Emerging Trends in Data Mining

Mrs. I. Faritha Beevi MCA., M. Phil., SET
Sadakathullah Appa College
Palayamkottai – 627 011 India.

Abstract

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. Data mining techniques are used in many research areas, including mathematics, cybernetics, genetics and marketing. While data mining techniques are a means to drive efficiencies and predict customer behaviour, if used correctly, a business can set itself apart from its competition through the use of predictive analysis. In general, the benefits of data mining come from the ability to uncover hidden patterns and relationships in data that can be used to make predictions that impact businesses. Data mining is one of the most widely used methods to extract data from different sources and organize them for better usage. In spite of having different commercial systems for data mining, a lot of challenges come up when they are actually implemented. With rapid evolution in the field of data mining, companies are expected to stay abreast with all the new developments. Complex algorithms form the basis for data mining as they allow for data segmentation to identify various trends and patterns, detect variations, and predict the probabilities of various events happening. The raw data may come in both analog and digital format, and is inherently based on the source of the data. Companies need to keep track of the latest data mining trends and stay updated to do well in the industry and overcome challenging competition.

Keywords – Data Mining, Cybernetics, genetics, predictive analysis, Complex Algorithms, Segmentation, Data Mining Trends.

I. Introduction

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [1]. The iterative process consists of the following steps:

•**Data cleaning**: It is also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

•**Data integration**: At this stage, multiple data sources, often heterogeneous, may be combined in a common source.

•**Data selection**: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

•**Data transformation**: It is also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

•**Data mining**: It is the crucial step in which clever techniques are applied to extract patterns potentially useful.

•**Pattern evaluation**: In this step, strictly interesting patterns representing knowledge are identified based on given measures.

•**Knowledge representation**: It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data. The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

II. Data Mining Processes

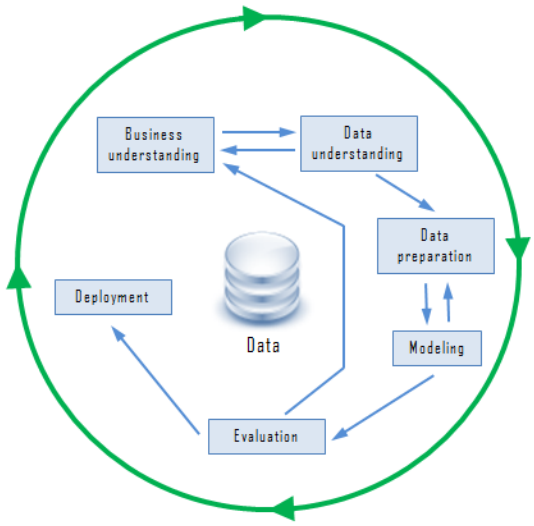
Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence (AI) and statistical [2].

Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace… etc, to increase their business efficiency. Therefore, the needs for a standard data mining process increased dramatically. A data mining process must be reliable and it must be repeatable by business people with little or no knowledge of data mining background. As the result,

in 1990, a cross-industry standard process for data mining (CRISP-DM) first published after going through a lot of workshops, and contributions from over 300 organizations.

The Cross-Industry Standard Process for Data Mining (CRISP-DM)

Cross-Industry Standard Process for Data Mining (CRISP-DM) consists of six phases intended as a cyclical process as the following figure:



Cross-Industry Standard Process for Data Mining (CRISP-DM)

Business understanding

In this phase, it is required to understand business objectives clearly and find out what are the business's needs. Next, we have to assess the current situation by finding the resources, assumptions, constraints and other important factors which should be considered. Then, from the business objectives and current situations, we need to create data mining goals to achieve the business objectives within the current situation. Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible.

Data understanding

First, the data understanding phase starts with initial data collection, which we collect from available data sources, to help us get familiar with the data. Some important activities must be performed including data load and data integration in order to make the data collection successfully. Next, the "gross" or "surface" properties of acquired data need to be examined carefully and reported. Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting, and visualization. Finally, the data quality must be examined.

Data preparation

The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.

Modelling

Modelling techniques have to be selected to be used for the prepared dataset. Next, the test scenario must be generated to validate the quality and validity of the model. Then, one or more models are created by running the modelling tool on the prepared dataset. Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives.

