AN OVERVIEW OF DATA MINING

Rishabh Kadyan , Nancy Arora , Paras Chhabra

Abstract—. The rapid progress of computers and databases has enable companies to store data about future use. The sheer amounts of data to be analyzed in order to make better decisions require dramatically improved new automated data modelling technologies. A concept of **Data Mining** is developed. There are two foundation of using data mining techniques: the availability of large amount of data and the data mining modelling techniques. In this paper I have explained the whole overview of Data Mining, that comprises all the techniques used,all the real time and the actual applications that are possible and through which some of the companies have made tremendous growth. In my paper I have written everything right from the history of data mining to the future prospects.

Index Terms— Analytical model, anomalous data, artificial neural network,clustering, data cleansing,data navigation, data visualization.

1 INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"Data Mining helps marketing professionals improve their understanding of customer behavior. In turn, this better understanding allows them to target marketing campaigns more accurately and to align campaigns more closely with the needs, wants and attitudes of customers and prospects.

There are several benefits of using data mining. Customized targeting at the right time Data Mining enables

Companies to reach consumers with the right product and the right offer at the right time.Data Mining is a process of looking for unknown relationships and patterns and extracting useful information volumes of data in data warehouse Data Mining, by its simplest definition, automates the detection of relevant patterns in a database.

.

2 HISTORY

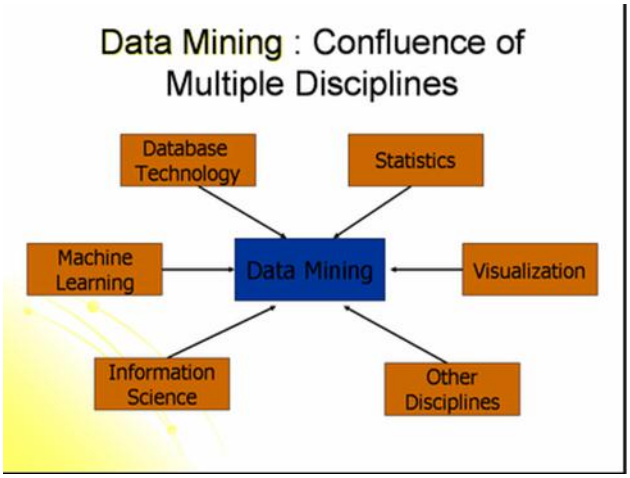
Data mining roots are traced back along three family lines. The longest of these three lines is classical statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Classical statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

Data mining's second longest family line is artificial intelligence, or AI. This discipline, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. The notable exceptions were certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS).The third family line of data mining is machine learning, which is more accurately described as the union of statistics and AI

- *Rishabh Kadyan, currently pursuing B.tech degree in Computer Science Engineering in Dronacharya College Of Engineering, Gurgaon, Haryana, India PH-9873318842. E-mail: @gmail.com*
- *Nancy Arora, currently pursuing.tech degree in Computer Science Engineering in Dronacharya College Of Engineering, Gurgaon,Haryana,India PH-9873397994. E-mail: arora.nancy29@gmail.com*
- *Paras Chhabra, currently pursuing.tech degree in Computer Science Engineering i n Dronacharya College Of Engineering, Gurgaon ,Haryana,India PH-9873364401. E-mail: @gmail.com*

While AI was not a commercial success, its techniques were largely co-opted by machine learning. Machine learning, able to take advantage of the ever-improving price/performance ratios offered by computers of the 80s and 90s, found more applications because the entry price was lower than AI. Machine learning could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.

Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns within. Data mining is finding increasing acceptance in science and business areas which need to analyze large amounts of data to discover trends which they could not otherwise find.



3 DATA MINING TECHNIQUES

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns. We will briefly examine those data mining techniques with example to have a good overview of them.

3.1 ASSOCIATION

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

3.2 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 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 make the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group.

3.3 CLUSTERING

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

3.4 PREDICTION

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. For instance,prediction analysis technique can be used in sale to predict profit for the future if we consider 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.

3.5 SEQUENTIAL PATTERNS

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

There are also some traditional and classical Techniques such as : Statistics, Neighbourhoods and Clustering

The Classics

These two sections have been broken up based on when the data mining technique was developed and when it became technically mature enough to be used for business, especially for aiding in the optimization of customer relationship man-

agement systems. Thus this section contains descriptions of techniques that have classically been used for decades the next section represents techniques that have only been widely used since the early 1980s.

This section should help the user to understand the rough differences in the techniques and at least enough information to be dangerous and well armed enough to not be baffled by the vendors of different data mining tools.

The main techniques that we will discuss here are the ones that are used 99.9% of the time on existing business problems. There are certainly many other ones as well as proprietary techniques from particular vendors - but in general the industry is converging to those techniques that work consistently and are understandable and explainable.

