Asian Journal of Applied Science and Technology (AJAST)
(Open Access Quarterly International Journal) Volume 2, Issue 2, Pages 947-953, 2018

An Overview of Data Mining Techniques and Their Application in Industrial Engineering

Keerthi Sumiran

Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee.
Roorkee- Haridwar Highway, Roorkee, Uttarakhand, 247667, India. Email: Keerthisumiran@tutanota.com

Article Received: 26 February 2018 *Article Accepted: 27 April 2018* *Article Published: 05 June 2018*

ABSTRACT

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. This paper contains an overview of data mining including the concepts behind what it is and the variations on how it is accomplished. It will cover the history on the evolution of data mining and how it got to where it is today will. The process of knowledge discovery will also be covered both leading up to and after the data mining phase. The different techniques used in data mining, and a more in-depth look at the algorithms for the predictive data mining technique known as sequential pattern mining. There is also a view at what visual data mining is and a short insight on web mining and how sequential data mining can be applied to it.

Keywords: Data Mining, Neural Networks, Sequential Patterns.

1. INTRODUCTION

In today's world, there are many ways to apply data mining concepts and techniques. The most known way of applying these data mining techniques is in the field of business (Agrawal, Imielinski, & Swami, 1993). Businesses have many uses of data mining. It can be used internally in the business, for example, to predict employee actions and used externally to identify possible customers and to help efficiently market its products (Agrawal, Imielinski, & Swami, 1993). Data mining can do these things in many different ways. Data mining is not limited to just business. For example, hessami et al (2017) used data mining techniques to provide an enhanced cost estimating and project development procedures for Metropolitan Organizations and they come up with decent model at the end of their project (Hessami, et al., 2017). It is being developed to work in the field of health care as well as many different sciences. One of the latest applications of data mining can be find in Torabi et al (2018) where they developed a new prediction model for energy production with wind turbines (Torabi, Kiaian Mousavy, Dashti, Saeedi, & Yousefi, 2018).

2. BACKGROUND – WHAT IS DATA MINING?

Data mining is a combination of many different disciplines. Figure 1 below gives a sample of some of these disciplines.

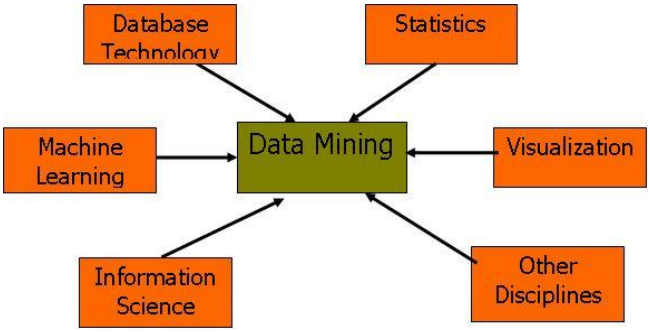

Figure 1: Some of the many different disciplines that are combined into data mining

 Online ISSN: 2456-883X

Data mining gets its name from the similarities there are between searching for valuable information in large data sets and mining a mountain for valuable ore. Data mining, in the simplest sense, is knowledge discovery (Bengio, Buhmann, Embrechts, & Zurada, 2012). It consists of the searching, analyzing and sifting through large data sets to find new patterns, trends, and relationships contained within. There are three general properties that the discovered knowledge should satisfy. It should be accurate, comprehensible, and interesting. Chu et al (2017) applied these 3 components and provide a model that is very useful in the oil and gas industry; by applying PSO-ANFIS approach, the model enables engineers to calculate the wax deposition produced in the pipelines (Chu, Sasanipour, Saeedi, Baghban, & Mansoori, 2017).

3. HISTORY OF DATA MINING

Data mining has developed greatly from what it started out as in the beginning. Data mining's roots can be traced back on three paths.

- The oldest of these paths being classic statistics. Statistics is the basis of data mining. There would not be any way of measuring the data without it. It is the oldest and most crucial part of data mining.
- Another path of data mining is artificial intelligence. Artificial intelligence focuses on what are called experience-based techniques for knowledge discovery. This attempts to apply human thought processes to the statistical problems.
- The third path of data mining is more of a combination of the previous two. It is machine learning. It brought experience-based techniques together with advanced statistical analysis (Zurada,1992).