Evaluation

In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to the new patterns that have been discovered in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

Deployment

The knowledge or information, which we gain through data mining process, needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization. In the deployment phase, the plans for deployment, maintenance, and monitoring have to be created for implementation and also future supports. From the project point of view, the final report of the project needs to summary the project experiences and reviews the project to see what need to improved created learned lessons.

The CRISP-DM offers a uniform framework for experience documentation and guidelines. In addition, the CRISP-DM can apply in various industries with different types of data [3].

III. Data Mining Techniques

One of the most important tasks in Data Mining is to select the correct data mining technique. Data mining technique has to be chosen based on the type of business and the type of problem your business faces. A generalized approach has to be used to improve the accuracy and cost effectiveness of using data mining techniques. There are basically seven main Data Mining techniques which are discussed in this article. There are also a lot of other Data Mining techniques but these seven are considered more frequently used by business people [4].

- Statistics
- Clustering
- Visualization
- Decision Tree
- Association Rules
- Neural Networks
- Classification

1. Statistical Techniques

Data mining techniques statistics is a branch of mathematics which relates to the collection and description of data. Statistical technique is not considered as a data mining technique by many analysts. But still it helps to discover the patterns and build predictive models. For this reason data analyst should possess some knowledge about the different statistical techniques [7]. In today's world people have to deal with large amount of data and derive important patterns from it. It also helps in providing information about the data with ease. Through statistical reports people can take smart decisions. There are different forms of statistics but the most important and useful technique is the collection and counting of data. There are a lot of ways to collect data like Histogram, Mean, Median, Mode, Variance, Max, Min, and Linear Regression.

2. Clustering Technique

Clustering is one among the oldest techniques used in Data Mining. Clustering analysis is the process of identifying data that are similar to each other [8]. This will help to understand the differences and similarities between the data. This is sometimes called segmentation and helps the users to understand what is going on within the database. There are different types of clustering methods. They are as follows

- Partitioning Methods
- Hierarchical Agglomerative methods
- Density Based Methods
- Grid Based Methods
- Model Based Methods

The most popular clustering algorithm is Nearest Neighbour. Nearest neighbour technique is very similar to clustering [9]. It is a prediction technique where in order to predict what an estimated value is in one record look for records with similar estimated values in historical database and use the prediction value from the record which is near to the unclassified record. This technique simply states that the objects which are closer to each other will have similar prediction values. Through this method you can easily predict the values of nearest objects very easily. Nearest Neighbour is the most easy to use technique because they work as per the thought of the people. They also work very well in terms of automation. They perform complex ROI calculations with ease. The level of accuracy in this technique is as good as the other Data Mining techniques. In business Nearest Neighbour technique is most often used in the process of Text Retrieval. They are used to find the documents that share the important characteristics with that main document that have been marked as interesting.

3. Visualization

Visualization is the most useful technique which is used to discover data patterns. This technique is used at the beginning of the Data Mining process [10]. Many researches are going on these days to produce interesting projection of databases, which is called Projection Pursuit. There are a lot of data mining techniques which will produce useful patterns for good data. But visualization is a technique which converts Poor data into good data letting different kinds of Data Mining methods to be used in discovering hidden patterns.

4. Induction Decision Tree Technique

A decision tree is a predictive model and the name itself implies that it looks like a tree. In this technique, each branch of the tree is viewed as a classification question and the leaves of the trees are considered as partitions of the dataset related to that particular classification. This technique can be used for exploration analysis, data pre-processing and prediction work. Decision tree can be considered as a segmentation of the original dataset where segmentation is done for a particular reason. Each data that comes under a segment has some similarities in their information being predicted. A decision tree provides results that can be easily understood by the user. Decision tree technique is mostly used by statisticians to find out which database is more related to the problem of the business. Decision tree technique can be used for Prediction and Data pre-processing [11]. The first and foremost step in this technique is growing the tree. The basic of growing the tree depends on finding the best possible question to be asked at each branch of the tree. The decision tree stops growing under any one of the below circumstances

- If the segment contains only one record
- All the records contain identical features
- The growth is not enough to make any further spilt

CART which stands for Classification and Regression Trees is a data exploration and prediction algorithm

which picks the questions in a more complex way [12]. It tries them all and then selects one best question which is used to split the data into two or more segments. After deciding on the segments it again asks questions on each of the new segment individually. Another popular decision tree technology is CHAID (Chi-Square Automatic Interaction Detector). It is similar to CART but it differs in one way. CART helps in choosing the best questions whereas CHAID helps in choosing the splits.

