Ethics in Knowledge Discovery consisting Large Databases

Simmi Bagga[1]*, Satinder Kaur[2]

[1]Assistant Professor, Sant Hira Dass Kanya Maha Vidyalaya, Kala Sanghian, Kapurthala, India
*Corresponding author E-mail: simmibagga12@gmail.com

[2]Assistant Professor, GNDU, RC, Sathiala, India

Abstract: Today, we are living in a large ocean of big data i.e. data which is voluminous, is created and communicated at very large velocity as well as having large variety of formats. There is a big problem to extract right knowledge from it. However, Knowledge discovery in databases provide a solution through data mining techniques but still there are major issues involved before planning a KDD process. This paper introduces these issues so that before implementing a systematic KDD, all aspects should be taken into mind**.**

Keywords: KDD, Data mining, Clustering, Privacy & Security.

I. INTRODUCTION

Knowledge discovery in databases refers to the broad process of finding knowledge from datum. The goal of the KDD process is to extract knowledge from data in the context of large databases. The knowledge discovery process is repetitive, interactive, and consists of nine steps. The process starts with determining the KDD goals and ends with the implementation of the discovered knowledge. KDD is a growing field. There are many knowledge discovery methodologies, some are in use and while few are under development. Some of these techniques are generic, while others are domain-specific. KDD provides the capability to discover new and meaningful information by using existing data. KDD quickly exceeds the human capacity to analyze large data sets. The amount of data that requires processing and analysis in a large database exceeds human capabilities, and the difficulty of accurately transforming raw data into knowledge surpasses the limits of traditional databases. Therefore, the full utilization of stored data depends on the use of knowledge discovery techniques.

Knowledge discovery is defined as ``the non-trivial extraction of implicit, unknown, and potentially useful information from data" [6]. This paper first introduces the common characteristics of each KDD process and then discuss how data mining is related to it. The next section introduces a systematic framework for a KDD process which comprises nine steps. Finally, ethics involved in KDD are discussed so that one can attain full potential through this technique.

A. Data Mining and KDD

Data Mining is main concerned with the analysis of data. Data Mining tools and techniques are used for finding patterns from the data set automatically with minimal user input and efforts. Data Mining tools and techniques can be successfully applied in various fields in various forms. Many organizations now start using Data Mining as a tool, to deal with the competitive environment for data analysis. By using Mining tools and techniques, various fields of business get benefit by easily evaluate various trends and pattern of market and to produce quick and effective market trend analysis. Data Mining is a powerful tool capable of handling decision making and for forecasting future trends of market. Data Mining applications are widely used in Health industry, Auditing, Telecommunication industry, Retail industry etc [11].

B. Characteristics of KDD

Although there are many approaches to KDD, six common and essential elements qualify each as a knowledge discovery technique. The following are basic features that all KDD techniques [5], [6]:

- All approaches deal with large amounts of data
- Efficiency is required due to volume of data
- Accuracy is an essential element
- All require the use of a high-level language
- All approaches use some form of automated learning
- All produce some interesting results

Data Mining can be used as a synonym for KDD but actually Data Mining is one of the main steps of KDD process [5]. Under their conventions, the knowledge discovery process takes the raw results from data mining (the process of extracting trends or patterns from data) and then accurately transforms them into useful and understandable information. Data Mining discovers patterns in a data set previously prepared in a specific way. Data Mining can be performed on data represented in qualitative, textual, multimedia form or any other form of data. Data mining tasks follow a traditional, hypothesis-driven data analysis approach, it is common place to employ an opportunistic, data driven approach that encourages the pattern detection algorithms to find useful trends, and relationships. Essentially, the two types of data mining approaches differ in whether they seek to build models or to find patterns. The first approach, concerned with building models is, apart from the large sizes of the data sets, similar to conventional exploratory statistical methods. The objective is the produce an overall summary of a set of data to identify and describe the main features of the shape of the distribution. The second type of data mining approach, pattern detection, seeks to identify small departures from the norm, to detect unusual patterns of behavior. In general, business databases pose a unique problem for pattern extraction because of their complexity.

II. NEED FOR DATA MINING

The motivation behind data mining whether commercial or scientific is the same. The need to find the accurate and perfect information for true decision-making is a great challenge. In science and engineering domains the size of data is the only reason why data mining techniques are gained popularity. Science data in the areas such as remote sensing, astronomy and computer simulation is routinely being measured in terabytes and petabytes. However what makes the analysis of these data set challenges is not just the size but also the complexity of data. Data mining applications are widely used in direct marketing, health industry, e-commerce, customer relationship management (CRM), telecommunication industry and financial sector. Data mining is available in various forms like text mining, web mining, audio & video data mining, pictorial data mining, relational databases, and social networks data mining.

A new concept of Business Intelligence data mining has evolved now, which is widely used by leading corporate houses to stay ahead of their competitors. Business Intelligence (BI) can help in providing latest information and used for competition analysis, market research, economical trends, consume behavior, industry research, geographical information analysis and so on [5]. Business Intelligence Data Mining helps in decision-making.

Data explosion is the one of the main recent problem [4]. These days automated data collection tools and mature data base technology leads to tremendous amount of data stored in the database, data warehouse and other information repositories. This means that we are drowning in data bus starving for knowledge. The solution to this problem is to extracting the interesting knowledge from the data by applying rules, regularities, patterns and constraints in the large data set. One of another problem like incomplete, noisy and inconsistent data are common when database size is large. There are several reason for these problems. Preprocessing allows the user to transform databases to a format that transform the selected data into a suitable shape to be used by particular algorithm. Usually in data mining tools the transformation perform in recasting of data or the calculation of new attributes from the existing ones. These are the main motivations of the data mining.

