DATA MINING

[1]Khan Mohammed Ali Rajab Ali, [2]Mulla Anas Abdul Rasheed, [3]Sayyed Faizan Hussain Mohd Shafi, [4]Shaikh Affan Shamim.
Department of Computer Engineering,
Anjuman-i-Islams A.R. Kalsekar, New Panvel, Navi Mumbai, India.

Abstract - Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning, and by many industrial companies as an important area with an opportunity of major revenues. Researchers in many different fields have shown great interest in data mining. Several emerging applications in information-providing services, such as data warehousing and online services over.

Keywords - Data mining tasks, Softwares for data mining, Sequential patterns, Descision trees.

I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science with an overall goal to extract information (with intelligent method) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term "data mining" is in fact a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.[6] It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book Data mining: Practical machine learning tools and techniques with Java (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons] Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

II. ACTUAL DATA MINING

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

III. RESEARCH

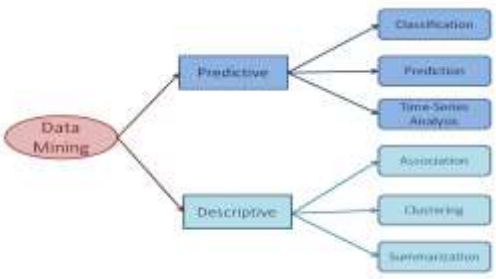
The premier professional body in the field is the Association for Computing Machinery's (ACM) Special Interest Group (SIG) on Knowledge Discovery and Data Mining (SIGKDD). Since 1989 this ACM SIG has hosted an annual international conference and published its proceedings, and since 1999 it has published a biannual academic journal titled "SIGKDD Explorations".

Computer science conferences on data mining include:

- CIKM Conference – ACM Conference on Information and Knowledge Management
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- KDD Conference – ACM SIGKDD Conference on Knowledge Discovery and Data Mining

Data mining topics are also present on many data management/database conferences such as the ICDE Conference, SIGMOD Conference and International Conference on Very Large Data Bases.

IV. DATA MINING TASKS



Data mining involves six common classes of tasks:

- Anomaly detection (outlier/change/deviation detection)– The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (dependency modeling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – 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 – 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 – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

V. BACKGROUND

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Baye's theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

VI. PROCESS

The knowledge discovery in databases (KDD) process is commonly defined with the stages:

- Selection
- Pre-processing
- Transformation
- Data mining
- Interpretation/evaluation.

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

or a simplified process such as (1) Pre-processing, (2) Data Mining, and (3) Results Validation.
Polls conducted in 2002, 2004, 2007 and 2014 show that the CRISP-DM methodology is the leading methodology used by data miners. The only other data mining standard named in these polls was SEMMA. However, 3–4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models, and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.

VII. FREE OPEN-SOURCE DATA MINING SOFTWARES AND APPLICATIONS

The following applications are available under free/open source licenses. Public access to application source code is also available..

- Massive Online Analysis (MAO): a real-time big data stream mining with concept drift tool in the Java programming language.
- MEPX - cross platform tool for regression and classification problems based on a Genetic Programming variant.
- ML-Flex: A software package that enables users to integrate with third-party machine-learning packages written in any programming language, execute classification analyses in parallel across multiple computing nodes, and produce HTML reports of classification results.
- MLPACK library: a collection of ready-to-use machine learning algorithms written in the C++ language.
- Orange: A component-based data mining and machine learning software suite written in the Python language.
- R: A programming language and software environment for statistical computing, data mining, and graphics. It is part of the GNU Project.
- Torch: An open source deep learning library for the Lua programming language and scientific_computing framework with wide support for machine learning algorithms.
- UIMA: The UIMA (Unstructured Information Management Architecture) is a component framework for analyzing unstructured content such as text, audio and video – originally developed by IBM.
- Weka: A suite of machine learning software applications written in the Java programming language.

VIII. PROPRIETARY DATA-MINNING SOFTWARES AND APPLICATIONS

The following applications are available under proprietary licenses.

- Angnoss KnowledgeSTUDIO: data mining tool.
- Clarabridge: text analytics product.
- KXEN Modeler: data mining tool provided by KXEN Inc.

- LIONsolver: an integrated software application for data mining, business intelligence, and modeling that implements the Learning and Intelligent OptimizatioN (LION) approach.
- Microsoft Analysis Service data mining software provided by Microsoft.
- NetOwl: suite of multilingual text and entity analytics products that enable data mining.
- OpenText Big Data Analytics: Visual Data Mining & Predictive Analysis by Open Text Corporation.
- Oracle Data Mining: data mining software by Oracle Corporation.
- Pseven: platform for automation of engineering simulation and analysis, multidisciplinary optimization and data mining provided by DATADVANCE.
- Qlucore Omics Explorer: data mining software.
- RapidMiner: An environment for machine learning and data mining experiments.
- SAS Enterprise Miner: data mining software provided by the SAS Institute.
- SPSS Modeller: data mining software provided by IBM.
- STATISTICA Data Miner: data mining software provided by StarSoft.
- Tanagra: Visualisation-oriented data mining software, also for teaching.
- Vertica: data mining software provided by Hewlett Packard.

IX. 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.

A. Association

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is 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 the customer and increase sales.

B. 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.

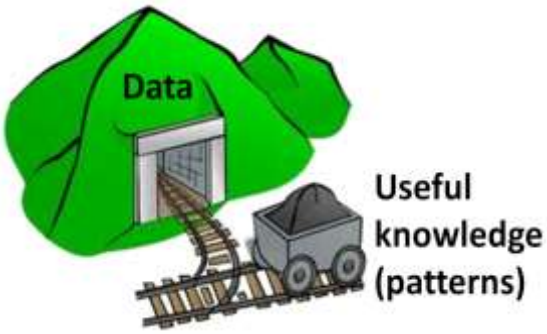
C. Clustering

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.

D. Prediction

The prediction, as its name implied, is one of a data mining techniques that discovers 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.

E. 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.

F. Decision trees

The A decision tree is one of the most commonly 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:

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.

X. REFERENCES

[1] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[2] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[3] Han, Kamber, Pei, Jaiwei, Micheline, Jian (June 9, 2011). Data Mining: Concepts and Techniques (3rd ed.).

[4] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases"

XI. CONCLUSION

Data mining, along with traditional data analysis, is a valuable tool that that is being used in Strategic Enrollment Management to achieve desired enrollment targets in colleges and universities. In situations where it has been applied, it has been proven to successfully predict enrollment, at least to a degree.