The term data mining was coined in the 1960s. Data collection abilities were starting in the 1960s. Data mining was used to find basic information from the collections such as total revenue over the last three years. The technologies that made this possible consisted of tapes, disks, and computers. The 1980s brought true databases into wider use. This allowed for easier data access via SQL. The 1990s came and brought the use of data warehouses and decision support (Berry, Lindoff, 1997). This was the time-period when a lot of the data mining that we see today was developed.

4. DATA MINING

The knowledge discovery process usually involves seven phases from start to finish.

- Phase 1 – Data Integration
    - Collect data from sources
- Phase 2 – Data Selection
    - Select useful data
- Phase 3 – Data Cleaning
    - Rid data of errors, missing values, inconsistent data
- Phase 4 – Data Transformation
    - Normalization, smoothing, other forms appropriate for data mining

- Phase 5 – Data Mining
    - Apply mining techniques to discover patterns
- Phase 6 – Pattern Evaluation / Presentation
    - Visualization and removing redundant patterns
- Phase 7 – Knowledge Discovery
    - Use to make decisions (Bengio, Buhmann, Embrechts, & Zurada, 2012)

The order of the first three phases is somewhat debatable. It depends on if you want to clean the data before you integrate it or not. Figure 2 below shows the process described above.

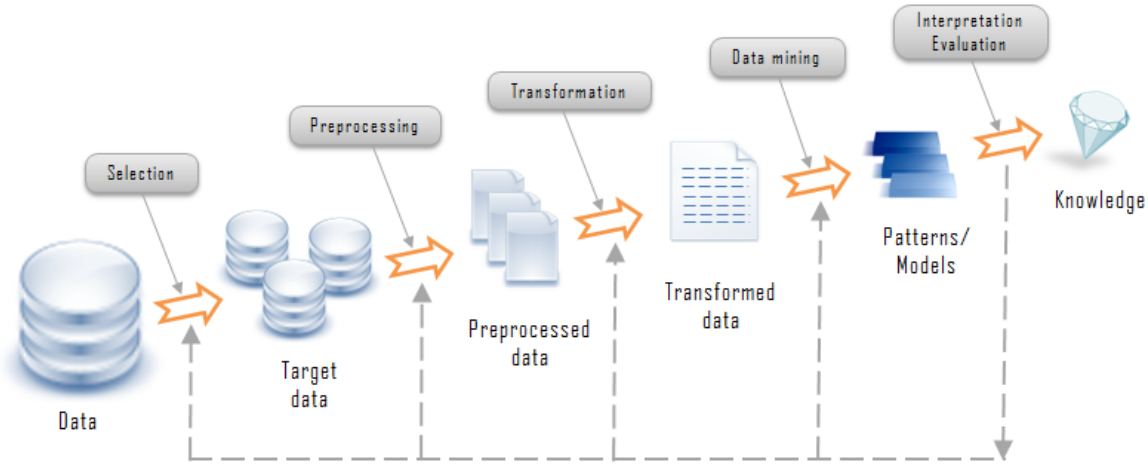

Figure 2: Visual of the phases involved with data mining

Data mining itself is generally split into two categories: descriptive data mining and predictive data mining. Descriptive data mining explores interesting patterns to describe the data while predictive data mining forecasts the behavior of the model based on available data set.

Predictive

Prediction is a data mining category that focuses on discovering a relationship between independent variables and a relationship between dependent and independent variables. Predictive data mining can be used to forecast explicit values based on patterns in the data. Predictive Data Mining is usually applied with the goal to identify a statistical or neural network model that can be used to predict some kind of interesting result. Prediction analysis techniques can be used, for instance, in sales to predict future profit based on previous sales activity.

Descriptive

Descriptive data mining describes a data set in a brief but comprehensive way and gives interesting characteristics of the data without having any predefined target. Descriptive techniques do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness, etc. of the data. These methods take the data given to them and show how things are related. They “describe” the data.

5. DIFFERENT TECHNIQUES

Association

One of the most known data mining technique is association. When using association, a pattern is discovered based on a relationship of a specific item with other items in the same transaction. Association is a predictive data mining technique.

An example would be to look at what products a customer usually buys together such as bread and sandwich meat. With this kind of relationship between the two products, a business could produce a marketing plan involving both products such as combining both products in a coupon for price reduction or for TV commercials.

Classification

Classification is a data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of the predefined sets of classes or groups. The classification technique makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. Classification is a descriptive data mining technique in which programmers create the smart software in a way that allows it to learn how to classify the data items into groups.