3.6 STATISTICS

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users perspective you will be faced with a conscious choice when solving a "data mining" problem as to whether you wish to attack it with statistical methods or other data mining techniques. For this reason it is important to have some idea of how statistical techniques work and how they can be applied.

Data, counting and probability

One thing that is always true about statistics is that there is always data involved, and usually enough data so that the average person cannot keep track of all the data in their heads. This is certainly more true today than it was when the basic ideas of probability and statistics were being formulated and refined early this century. Today people have to deal with up to terabytes of data and have to make sense of it and glean the important patterns from it. Statistics can help greatly in this process by helping to answer several important questions about your data:

- What patterns are there in my database?
- What is the chance that an event will occur?
- Which patterns are significant?
- What is a high level summary of the data that gives me some idea of what is contained in my database?

Certainly statistics can do more than answer these questions but for most people today these are the questions that statistics can help answer. Consider for example that a large part of statistics is concerned with summarizing data, and more often than not, this summarization has to do with counting. One of the great values of statistics is in presenting a high level view of the database that provides some useful information without requiring every record to be understood in detail. This aspect of statistics is the part that people run into every day when they read the daily newspaper and see, for example, a pie chart reporting the number of US citizens of different eye colors, or the average number of annual doctor visits for people of different ages. Statistics at this level is used in the reporting of important information from which people may be able to make useful decisions. There are many different parts of statistics but the idea of collecting data and counting it is often at the base of even these more sophisticated techniques. The first step then in understanding statistics is to understand how the data is collected into a higher level form - one of the most notable ways of doing this is with the histogram.

3.7 NEAREST NEIGHBOUR

Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining. Most people have an intuition that they understand what clustering is - namely that like records are grouped or clustered together. Nearest neighbor is a prediction technique that is quite similar to clustering - its essence is that in order to predict what a prediction value is in one record look for records with similar predictor values in the historical database and use the prediction value from the record that it "nearest" to the unclassified record.

3.8 CLUSTERING

Clustering for Clarity

Clustering is the method by which like records are grouped together. Usually this is done to give the end user a high level view of what is going on in the database. Clustering is sometimes used to mean segmentation - which most marketing people will tell you is useful for coming up with a birds eye view of the business. Two of these clustering systems are the PRIZM™ system from Claritas corporation and MicroVision™ from Equifax corporation. These companies have grouped the population by demographic information into segments that they believe are useful for direct marketing and sales. To build these groupings they use information such as income, age, occupation, housing and race collect in the US Census. Then they assign memorable "nicknames" to the clusters. Some examples are shown in Table 1.2.

Name	Income	Age	Education	Vendor
Blue Blood Estates	Wealthy	35-54	College	Claritas Prizm™
Shotguns and Pickups	Middle	35-64	High School	Claritas Prizm™
Southside City	Poor	Mix	Grade School	Claritas Prizm™
Living Off the Land	Middle-Poor	School Age Families	Low	Equifax MicroVision™
University USA	Very low	Young - Mix	Medium to High	Equifax MicroVision™
Sunset Years	Medium	Seniors	Medium	Equifax MicroVision™

Table 1.2 Some Commercially Available Cluster Tags

This clustering information is then used by the end user to tag the customers in their database. Once this is done the business user can get a quick high level view of what is happening within the cluster. Once the business user has worked with these codes for some time they also begin to build intuitions about how these different customers clusters will react to the marketing offers particular to their business. For instance some of these clusters may relate to their business and some of them may not. But given that their competition may well be using these same clusters to structure their business and marketing offers it is important to be aware of how you customer base behaves in regard to these clusters.

4 APPLICATION OF DATA MINING

There are a number of applications that data mining has. The first is called market segmentation. With market segmentation, you will be able to find behaviours that are common among your customers. You can look for patterns among customers that seem to purchase the same products at the same time. Another application of data mining is called customer churn. Customer churn will allow you to estimate which customers are the most likely to stop purchasing your products or services and go to one of your competitors. In addition to this, a company can use data mining to find out which purchases are the most likely to be fraudulent.

For example, by using data mining a retail store may be able to determine which products are stolen the most. By finding out which products are stolen the most, steps can be taken to protect those products and detect those who are stealing them. While direct mail marketing is an older technique that has been used for many years, companies who combine it with data mining can experience fantastic results.

For example, you can use data mining to find out which customers will respond favorably to a direct mail marketing strategy. You can also use data mining to determine the effectiveness of interactive marketing. Some of your customers will be more likely to purchase your products online than offline and you must identify them.

