Data Archeology "Knowledge discovery in Databases"

Mrs. Jasmeet Kaur[1] and Mrs.Jigisha Pandya[2]

[1,2]Lecturer, G.H Raisoni College of Engineering and Management, Pune

ABSTRACT

In the present era challenging problems cannot be solved in a reasonable amount of time with conventional computers. While grand challenge problems can be found in many domains like, computing software changes, science applications etc.

These problems often require numerous complex calculations and collaboration among people with multiple disciplines and geographic locations. Many grand challenge problems involve the analysis of very large volumes of data. Data mining, also popularly known as Knowledge Discovery in Databases (KDD) is a well established field of computer science concerned with the automated search of large volumes of data for patterns that can be considered knowledge about the data.

Data Archeology offers automated discovery of previously unknown patterns as well as automated prediction of trends and behaviors; its technologies are complimentary to existing decision support tools and provide the business analyst and marketing professional with a new way of analyzing the business. After a general introduction of the knowledge discovery, this paper concludes the applications of data mining and steps, stages, techniques of knowledge discovery have been presented.

Key words: Bayesian, data mining, Exploration, Knowledge

Introduction

Knowledge Discovery in Databases (KDD), also popularly known as Data mining, while data mining and knowledge discovery in data bases are frequently treated as synonyms, data is actually part of the knowledge discovery process.

It is the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. However, really knowledge discovery in Databases turns databases into knowledge bases which is one of the fundamental components of expert systems. Instead of the computer just blindly pulling data from a database, the computer is able to take all the data and interpret it, which is a huge step to make. If it was not for existing AI technologies this field could not have emerged as quickly; if at all. It is continuing to receive enormous attention by both commercial and scientific communities to receive enormous attention by both commercial and scientific communities for three reasons

1. Both the number and size of databases in much organization are growing at a staggering rate terabyte databases once unthinkable are now becoming a reality in verity of domains including marketing, finance, sales, health care, earth science and various government applications.
2. Organizations have revised that there is valuable knowledge which is buried in the data which if discovered could provide those organizations with competitive advantage.
3. Some of the enabling technologies have only recently become mature enough to make data mining possible on large data sets.

Applications of KDD

Knowledge Discovery in Databases concept is now a days applicable in every field. Applications of KDD approach are 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.

1. 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.
2. 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.
3. The insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring).

4. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together.
5. Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor.
6. 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. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use Knowledge Discovery in Databases of chemical compounds and genetic material to help guide research on new treatments for diseases.

In the public sector, KDD applications were initially used as a means to detect fraud and waste, but they have grown also to be used for purposes such as measuring and improving program performance.

Knowledge Discovery in Databases can be used to detect unusual patterns, terrorist activities and fraudulent behavior. A critical application of data mining technology is counterterrorism-the development of countermeasures to threats occurring from the terrorist activities.

Methodology

KDD or Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of KDD or data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications.

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important. If not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision making.

The process of Knowledge Discovery in Databases consists of three stages:

Stage 1: Exploration.

This stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records and - in case of data sets with large numbers of variables ("fields") - performing some preliminary feature selection operations to bring the number of variables to a manageable range (depending on the statistical methods which are being considered). Then, depending on the nature of the analytic problem, this first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and/or the general nature of models that can be taken into account in the next stage.

Stage 2: Model building and validation.

This stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples). This may sound like a simple operation, but in fact, it sometimes involves a very elaborate process. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best

Stage 3: Deployment.

That final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

Knowledge Discovery in Databases refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.

Knowledge discovery process (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process generalizes to non database sources of data, although it emphasizes databases as a primary source of data. It consists of many steps (one of them is DM), each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain.

Following steps shows data mining as a step in an iterative knowledge discovery process.

1. *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.
2. *Data Integration:* At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. *Data Selection :* At this step, the data relevant to the analysis is decided on and retrieved from the data collection
4. *Data Transformation:* It is also known as data consolidation, it is phase in which the selected data is transformed into forms appropriate for the mining procedure.
5. *Data Mining:* It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6. *Pattern evaluation:* In this step, strictly interesting patterns representing knowledge are identified based on given measures.

KDD Techniques

There are many different approaches that are classified as KDD techniques. There are quantitative approaches, such as the probabilistic and statistical approaches. There are approaches that utilize visualization techniques. There are classification approaches such as Bayesian classification, inductive logic, data cleaning/pattern discovery, and decision tree analysis. Other approaches include deviation and trend analysis, genetic algorithms, neural networks, and hybrid approaches that combine two or more techniques.

