Credit Marking in Banking

Tamanna
Research Scholar, Dept. Of Computer Science, Sat Priya Group of Institutions, Rohtak, Haryana, India.

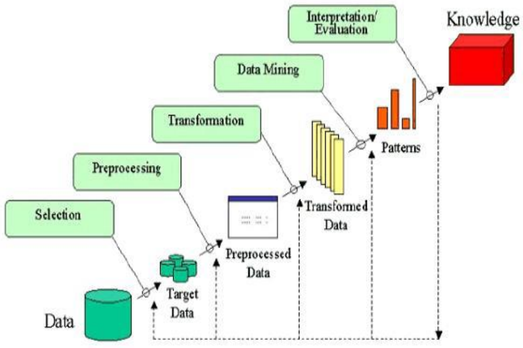
Dr. Harish Mittal
Director, Sat Priya Group of Institutions, Rohtak, Haryana, India.

Abstract – Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides. Data mining is used in various fields like, in credit card fraud detection, education, healthcare etc. In this review paper, a review of the fundamental terminology and the concepts of data mining are done.

Index Terms – Data Mining, Knowledge Discovery, Fraud detection.

1. INTRODUCTION

Data mining is a process that takes data as input and outputs knowledge. One of the earliest and most cited definitions of the data mining process, which highlight of its distinctive characteristics, is provided by Fayyad, Piatetsky-Shapiro and Smyth (1996), who it as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Note that because the process must be non-trivial, simple computations and statistical measures are not considered data mining. Thus predicting which salesperson will make the most future sales by calculating who made the most sales in the previous year would be considered data mining. The connection between "patterns in data" and "knowledge" will be discussed shortly. Although not stated explicitly in this definition, it is understood that the process must be at least partially automated, relying heavily on specialized computer algorithms (i.e., data mining algorithms) that search for patterns in the data. It is important to point out that there is some ambiguity about the term "data mining", which is in large part purposeful. This term originally referred to the algorithmic step in the data mining process, which initially was known as the Knowledge Discovery in Databases (KDD) process. Data mining is one of the most important steps of KDD process and it is the process of extracting hidden information from large database and transforms it into understandable format by considering different Perspectives.



Data Selection

This stage includes the study of the application domain, and the selection of the data. The domain's study intends to contextualize the project in the company's operations, by understanding the business language and defining the goals of the project. In this stage, it is necessary to evaluate the minimum subset of data to be selected, the relevant attributes and the appropriate period of time to consider.

Data Pre-processing

This stage includes basic operations, such as: removing noise or outliers, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data attributes, and accounting for time sequence information and known changes. This stage also includes issues regarding the database management system, such as data types, schema, and mapping of missing and unknown values.

Data Transformation

This stage consists of processing the data, in order to convert the data in the appropriate formats for applying data mining algorithms. The most common transformations are: data normalization, data aggregation and data discretization. To normalize the data, each value is subtracted the mean and divided by the standard deviation. Some algorithms only deal with quantitative or qualitative data. Therefore, it may be necessary to discredit the data, i.e. map qualitative data to quantitative data, or map quantitative data to qualitative data. In metadata and data warehouse, a data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system.

Data Mining

This stage consists of discovering patterns in a dataset previously prepared. Several algorithms are evaluated in order to identify the most appropriate for a specific task. The selected one is then applied to the pertinent data, in order to find indirect relationships or other interesting patterns.

Interpretation/Evaluation

This stage consists of interpreting the discovered patterns and evaluating their utility and importance with respect to the application domain. In this stage it can be concluded that some relevant attributes were ignored in the analysis, thus suggesting the need to replicate the process with an updated set of attributes.