5. Neural Network

Neural Network is another important technique used by people these days. This technique is most often used in the starting stages of the data mining technology [13]. Artificial neural network was formed out of the community of Artificial intelligence. Neural networks are very easy to use as they are automated to a particular extent and because of this the user is not expected to have much knowledge about the work or database. There are two main parts of this technique – the node and the link. A neural network is a collection of interconnected neurons, which could form a single layer or multiple layers. The formation of neurons and their interconnections are called architecture of the network. There are a wide variety of neural network models and each model has its own advantages and disadvantages. Every neural network model has different architectures and these architectures use different learning procedures. Neural networks are very strong predictive modelling technique. But it is not very easy to understand even by experts. It creates very complex models which is impossible to understand fully. Thus to understand the Neural network technique companies are finding out new solutions. Neural network has been used in various kinds of applications. This has been used in the business to detect frauds taking place in the business.

6. Association Rule Technique

This technique helps to find the association between two or more items. It helps to know the relations between the different variables in databases. It discovers the hidden patterns in the data sets which is used to identify the variables and the frequent occurrence of different variables that appear with the highest frequencies. Association rule offers two major information's are support and confidence [14]. This technique follows a two step process. There are three types of association rule. They are

- Multilevel Association Rule
- Multidimensional Association Rule
- Quantitative Association Rule

This technique is most often used in retail industry to find patterns in sales. This will help increase the conversion rate and thus increases profit.

7. Classification

Data mining techniques classification is the most commonly used data mining technique which contains a set of pre classified samples to create a model which can classify the large set of data. This technique helps in deriving important information about data and metadata (data about data). This technique is closely related to cluster analysis technique and it uses decision tree or neural network system [15]. The two main processes involved in this technique are learning and classification. There are different types of classification models. They are as follows

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

IV. Trends in Data Mining

Mining personal data: In the past, the level of personal information that business organizations were able to collect on their consumers was limited. It was possible to see what users shared, their emails, names, gender and in some cases location. However, today there are endless possibilities when it comes to finding and digging up personal information online. As an example you can look at Facebook allowing all apps, or companies that have received a permission from the users to see their personal details using Facebook, to access various personal data from sexual preferences, to location, education and many others. Additionally, Facebook has allowed many large advertising companies to target people on various personal details without asking for any consent from the Facebook users. Most people weren't even aware of how certain ads were targeting them. However, soon enough, from this trend, yet another one appeared [5].

More transparency is required: Despite the fact that everyone is gathering valuable user data today, the people whose data is collected, the consumers, are usually unaware of this, but people have started educating themselves about online data mining techniques. Given the fact that this lack of transparency has been around since the beginning of the data mining practice, it's only logical that there are consumers who have educated themselves about data mining and they don't like the fact that they aren't notified that someone is gathering their data. There are many companies and organizations that have already recognized this, which is why they are trying to do something about this unwillingness that has been building up in consumers. Users will soon have analytic tools that they can use to see which information was gathering on them by ad engines. In the future, companies and marketers alike will have to consider consumers as well, and make sure that they don't gather any data their consumers don't want them to.

Life science data mining: Yet another trend that is expected to explode this year is the increased data for life sciences [6]. People create enormous amounts of data every day and it's only expected that these numbers will grow in the future, given the fact that people will only adopt more and more technology interfaces as time goes by. Gathering, structuring and understanding all of this data can be of great help to the life sciences sector and make it more efficient and focused in their development and research processes.

Predictive data mining: A new growing trend of predictive analytics is visible in various sectors and it consists of extracting valuable information from existing data to be able to forecast future outcomes and probabilities [7]. This is a data mining technique that focuses only on past data. This is a method of determining and finding certain categories of data and analyzing it for the purpose of estimating certain outcomes. Additionally it also includes risk assessment and giving a couple of alternative scenarios. Various organizations use this data mining technique to decide their next business moves, launch products, marketing campaigns etc. These analytical results can help a business backup their new campaigns or even help them determine their future steps. All of these trends show that data mining is going to be exciting right up until the very end, at least when it comes to data mining. All industries are affected by data mining and its changes. Smart investors who are brave enough should try to grab the opportunity to invest in data, as its possibilities are endless at the moment. Who knows where it can take us in the future and how important it will be.