(i) Probabilistic Approach

This family of KDD techniques utilizes graphical representation models to compare different knowledge representations. These models are based on probabilities and data independencies. They are useful for applications involving uncertainty and applications structured such that a probability may be assigned to each “outcome" or bit of discovered knowledge. Probabilistic techniques may be used in diagnostic systems and in planning and control systems[1]. Automated probabilistic tools are available both commercially and in the public domain.

(ii) Statistical Approach

The statistical approach uses rule discovery and is based on data relationships. An ``inductive learning algorithm can automatically select useful join paths and attributes to construct rules from a database with many relations" [2]. This type of induction is used to generalize patterns in the data and to construct rules from the noted patterns. Online analytical processing (OLAP) is an example of a statistically-oriented approach. Automated statistical tools are available both commercially and in the public domain.

An example of a statistical application is determining that all transactions in a sales database that start with a specified transaction code are cash sales. The system would note that of all the transactions in the database only 60% are cash sales. Therefore, the system may accurately conclude that 40% are collectibles.

(iii) Classification Approach

Classification is probably the oldest and most widely-used of all the KDD approaches [3]. This approach groups data according to similarities or classes. There are many types of classification techniques and numerous automated tools available.

(iv) Bayesian Approach

The Bayesian Approach to KDD ``is a graphical model that uses directed arcs exclusively to form an [sic] directed acyclic graph" [1]. Although the Bayesian approach uses probabilities and a graphical means of representation, it is also considered a type of classification.

Bayesian networks are typically used when the uncertainty associated with an outcome can be expressed in terms of a probability. This approach relies on encoded domain knowledge and has been used for diagnostic systems. Other pattern recognition applications, including the Hidden Markov Model, can be modeled using a Bayesian approach [4]. Automated tools are available both commercially and in the public domain.

(v) Pattern Discovery and Data Cleaning

Pattern Discovery and Data Cleaning is another type of classification that systematically reduces a large database to a few pertinent and informative records [5]. If redundant and uninteresting data is eliminated, the task of discovering patterns in the data is simplified. This approach works on the premise of the old adage, ``less is more". The pattern discovery and data cleaning techniques are useful for reducing enormous volumes of application data, such as those encountered when analyzing automated sensor recordings. Once the sensor readings are reduced to a manageable size using a data cleaning technique, the patterns in the data may be more easily recognized. Automated tools using these techniques are available both commercially and in the public domain.

(vi) Decision Tree Approach

The Decision Tree Approach uses production rules, builds a directed a cyclical graph based on data premises, and classifies data according to its attributes. This method requires that data classes are discrete and predefined [3]. According to the primary use of this approach is for predictive models that may be appropriate for either classification or regression techniques. Tools for decision tree analysis are available commercially and in the public domain

Knowledge Representation

It is the final phase in which the discovered knowledge is visually represented to the user; the essential step uses visualization techniques to help user 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 evolution 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.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is however, a misnomer, since mining for gold in rocks usually called “good mining” and not “rock mining”, thus by analogy, data mining should have been called “knowledge mining” instead Nevertheless, data mining became the accepted customary term, and very rapidly a trend the even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are data dredging, knowledge extraction and pattern discovery.

Conclusion

1. It is concluded that large amounts of data are stored in repositories, using techniques like pattern recognition and mathematical techniques
2. The number and size of databases in many organizations are growing at a staggering rate. Terabyte databases are now becoming a reality in a variety of domains such as marketing, sales, finance, healthcare, earth science, molecular biology.
3. There is valuable knowledge, which is buried in the data, which, if discovered, could be advantageous.
4. A few of the technologies have recently become capable enough to make data mining possible on larger datasets.

Results

The concept of KDD or *Data Mining (popularly called)* is becoming increasingly popular as a business information management tool where it is expected to reveal knowledge structures that can guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques specifically designed to address the issues relevant to business Knowledge Discovery in Databases, but KDD is still based on the conceptual principles of statistics including the traditional Exploratory Data Analysis (EDA) and modeling and it shares with them both some components of its general approaches and specific techniques.

References

Buntine, W., (1996), Graphical Models for Discovering Knowledge, *Advances in Knowledge Discovery and Data Mining*, 59-82.

Buntine, W,(1996),A Guide To The Literature On Learning Probabilistic Networks From Data, *IEEE Transactions on Knowledge and Data Engineering* 8(2), 195-210

Guyon, I., Matic, N., Vapnik, V., (1996), Discovering Informative Patterns and Data Cleaning, Advances, 181-203.

Hsu, C.N., Knoblock, C.A., (1996), Using Inductive Learning to Generate Rules for Semantic Query Optimization, *Advances in Knowledge Discovery and Data Mining*, 425-445.

Quinlan, J.R.,(1993),*Programs For Machine Learning*,4(5).