Available at: www.researchpublications.org

Knowledge Discovery in Databases using Data Mining

Ms. Meghana A. Deshmukh
Dept. of Computer Science and Engineering
Prof. Ram Meghe Institute of Technology and Research
Badnera,Amravati.
meghnadeshmukh9@gmail.com

Prof. R. R. Tuteja
Dept. of Computer Science and Engineering
Prof. Ram Meghe Institute of Technology and Research
Badnera,Amravati.
ramu.tuteja@gmail.com

Abstract

The paper is about the mining of data and finding essential information from large amounts of data. Extracting the knowledge from huge amounts of data is known as Data Mining. Use of algorithms to extract the information and patterns is derived by the KDD process. A process of finding useful information and patterns in data is Knowledge Discovery in Databases. Research in data mining continues growing in business and in learning organization over coming decades. Knowledge Discovery and Data Mining are powerful automated data analysis tools and they are predicted to become the most frequently used analytical tools in the near future.

Keywords
Data Mining1, Knowledge Discovery in Databases (KDD)2.

I. INTRODUCTION

Knowledge Discovery and Data Mining are powerful automated data analysis tools and they are predicted to become the most frequently used analytical tools in the near future [3]. Knowledge discovery and data mining are the landmarks of the information age. Acquiring, storing, and understanding data have posed great challenges and brought a lot of promises. Knowledge discovery in databases (KDD) and data mining (DM) have emerged as high profile, rapidly evolving, badly needed, conceptually advanced, and practically important areas.

Databases can contain vast quantities of data describing decisions, performance and operations. In many cases the database contains critical information concerning past business performance which could be used to predict the future. Data mining (also known as Knowledge Discovery) technology helps businesses discover hidden data patterns and provides predictive information which can be applied to benefit the business. The basic approach is to access a database of historical data and to identify relationships which have a bearing on a specific issue, and then extrapolate from these relationships to predict future performance or behaviour.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data

Prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

II. TECHNICAL BACKGROUND

Data Mining and Knowledge Discovery in Databases are terms used interchangeably. Other terms often used are data or information harvesting, data archeology, functional dependency analysis, knowledge extraction and data pattern analysis. A high level definition of Data Mining is: the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining is not a simple process and there is no tool that can do the job automatically. Data mining can be aided by tools, but it requires both human data mining expertise and human domain expertise. Data mining consists of a number of operations, each of which are supported by a variety of technologies, such as rule induction, neural networks, conceptual clustering [4].

There is an immense diversity of current research on knowledge discovery in databases.

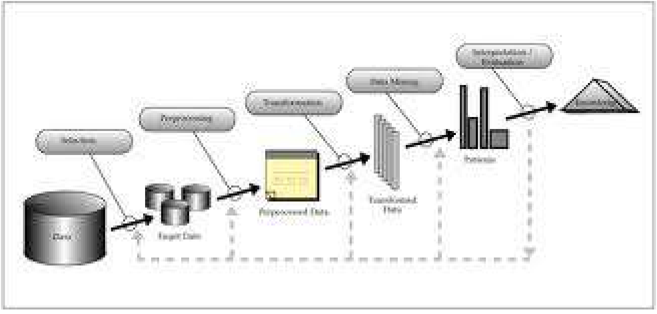
III. Data mining and Knowledge discovery in database

A. Knowledge discovery in database:

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. *Knowledge Discovery in Databases* brings together current research on the exciting problem of discovering useful and interesting knowledge in databases [6].

.

1) **KDD process**:



The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the interactive nature of the process.

Steps are:

- Data cleaning: 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: 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: 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 [6].

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.

2) **Current KDD Applications**

- Science - SKYCAT: used to aid astronomers by classifying faint sky objects
- Marketing - AMEX: used customer group identification and forecasting. Claims 10%-15% increase in card usage.
- Investment - Many use. Few tell. - LBS Capital Management: uses and expert system/neural network to manage $600 million portfolio. Results outperform market.
- Fraud Detection - HNC Falcon, Nestor Prism: credit card fraud detection - FAIS: US Treasury money-laundering detection system
- Manufacturing - CASSIOPEE: a trouble-shooting system used in Europe to diagnose 737 problems by deriving families of faults by clustering
- Telecommunications - TASA (Telecommunications Alarm-Sequence Analyzer): locates patterns of frequently occurring alarm episodes and represents the patterns as rules
- Data Cleaning - MERGE-PURGE: used by Washington State to locate and remove duplicate welfare claims
- Sports - ADVANCED SCOUT: helps NBA coaches analyze data to organize and interpret game data ==> player selection and team management
- Information Retrieval - Intelligent Agents have been designed to navigate the internet and return information pertinent to some non-trivial query +

3) **Advantages of KDD**

- Merges machine learning, pattern recognition, statistics, database, high performance computing with unified goal of extracting high-level knowledge from low-level data in the context of large datasets.
- Differs from much of ML, etc. in that it places special emphasis on finding understandable patterns that can be interpreted as useful or interesting knowledge.
- Fundamentally a statistical endeavor. Statistics provide a language and framework for quantifying the uncertainty that results when one tries to infer general patterns from a particular sample of an overall population.