While many businesses use data mining to help increase their profits, many of them don't realize that it can be used to create new businesses and industries. One industry that can be created by data mining is the automatic prediction of both behaviors and trends. Imagine for a moment that you were the owner of a fashion company, and you were able to precisely predict the next big fashion trend based on the behavior and shopping patterns of your customers? It is easy to see that you could become very wealthy within a short period of time. You would have an advantage over your competitors. Instead of simply guessing what the next big trend will be, you will determine it based on statistics, patterns, and logic.

Another example of automatic prediction is to use data mining to look at your past marketing strategies. Which one worked the best? Why did it work the best? Who were the customers that responded most favorably to it? Data mining will allow you to answer these questions, and once you have the answers, you will be able to avoid making any mistakes that you made in your previous marketing campaign.

Data mining can allow you to become better at what you do. It is also a powerful tool for those who deal with finances. A financial institution such as a bank can predict the number of defaults that will occur among their customers within a given period of time, and they can also predict the amount of fraud that will occur as well.

Another potential application of data mining is the automatic recognition of patterns that were not previously known. Imagine if you had a tool that could automatically search your database to look for patterns which are hidden. If you had access to this technology, you would be able to find relationships that could allow you to make strategic decisions.

Because your decisions are based on logic, you would increase the chances of being successful. While data mining is a very valuable tool, it is important to realize that it is not a panacea. Even if an automated technology should be invented, it will not guarantee the success of you or your company. However, it will tip the odds in your favor.

Practical Applications of Data Mining :Mining Object, Spatial, Multimedia, Text, and web media. An important feature of object-relational and object-oriented databases is their capability of storing, accessing, and modelling complex structure-valued data, such as set- and list-valued data and data with nested structures. A set-valued attribute may be of homogeneous or heterogeneous type. Typically, set-valued data can be generalized by Generalization of each value in the set to its corresponding higher-level concept Generalization of a set-valued attribute. Suppose that the expertise of a person is a set-valued attribute containing the set of values {tennis, hockey, NFS, violin, prince of Persia}. This set can be generalized to a set of high-level concepts, such as {sports, music, computer games} or into the number 5 (i.e., the number of activities in the set). Moreover, a count can be associated with a generalized value to indicate how many elements are generalized to that value, as in {sports(3), music(1), computer games(1)}, where sports(3) indicates three kinds of sports, and so on.

Aggregation and Approximation in Spatial and Multimedia Data GeneralizationAggregation and approximation are another important means of generalization. They are especially useful for generalizing attributes with large sets of values, complex structures, and spatial or multimedia data.

Example: Spatial aggregation and approximation. Suppose that we have different pieces of land for various purposes of agricultural usage, such as the planting of vegetables, grains, and fruits. These pieces can be merged or aggregated into one large piece of agricultural land by a spatial merge. However,

such a piece of agricultural land may contain highways, houses, and small stores. If the majority of the land is used for agriculture, the scattered regions for other purposes can be ignored, and the whole region can be claimed as an agricultural area by approximation.

Generalization of Object Identifiers and Class/Subclass HierarchiesAn object identifier can be generalized as follows. First, the object identifier is generalized to the identifier of the lowest subclass to which the object belongs. The identifier of this subclass can then, in turn, be generalized to a higherlevel class/subclass identifier by climbing up the class/subclass hierarchy. Similarly, a class or a subclass can be generalized to its corresponding superclass(es) by climbing up its associated class/subclass hierarchy.

Some successful application areas include:

- A pharmaceutical company can analyze its recent sales force activity and their results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.

- A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.

- A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using data mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database such as those provided by Dun & Bradstreet can yield a prioritized list of prospects by region.

- A large consumer package goods company can apply data mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

5 ADVANTAGES AND DISADVANTAGES OF DATA MINING

5.1 ADVANTAGES OF DATA MINING

5.1.1 Marking/Retailing

Data mining can aid direct marketers by providing them with useful and accurate trends about their customers' purchasing behavior. Based on these trends, marketers can direct their marketing attentions to their customers with more precision. For example, marketers of a software company may advertise about their new software to consumers who have a lot of software purchasing history. In addition, data mining may also help marketers in predicting which products their customers may be interested in buying. Through this prediction, marketers can surprise their customers and make the customer's shopping experience becomes a pleasant one.

Retail stores can also benefit from data mining in similar ways. For example, through the trends provide by data mining, the store managers can arrange shelves, stock certain items, or provide a certain discount that will attract their customers.

5.1.2 Banking/Crediting

Data mining can assist financial institutions in areas such as credit reporting and loan information. For example, by examining previous customers with similar attributes, a bank can estimated the level of risk associated with each given loan. In addition, data mining can also assist credit card issuers in detecting potentially fraudulent credit card transaction. Although the data mining technique is not a 100% accurate in its prediction about fraudulent charges, it does help the credit card issuers reduce their losses.