2. ARCHITECTURE OF DATA MINING

The architecture of a typical data mining system may have the following major components:

- Database, Data Ware House, World Wide Web, or Other Information Repository.
- Database or Data Ware House Server.
- Knowledge base
- Data mining engine
- Pattern evaluation module
- User interface

Database, Data Ware House, World Wide Web, or Other Information Repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

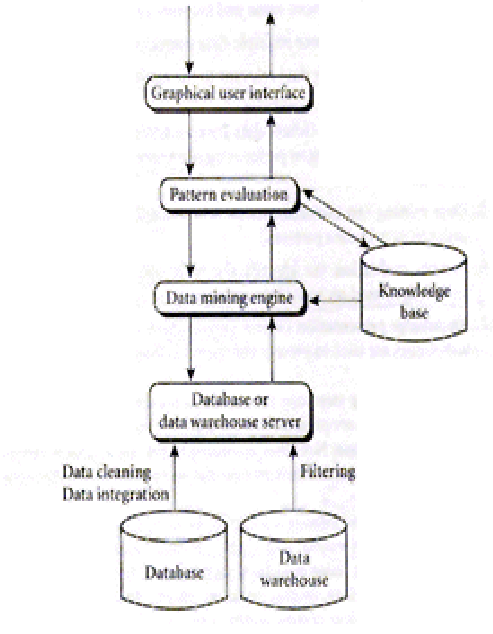
Database or Data Ware House Server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base: This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern evaluation module: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns [2] Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

User interface: This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.



2.1. Overview of Data Mining Tasks

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as[1,2]

Anomaly detection (Outlier/change/deviation detection)**:** The identification of unusual data records, that might be interesting or data errors that require further investigation.

Clustering: It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification: It is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression: It attempts to find a function which models the data with the least error.

Summarization: It provides a more compact representation of the data set, including visualization and report generation.

2.2. Data Mining Life Cycle

The life cycle of a data mining project consists of six phases [2,4]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase. The main phases are:

1. **Business Understanding:** This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. **Data Preparation:** It covers all activities to construct the final dataset from the initial raw data.
3. **Modeling:** In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
4. **Evaluation:** In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives.

At the end of this phase, a decision on the use of the data mining results should be reached.

5. **Deployment:** The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implement **User interface** presenting a repeatable data mining process across the enterprise.

3. DATA MINING TECHNIQUES

For achieving the data mining tasks listed in the previous section, a number of data mining techniques have been contributed by various disciplines.

Decision tree: A decision tree is a decision support tool that uses a tree-like graph of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Decision trees represent a supervised approach to classification. It is a structure which includes a root node, branches and leaf nodes. A decision tree is a simple tree structure where non-terminal/internal nodes represent tests on one or more attributes, each branch denotes the outcome of the test and each terminal node/leaf node holds a class label. The paths from root to leaf represent classification rules. The topmost node in the tree is root node. It is easy to comprehend and it does not require any domain knowledge. Moreover, the learning and the classification steps of a decision tree are simple and fast.

Clustering: It is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is commonly used for segmentation. Clustering algorithms allow entities described by a large number of attributes to be partitioned into a few distinct groups or segments.

Fuzzy Logic: Fuzzy logic is a form of multi valued logic in which the truth values of variables may be any real number between 0 and 1. By contrast, in Boolean logic, the truth values of variables may only be 0 or 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. The term "fuzzy logic" was introduced with the 1965 proposal of fuzzy set theory by Lotfi A. Zadeh. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions. In linguistic form, the imprecise concepts like “slightly”, “quite”, “very” are definable using fuzzy logic. It allows partial.

Nearest-Neighbor: Nearest-neighbor learners (Cover and Hart 1967) are very different from any of the learning methods just described in that no explicit model is ever built. That is, there is no training phase and instead all of the work associated with making the prediction is done at the time an example is presented. Given an example the nearest-neighbour method first determines the *k* most similar examples in the training data and then determines the prediction based on the class values associated with these *k* examples, where *k* is a user specified parameter. The simplest scheme is to predict the class value that occurs most frequently in the *k* examples, while more sophisticated schemes might use weighted voting, where those examples most similar to the example to be classified are more heavily weighted. People naturally use this type of technique in everyday life. For example, realtors typically base the sales price of a new home on the sales price of similar homes that were recently sold in the area. Nearest-neighbor learning is sometimes referred to as instance-based learning. Nearest-neighbor algorithms are typically used for classification tasks, although they can also be used for regression tasks.