V. Future Trends in Data Mining

Businesses which have been slow in adopting the process of data mining are now catching up with the others. Extracting important information through the process of data mining is widely used to make critical business decisions [8]. In the coming decade, we can expect data mining to become as ubiquitous as some of the more prevalent technologies used today. Some of the key data mining trends for the future include -

1. **Multimedia Data Mining**
   This is one of the latest methods which is catching up because of the growing ability to capture useful data accurately. It involves the extraction of data from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. and the data is converted into a numerical representation in different formats [9]. This method can be used in clustering and classifications, performing similarity checks, and also to identify associations.
2. **Ubiquitous Data Mining**
   This method involves the mining of data from mobile devices to get information about individuals [10]. In spite of having several challenges in this type such as complexity, privacy, cost, etc. this method has a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.
3. **Distributed Data Mining**
   This type of data mining is gaining popularity as it involves the mining of huge amount of information stored in different company locations or at different organizations [11]. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based upon them.
4. **Spatial and Geographic Data Mining**
   This is new trending type of data mining which includes extracting information from environmental, astronomical, and geographical data which also includes images taken from outer space [12]. This type of data mining can reveal various aspects such as distance and topology which is mainly used in geographic information systems and other navigation applications.
5. **Time Series and Sequence Data Mining**
   The primary application of this type of data mining is study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events which occur outside the normal series of events [13]. This method is mainly being use by retail companies to access customer's buying patterns and their behaviours.

VI. Conclusion

Data mining is a technology that has emerged to provide organizations whether large or small the opportunity to discovery-hidden trends and patterns in their data. This realization has come about as a result of the increasing loads of data being stored in organization's databases. To take advantage of this storage data mining can use a data warehouse to manage the data before applying a data mining application. In summary, data mining works by simply learning from data it can. Models are created automatically from the data and represent an unbiased distillation of the business experience. In this paper we try to briefly review the various data mining trends from its inception to the future. We found that Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. So, data mining will be more and more useful in future.

VII. References and Links

Links

[1]http://www.exinfm.com/pdffiles/intro_dm.pdf

[2]http://www.zentut.com/data-mining/data-mining-processes/

[3]http://web.mit.edu/profit/PDFS/AldanaW.pdf

[4]https://www.educba.com/7-data-mining-techniques-for-best-results/

[5]https://www.smallbizdaily.com/emerging-trends-data-mining/

[6]https://www.worldscientific.com/worldscibooks/10.1142/6268

[7]https://www.techopedia.com/definition/30597/predictive-data-mining

[8]https://www.flatworldsolutions.com/data-management/articles/data-mining-future-trends.php

[9]http://airccse.org/journal/ijcga/papers/5115ijcga05

[10]https://www.igi-global.com/dictionary/ubiquitous-data-mining/34845

[11]https://www.csee.umbc.edu/~hillol/PUBS/review.pdf

[12]https://www.nap.edu/read/10661/chapter/5

[13]https://www.slideshare.net/dataminingtools/mining-stream-time-series-and-sequence-data

References

1. Data Mining: Extending the Information Warehouse Framework: Data Management Solutions. IBM Whitepaper. IBM. 2000.

2. Data Mining: An Introduction. SPSS Whitepaper. SPSS. 2000.

3. An Introduction to Data Mining. Pilot Software Whitepaper. Pilot Software. 1998.

4.Data Mining Market Share. Data Mining News. Volume 1, Number 18. May

5. W.H. Inmon. Tech Topic: What is a Data Warehouse? Prism Solutions. Volume 1. 1995.

6.Data Mining News. Intelligent Data Analysis Group. 2000.

7. Edelstein, Herb. Data Mining News "Two Crows Releases 1999 Technology Report". Volume 2, number 18. 10 May 1999.

8. Cisco, General Growth Properties Link Mall Retailers Online. InternetNews.Com. May 12, 2000.

9. Offline Spending by Internet Brands Passes $1 Billion. Internet Newsletter. 1999.

10. Data Mining Concepts and Techniques – Jiawei Han & Micheline Kamber.

11. DataMining: concepts,Models,Methods and Algorithms, Wiley-Interscience,Hoboken,Nj (2003).

12. Modern Data Warehousing, Mining and Visualization Core Concepts by George M. Marakas.