5.1.3 Law enforcement

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

5.1.4 Researchers

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing them more time to work on other projects.

5.2 DISADVANTAGES OF DATA MINING

5.2.1 Privacy Issues

Personal privacy has always been a major concern in this country. In recent years, with the widespread use of Internet, the concerns about privacy have increase tremendously. Because of the privacy issues, some people do not shop on Internet. They are afraid that somebody may have access to their personal information and then use that information in an unethical way; thus causing them harm.

Although it is against the law to sell or trade personal information between different organizations, selling personal information have occurred. For example, according to Washing Post, in 1998, CVS had sold their patient's prescription purchases to a different company.7 In addition, American Express also sold their customers' credit care purchases to another company. What CVS and American Express did clearly violate privacy law because they were selling personal information without the consent of their customers. The selling of personal information may also bring harm to these customers because you do not know what the other companies are planning to do with the personal information that they have purchased.

5.2.2 Security issues

Although companies have a lot of personal information about us available online, they do not have sufficient security systems in place to protect that information. For example, recently the Ford Motor credit company had to inform 13,000 of the consumers that their personal information including Social Security number, address, account number and payment history were accessed by hackers who broke into a database belonging to the Experian credit reporting agency.9 This incidence illustrated that companies are willing to disclose and share your personal information, but they are not taking care of the information properly. With so much personal information available, identity theft could become a real problem.

5.2.3 Misuse of information/inaccurate information

Trends obtain through data mining intended to be used for marketing purpose or for some other ethical purposes, may be misused. Unethical businesses or people may used the information obtained through data mining to take advantage of vulnerable people or discriminated against a certain group of people. In addition, data mining technique is not a 100 percent accurate; thus mistakes do happen which can have serious consequence.

6 CONCLUSION AND FUTURE WORKS

With the increase of economic globalization and evolution of information technology, financial data are being generated and accumulated at an unprecedented pace. As a result, there has been a critical need for automated approaches to effective and efficient utilization of massive amount of financial data to support companies and individuals in strategic planning and investment decision making. Data mining techniques have been used to uncover hidden patterns and predict future trends and behaviors in financial markets. The competitive advantages achieved by data mining include increased revenue, reduced cost, and much improved marketplace responsiveness and awareness. There has been a large body of research and practice focusing on exploring data mining techniques to solve financial problems. There are many contributions on this field. Study of implementing data mining approaches and integrating them into stock market research on Tehran stock market is an example for future research and implementation. Another future extension of this research involves incorporating some other variables into the criteria for data mining techniques. These include new variables that reflect future information and those that reflect the impacts of other stock markets to the market of concern.

The research reviewed in this paper has mainly concentrated on applications of the algorithms. The quality of the data and data preparation issues, particularly relating to financial databases has not been discussed. Major effort is needed in the data preparation process, as this is often simply based on practitioner's instinct and experience. A more generic process for data cleaning is essential to enable the growth of data mining in financial market. The stock market data mining research often does not consider the quality of the rules or knowledge discovered. The knowledge generated is sometimes cumbersome and the relationships obtained are too complex to understand. Future research effort is therefore also needed to enhance the expressiveness of the knowledge. Further research is needed to develop generic guidelines for a variety of different data and types of problems, which are commonly faced by financial markets. To be successful, a data mining project should be driven by the application needs and results should be tested quickly. Financial applications provide a unique environment where efficiency of the methods can be tested instantly, not only by using traditional training and testing data but making real stock forecast and testing it the same day. This process can be repeated daily for several months collecting quality estimates. This paper states problems of data mining in finance (stock market) and specific requirements for data mining methods including in making interpretations, incorporating relations and probabilistic learning. The data mining techniques outlined in this paper advances pattern discovery methods that deals with complex numeric and non-numeric data, involving structured objects, text and data in a variety of discrete and continuous scales (nominal, order, absolute and so on). Also, this paper shows benefits of using such techniques for stock market forecast. Currently the success of data mining exercises has been reported in literature extensively.

REFERENCES

[1] Agrawal R, Imilienski T, Swami A (1993). Mining association rules between sets of items in large databases, In Proceedings of the ACM SIGMOD international conference on management of data.
[2] Basaltoa N, Bellottib R, De Carlob F, Facchib P, Pascazio S (2005). Clustering stock market companies via chaotic map synchronization, Physica A.
[3] Berry MJA, Linoff GS (2000). Mastering data mining, New York: Wiley.
[4] Boris K, Evgenii V (2005). Data Mining for Financial Applications, the
[5] Data Mining and Knowledge Discovery Handbook.
[6] www.the-data-mine.com
[7] www.dataminingtools.net