4. DATA PREPROCESSING

It is an important step in data mining step in data mining process. The main period of data mining is to develop crude information. More than half of time, that is, in the middle of 60-90% of aggregate time is consumed in understanding and planning information. Subsequently suggests the noteworthiness of this stage. Under it numerous exercises are to be performed, for example, portrayal of information; disposal of commotion which is irregular piece of slip; joining of two or more information set to structure a solitary one; changing estimations of variable estimations of variable to obliged scale; decreasing the information by killing unimportant information. The phrase “garbage in, garbage out” is particularly applicable to data mining and machine learning projects. Data gathering methods are often loosely controlled, resulting in out of range value (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set. Kotsiantis et al. (2006) present a well-known algorithm for each step of data pre-processing.

5. DATA CLEANING

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data or coarse data. After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data. The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities.

6. DATA NORMALIZATION

Data normalization is the process of reducing data to its canonical form. For instance, Database normalization is the process of organizing the fields and tables of a relational database to minimize redundancy and dependency. In the field of software security, a common vulnerability is unchecked malicious input. The mitigation for this problem is proper input validation. Before input validation may be performed, the input must be normalized, i.e., eliminating encoding (for instance HTML encoding) and reducing the input data to a single common character set.

7. DATA TRANSFORMATION

In metadata and data warehouse, a data transformation converts a set of data values from the data format of a source data system into the data format of a destination data system. Data transformation can be divided into two steps:

- Data mapping maps data elements from the source data system to the destination data system and captures any transformation that must occur
- Code generation that creates the actual transformation program

Data element to data element mapping is frequently complicated by complex transformations that require one-to-many and many-to-one transformation rules. The code generation step takes the data element mapping specification and creates an executable program that can be run on a computer system. Code generation can also create transformation in easy-to-maintain computer languages such as Java or XSLT.A master data recast is another form of data transformation where the entire database of data values is transformed or recast without extracting the data from the database. All data in a well-designed database is directly or indirectly related to a limited set of master database tables by a network of foreign key constraints. Each foreign key constraint is dependent upon a unique database index from the parent database table. Therefore, when the proper master database table is recast with a different unique index, the directly and indirectly related data are also recast or restated. The directly and indirectly related data may also still be viewed in the original form since the original unique index still exists with the master data. Also, the database recast must be done in such a way as to not impact the applications architecture software. When the data mapping is indirect via a mediating data model, the process is also called datamediation.Other forms of data, typically associated with signal processing (including audio and imaging), can be normalized in order to provide a limited range of values within a norm.

8. FEATURE EXTRACTION

In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative, non-redundant, facilitating the subsequent learning and generalization steps, in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named features vector). This process is called feature extraction. The extracted features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

9. CLASSIFICATION

Classification consists of assigning a class label to a set of unclassified cases. It is of two types:

Supervised Classification: The set of possible classes is known in advance. The input data, also called the training set, consists of multiple records each having multiple attributes or features. Each record is tagged with a class label. The objective of classification is to analyze the input data and to develop an accurate description or model for each class using the features present in the data. This model is used to classify test data for which the class descriptions are not known. (1)

Unsupervised Classification: Set of possible classes is not known. After classification we can try to assign a name to that class. Unsupervised classification is called clustering.

The predictive data mining task that involves assigning an example to one of a set of predefined classes is called clustering. It classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data. Classification consists of assigning a class label to a set of unclassified cases. In Supervised Classification, the set of possible classes is known in advance. In unsupervised Classification, set of possible classes is not known. Unsupervised classification is called clustering.

Bayesian Classifier: The Bayesian classifier considers the following assumptions:

- The classes are mutually exclusive and exhaustive.
- The attributes are independent given the class.

