ModelingTaiyuan) 2010. Oct. 22-24. IEEE Xplore Press. p. .

[Technol et al. ()] 'Data mining for effective risk analysis in a bank intelligence scenario'. Technol , Appli , G Costa , F Folino , A Locane , G Manco , R Ortale . 10.1109/ICDEW.2007.4401083. *Preccedings of the 23 rd International Conference on Data Engineering Workshop*, (Istanbul) 2007. Apr. 17-20. IEEE Xplore Press. 2 p. .

[Deshpande and Thakare ()] 'Data mining system and applications: A review'. M S P Deshpande , D V M Thakare . *Int. J. Distrib. Parallel Syst* 2010. 1 p. .

[Ramageri ()] 'Data mining techniques and applications'. B M Ramageri . *Ind. J. Comput. Sci. Eng* 2010. 1 p. .

[Kaur and Sing ()] 'Data mining: An overview'. G Kaur , L Sing . *Int. J. Comput. Sci. Technol* 2011. 2 p. .

[He et al. ()] 'Domaindriven classification based on multiple criteria and multiple constraint-level programming for intelligent credit scoring'. J He , Y Zhang , Y Shi , G Huang . 10.1109/TKDE.2010.43. *IEEE Trans. Knowl. Data Eng* 2010. 22 p. .

[Ingle and Meshram ()] *E-Investment Ionita, I. and L. Ionita, 2011. A decision support based on data mining in e-banking*, D R Ingle , B B Meshram . 2012. (IEEE Preccedings of the 10th Reodunet International)

[Anuar et al. ()] 'Identifying false alarm for network intrusion detection system using hybrid data mining and decision tree'. N B Anuar , H Sallehudin , A Gani , O Zakari . *Malyasian J. Comput. Sci* 2008. 21 p. 110115.

[Chopra et al. ()] 'Implementation of data mining techniques for strategic CRM issues'. B Chopra , V Bhambri , B Krishnan . *Int. J. Comput* 2011.

[Aburrous et al. ()] 'Intelligent phishing detection system for e-banking using fuzzy data mining'. M Aburrous , M A Hossain , K Dahal , F Thabtah . DOI:0.1016/j.eswa.2010.04.044 *Expert Syst. Appli* 2010. 37 p. .

[Geng and Hamilton ()] 'Interestingness measures for data mining: A survey'. L Geng , J H Hamilton . 10.1145/1132960.1132963. *ACM Comput. Surveys* 2006. 38 p. .

[Kumar et al. (2011)] 'Performance evaluation of decision tree versus artificial neural network based classifiers in diversity of datasets'. P Kumar , S V Nitin , D S Chauhan . 10.1109/WICT.2011.6141349. *Proccedigs of the World Congress on Information and Communication Technologies*, (cedigs of the World Congress on Information and Communication TechnologiesMumbai) 2011. Dec. 11-14. IEEE Xplore Press. p. .

[Naeini et al. (2010)] 'Stock market value prediction using neural networks'. M P Naeini , H Taremian , H B Hashemi . 10.1109/CISIM.2010.5643675. *Proccedings of the International Conference on Computer Information Systems and Industrial Management Applications*, (cedings of the International Conference on Computer Information Systems and Industrial Management ApplicationsKrackow) 2010. Oct. 8-10. IEEE Xplore Press. p. .

[Ngai et al. ()] 'The application of data mining techniques in financial fraud detection: Aclassification framework and an academic review of literature'. E W T Ngai , H Yong , Y H Wong , C Yijun , S Xin . 10.1016/j.dss.2010.08.006. *Decision Support Syst* 2011. 50 p. .

[Kazi and Ahmed ()] 'Use of data mining in banking'. I M Kazi , . Q B Ahmed . *Int. J. Eng. Res. Appli* 2012. 2 p. .