An example of classification would be: we could use classification to sort all people in the town into the categories of dog person or cat person. It would have information on known dog and cat people and would take in the information from the new group and compare different traits and place people in the appropriate, predefined, areas.

Clustering

Clustering is a data mining technique that partitions a data set into groups where objects from the same group are as similar as possible, and objects from other groups are well differentiated. Different from classification, the clustering technique also defines the classes and put objects in them, while in the classification technique, objects are assigned into predefined classes. Clustering is a descriptive data mining technique but is sometimes considered to be a predictive technique. This may be because in some cases, the data undergoes clustering as a pre-step before a predictive technique is applied.

Take a library as an example of clustering. There are a wide range of topics available. By using the clustering technique, books that have similarities can be in one cluster or on one shelf and be labeled with a meaningful name. If a reader wants to grab books on a specific topic, he or she would only go to that shelf instead of searching the whole library.

Regression

Regression is a data mining technique that is used to predict numbers from data sets that have known target values. Regression is a predictive data mining technique.

Examples of situations where regression could be applied are: sales, distance, temperature, etc. Regression could be used to predict the value of a house based on location, number of rooms, lot size, etc by observing data on past houses over time. The known target for this example would be the house value.

Sequential Patterns

Sequential pattern mining discovers frequent subsequences as patterns in a sequence database. The uncovered patterns are used for further analysis to recognize relationships among data. Sequential pattern analysis is a predictive data mining technique. What is a sequence database, you may ask? It is a database that stores a number of records as sequences of ordered events that may or may not have a definite notion of time.

Sequential Pattern Mining Algorithms

A reliable sequential pattern mining algorithm should provide acceptable performance measures such as low execution time and low memory utilization when mined with low minimum support values and should be scalable. There are three categories that the main sequential pattern mining techniques fall into.

- Apriori-based
- Pattern-growth
- Early-pruning

Apriori-based algorithms follow the apriori property which states that "all nonempty subsets of a frequent itemset must also be frequent." To be "frequent" a set must have support that is greater than the specified minimum support. Some algorithms that are apriori-based include AprioriAll, GSP, PSP, and SPAM. Pattern-growth algorithms perform a divide-and-conquer strategy. Sequence databases are recursively projected into a set of smaller projected databases based on the current sequential pattern(s). The sequential patterns are grown in each projected database by exploring only locally frequent fragments. Some pattern-growth algorithms include FreeSpan, PrefixSpan, WAP-mine, and FS-Miner. Early-pruning algorithms utilize a sort of position induction to prune candidate sequences very early in the mining process and to avoid support counting as much as possible. Some examples of early-pruning algorithms are LAPIN, HVSM, and DISC-all.

Visual Data Mining

Visual data mining combines traditional mining methods and information visualization techniques. The user is directly involved in the exploration process. It takes advantage of the automatic calculations and the capabilities of human perception to extract structures from pictures. A visual data mining system should have simplicity, reliability, reusability, availability, and security. Simplicity is the most important when creating a visual data mining system. The interface must be easy to use and the data displayed must be easily interpreted. These systems should aid in the guiding of the user through the process of knowledge discovery. Programmers must build these interfaces in such a way that it allows for accurate visual presentations of the data when used by humans to interpret

the data. Examples of possible views when working with a type of visual data mining system are shown here in Figure 3 and Figure 4 below.

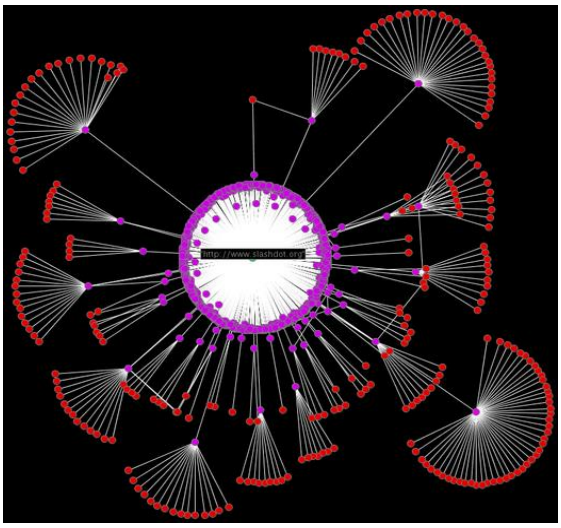

Figure 3: This depicts a possible visual that a user might see.