These are called “Naïve” classifier because of this assumption. Bayesian classifier is defined by a set C of classes and a set A of attributes. A generic class belonging to C is denoted by cj and a generic attribute belonging to A as Ai. Consider a database D with a set of attribute values and the class label of the case. The training of the Bayesian Classifier consists of the estimation of the conditional probability distribution of each attribute, given the class. Incomplete databases seriously compromise the computational efficiency of Bayesian classifiers. One approach is to throw away all the incomplete entries. Another approach is to try to complete the database by allowing the user to specify the pattern of the data. Robust Bayesian Classifier makes no assumption about the nature of the data. It provides probability intervals that contain estimates learned from all possible completions of the database.

10. APPLICATIONS OF DATA MINING

In this section, we have focused some of the applications of data mining in respective domains:

Banking/Finance:

Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection. Data mining is used to identify customer’s loyalty by analyzing the data of customer’s purchasing activities such as the data of frequency of purchase in a period of time, total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. The higher of the score, the more relative loyal the customer is. To help bank to retain credit card customers, data mining is applied. By analyzing the past data, data mining can help banks predict customers that likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers. Credit card spending by customer groups can be identified by using data mining. The hidden correlations between different financial indicators can be discovered by using data mining. From historical market data, data mining enables to identify stock trading rules.

Data mining is used for market basket analysis:

Data mining technique is can be used in MBA(Market Basket Analysis).When the customer wants to buy some products, then this technique can help us to find out the associations between different items which the customer put in their shopping cart or baskets. Here the discovery of such associations can be identified which promotes the business techniques. The retailers uses the data mining techniques to identify the customers buying pattern .In this way this technique is used for profits of the business and also helps to identify the behavior of customers .

The data mining is used as emerging trends in the education system:

In the field of education data mining is tremendously used and is an emerging field. As every year millions of students are enrolled across the country with huge number of higher education aspirants, we believe that data mining technology can help bridging knowledge gap in higher educational systems. Data Mining helps to identify hidden patterns, associations, and anomalies from educational data and can improve decision making processes in higher educational systems. This improvement can bring advantages such as maximizing educational system efficiency, decreasing student's drop-out rate, an increasing student's promotion rate, increasing student's retention rate in, increasing student's transition rate, increasing educational improvement ratio, increasing student's success ratio, increasing student's learning outcome, and reducing the cost of system processes. In recent era we are using the KDD and the data mining tools for extracting the knowledge. The decision tree classification is frequently used in this type of applications.

Manufacturing Engineering:

When data is retrieved from manufacturing system it is used for different purposes like to find the errors in the data or product, to enhance the design methodology, to make the good quality product. The new methodology was proposed as CRISP-DM which will provides the high level detail steps of instructions for using the data mining in the engineering.

Customer Relationship Management:

Data mining technique is used in CRM .Now a days it is one of the hot topic to research in the industry because CRM have attracted both the practitioners and academics. It aims to give a research summary on the application of data mining in the CRM domain and techniques which are most often used. Research on the application of data mining in CRM will increase significantly in the future based on past publication rates and the increasing interest in the area.

Language research and Language engineering:

Sometimes linguistic information is needed about a text. A linguistic profile that contains large number of linguistic features can be generated from text file automatically using data mining. This technique found quite effective for authorship verification and recognition. The linguistic profiling of text effectively used to control the quality of language and for the automatic language verification. This method verifies automatically the text is of native quality.

Education:

Data mining methods are used in the web Education which is used to improve courseware. The relationships are discovered among the usage data picked up during student's sessions. This knowledge is very useful for the teacher or the author of the course, who could decide what modifications will be the most appropriate to improve the effectiveness of the course. Data mining techniques are one of the best learning methods. Web Education which will rapidly grow in by the application of data mining methods to educational systems can be both feasible and enhanced in the learning process.

Healthcare:

Data mining applications in health have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical for the healthcare industry to look into how data can be better captured, stored, prepared and mined. In health care, Data Mining is used for the diagnosis and prognosis of diseases and to identify the relationship that occurs among several diseases. As healthcare data are not limited to just quantitative data ,it is also necessary to explore the use of data mining to expand the scope of what health care data mining can do.