- Because data-mining algorithms typically assume data are in main memory, KDD relies on database techniques for gaining efficient data access to large datasets.

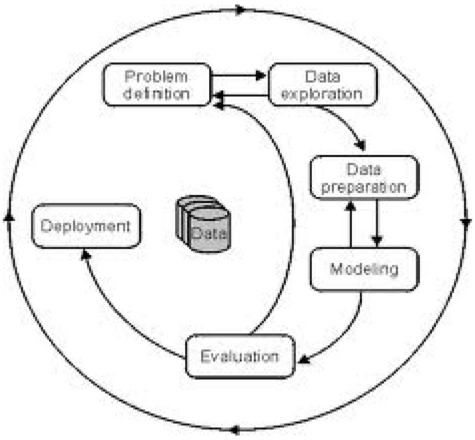
- A set of principles from the database field for dealing with large datasets is OLAP, Online Analytical Processing. OLAP tools focus on simplifying and supporting interactive data analysis; the goal of KDD tools is to automate as much of the process as possible.

B. Data mining

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 is usually called "gold 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 that 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.

1) **Working of data mining**:

Data mining is an iterative process that typically involves the following phases:



Problem definition: A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required.

Data exploration: Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

Data preparation: Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

Modeling: Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

Evaluation: Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions: 1. Does the model achieve the business objective? 2. Have all business issues been considered? At the end of the evaluation phase, the data mining experts decide how to use the data mining results. Deployment Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets [1].

2) **Issues in Data Mining**

Security and social issues: Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behaviour understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

User interface issues: The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

Mining methodology issues: These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs, the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently.

Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the accuracy of the results. As a consequence

, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information.

More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the *curse of dimensionality*. This "curse" affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

Performance issues: Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to reanalyze the complete dataset.

Data source issues: Verious issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community [6].

IV. Research and Application Challenges:

We outline some of the current primary research and application challenges for KDD. This list is by no means exhaustive and is intended to give the reader a feel for the types of problem that KDD practitioners wrestle with.

Larger databases: Databases with hundreds of fields and tables and millions of records and of a multigigabyte size are commonplace, and terabyte (1012 bytes) databases are beginning to appear. Methods for dealing with large data volumes include more efficient algorithms, sampling, approximation, and massively parallel processing.

High dimensionality: Not only is there often a large number of records in the database, but there can also be a large number of fields (attributes, variables); so, the dimensionality of the problem is high. A high-dimensional data set creates problems in terms of increasing the size of the search space for model induction in a combinatorially explosive manner. In addition, it increases the chances that a data-mining algorithm will find spurious patterns that are not valid in general. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables.

Overfitting: When the algorithm searches for the best parameters for one particular model using a limited set of data, it can mod- el not only the general patterns in the data but also any noise specific to the data set, resulting in poor performance of the model on test data. Possible solutions include crossvalidation, regularization, and other sophisticated statistical strategies.

Assessing of statistical significance: A problem (related to overfitting) occurs when the system is searching over many possible models. For example, if a system tests models at the 0.001 significance level, then on aver- age, with purely random data, N/1000 of these models will be accepted as significant.

V. KDD vs Data-Mining

- KDD and Data minig are not the same thing.

KDD: The nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [Fayyad, et al]

KDD is the overall process of discovering useful knowledge from data.

Data mining: An application of specific algorithms for extracting patterns from data.

Data mining is a step in the KDD process [5].

Advantages:

1. The analysis and dependency of any variable can be done.

2. The clustering of finite sets if data is the biggest advantage of data mining [1].

Disadvantages:

1. It has security and privacy issues.

2. It is not accurate.

VI. Conclusion

Building models is only one step in knowledge discovery. It’s vital to properly collect and prepare the data, and to check the models against the real world. The “best” model is often found after building models of several different types.

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. However, data mining tools need to be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of applications, from improving revenues to reducing costs.

References

[1] “Data Mining, Applications and Knowledge Discovery” International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth “From Data Mining to Knowledge Discovery in Databases”.

[3] “On the Impact of Knowledge Discovery and Data Mining” crpit.com/confpapers/CRPITV1Wahlstrom.pdf.

[4] http://www.aiai.ed.ac.uk/links/dm.html#intro.

[5] seclab.cs.ucdavis.edu/projects/misuse/meetings/KDD.html.KDD Overview Notes.

[6] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/."Introduction to Data Mining".

[7] Agrawal, R., and Psaila, G. 1995. Active Data Min- ing. In Proceedings of the First International Con- ference on Knowledge Discovery and Data Mining (KDD-95), 3–8. Menlo Park, Calif.: American Asso- ciation for Artificial Intelligence.

[8] "Introduction to Data Mining and Knowledge Discovery" by Two Crows Corporation. www.twocrows.com/intro-dm.pdf

[9] "The KDD Process for Extracting Useful". Knowledge from Volumes of Data. shawndra.pbworks.com/.../...