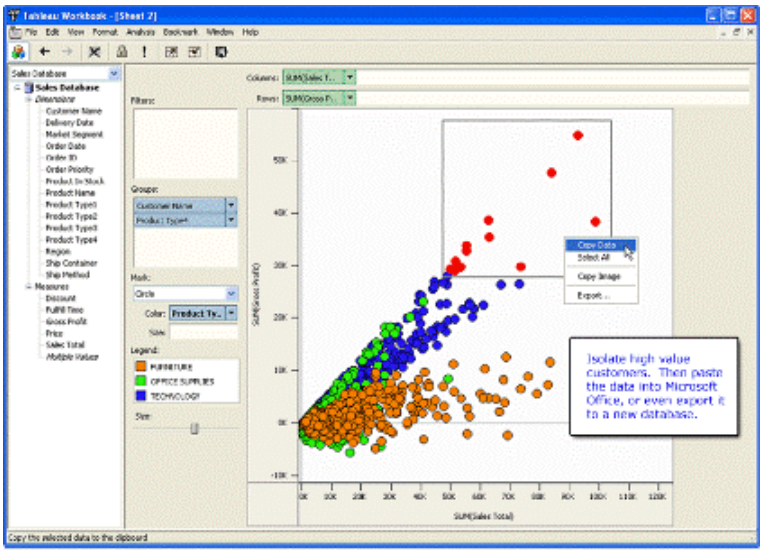

Figure 4: Another depiction of possible visual representation of data.

Web Mining

Web mining allows the searching for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines. Structure mining is used to examine data related to the structure of a particular Web site. Usage mining is used to examine data related to a particular user's browser as well as data gathered by forms that the user may have submitted during Web transactions. Web mining, as a sequential pattern mining application, is concerned with finding user navigational patterns on the World Wide Web by extracting knowledge from web logs. An example of applying sequential pattern mining in this topic would be searching for a pattern in sites visited. Given websites *a* thru *f*, if a log pattern showed a frequent sequence, *abac*, for instance, you would see that a user that visits *a* tends to visit *b* and then re-visit *a* before continuing on to *c*.

6. CONCLUSION

Data mining today is of high importance in the business and scientific fields. This approach at problem solving is only limited by the technology that exists at the time. More and more data is being accumulated every day. Data mining will continue to grow and develop beyond what it is today. Better, faster, more efficient algorithms will be surfacing. One important thing to keep in mind is the importance of data mining on the ever developing world. It is suggested to do further researches on other applications of data mining and how they effect on different industries.

REFERENCES

Agrawal, R., Imielinski, T., Swami, A. (1993), "Database Mining: A Performance Perspective", *IEEE Transactions on Knowledge and Data Engineering*, pp. 914- 925, December 1993.

Bengio Y., Buhmann J. M., Embrechts M., and Zurada J.M (2012). Introduction to the special issue on neural networks for data mining and knowledge discovery. *IEEE Trans. Neural Networks.*

Berry, J. A., Lindoff, G. (1997), Data Mining Techniques, Wiley Computer Publishing (ISBN 0-471-17980-9).

Berson (2011), "Data Warehousing, Data-Mining & OLAP", TMH.

Craven M. W. and Shavlik J. W. (1997). Using neural networks for data mining. Future Generation Computer Systems, 13:211–229.

Chu, Z.-Q., Sasanipour, J., Saeedi, M., Baghban, A., & Mansoori, H. (2017). Modeling of wax deposition produced in the pipelines using PSO-ANFIS approach. *Petroleum Science and Technology*, 1974-1981.

Haykin, S. (1999), Neural Networks, Prentice Hall International Inc.

Hessami, A. R., Sun, D., Odreman, G. J., Zhou, X., Nejat, A., & Saeedi, M. (2017). *Project Scoping Guidebook for Metropolitan Planning Organization Transportation Projects.* Kingsville, Texas: Texas A&M University Kingsville (TAMUK).

Khajanchi, Amit (2013), Artificial Neural Networks: The next intelligence.

Torabi, A., Kiaian Mousavy, S., Dashti, V., Saeedi, M., & Yousefi, N. (2018). A New Prediction Model Based on Cascade NN for Wind Power Prediction. *Computational Economics*.

Zurada J.M. (1992), "An introduction to artificial neural networks systems", St. Paul: West Publishing.