Intrusion Detection in the Network:

The intrusion detection in the Network is very difficult and needs a very close watch on the data traffic. The intrusion detection plays an essential role in computer security. The classification method of data mining is used to classify the network traffic either normal traffic or abnormal traffic. If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

Sports Data mining:

The data mining and its technique is used for an application of Sports Centre. Data mining is not only used in the business purposes but also it used in the sports .A huge number of games are available where each and every day the national and international games are to be scheduled, where a huge number of data are to be maintained .The data mining tools are applied to give the information as and when its required. The open source data mining tools like WEKA and RAPID MINER are frequently used for sports. This means that users can run their data through one of the built-in algorithms, see what results come out, and then run it through a different algorithm to see if anything different stands out. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. Data mining can be used for prediction of performance, selection of players, coaching and training and for the strategy planning. The data mining techniques are used to determine the best or the most optimal squad to represent a team in a team sport.

Intelligence Agencies:

The Intelligence Agencies collect and analyze information to investigate terrorist activities. One of the challenges to law enforcement and intelligent agencies is the difficulty of analyzing large volume of data involved in criminal and terrorist activities. Now a day the intelligence agencies are using the sophisticated data mining algorithms which makes it easy, to handle the very large databases for organizations. The different data mining techniques are used in crime data mining.

Data Mining helps to generate different types of information in the organization like personal details of the persons along with the vehicle details which can help to identify terrorist activities. The Clustering techniques are used (Association rule mining) for the different objects (like persons, organizations, vehicles etc.) in crime records. The classification technique is used to detect email spamming and String comparator is used to detect deceptive information in criminal record.

Data mining system implemented at the Internal Revenue Service:

The data mining system implemented at the Internal Revenue Service to identify high-income individuals engaged in abusive tax shelters show significantly good results. Data mining can be used to identify and rank possibly abusive tax avoidance transactions. To enhance the quality of product data mining techniques can be effectively used. The data mining technology SAS/EM is used to discover the rules those are unknown before and it can improve the quality of products and decrease the cost. A regression model and the neural network model can also be used for this purpose.

The Digital Library Retrieves:

The data mining application can be used in the field of the Digital Library where the user finds or collects stores and preserves the data which are in the form of digital mode. The data and information are available in different formats. These formats include Text, Images, Video, Audio, Picture, Maps, etc.

Data mining has been proven as a valuable tool for the banking and retail industries [4], which identify useful information from a large size data.

REFERENCES

[1] S. Kotsiantis, “Credit Risk Analysis using Hybrid Data Mining Model”, Int. Journal Intelligent Systems Technologies and Applications , Vol. 2, No. 4, 2007.

[2] L.K. Sheng and T.Y. Wah, “A comparative study of data mining techniques in predicting consumers’ credit card risk in banks”, African Journal of Business Management , Vol. 5, No. 20, Available online at http://www.academicjournals.org/AJBM, ISSN 1993-8233, 2011, pp. 8307-8312.

[3] R.F. Lopez, “Effects of missing data in credit risk scoring, A comparative analysis of methods to achieve robustness in the absence of sufficient data”, Journal of the Operational Research Society , Vol. 61, No. 3, 2010, pp. 486 -501.

[4] A.M. Hormozi and S. Giles, “Data Mining: A Competitive Weapon for Banking and Retail Industries”, Information Systems Management, Spring , Vol. 21, No. 2, ProQuest Research Library, 2004.

[5] G.C. Peng,”Credit scoring using data mining techniques”, Journal of Singapore Management Review , ISSN: 0129-5977, 2004.

[6] Y. Liu and M. Schumann, “Data mining feature selection for credit scoring models”, Journal of the Operational Research Society , Vol. 56, 1099–1108 r 2005 Operational Research Society Ltd, 2005.

[7] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, San Francisco, Morgan Kaufmann Publishers, 2012.