III. SYSTEMATIC FRAMEWORK FOR KDD

Knowledge Discovery in Databases (KDD) is a non-trivial, multi-step process. The KDD process is interactive and iterative, involving various steps with many decisions being made by the user. 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. The process consists of data integration, preparation and transformation, data mining as well as evaluation and presentation of the data mining results.

Knowledge is a pattern that is sufficiently interesting to the user .The user specifies the measure of interestingness and the certainty criterion. Discovered knowledge is the output from a program that analyzes a data set and generates patterns. A pattern's certainty is the measure of confidence in discovered knowledge represented by the pattern. Pattern's certainty is higher if data in the data set considered are good representatives of data in the database, if they contain little or no noise at all, if they are valid, reliable, complete, precise, and contain no contradictions [9].

A. *Developing and Understanding of Application Domain and identify the goal*

This is the initial step of KDD. This step mainly deals to find the business goals, objectives, critical success factors from organization after spending some time at the place and to sift through the raw data [1]. Then the real goal of the discovery will be found.

B. *Selection and Creating a target data set*

The dataset from a subset of samples or variables on which to make discoveries is selected or created. In the step of data discovery, we have to decide whether quality of data is satisfactory for the goal. If important attributes missing in the dataset the entire process may fail [3]. It includes finding what data is available, obtaining additional data if required and then integrating all the data for the knowledge discovery into one data set. This process is the important step of KDD because the whole process depends on the available data set.

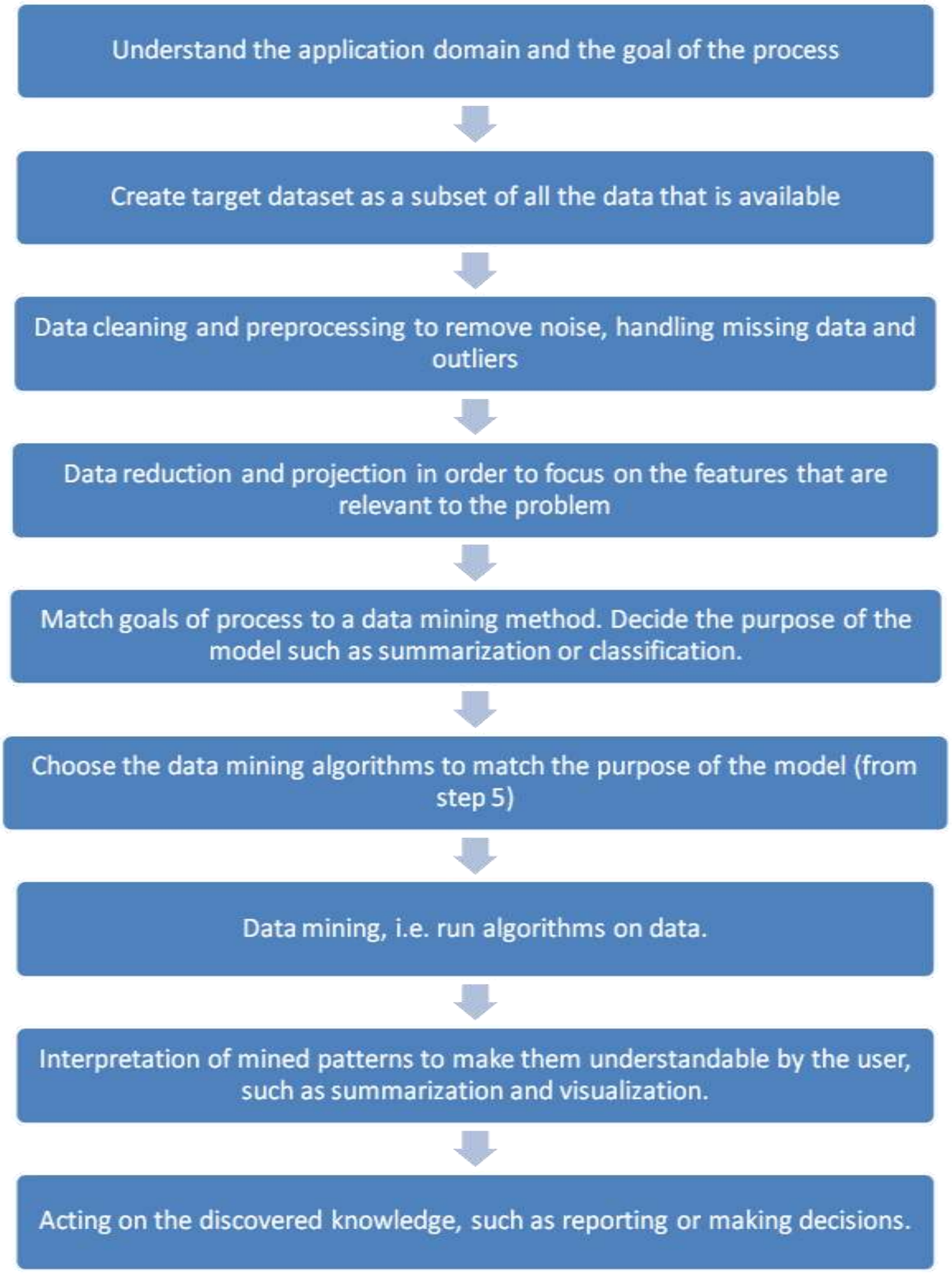
Understand the application domain and the goal of the process
Create target dataset as a subset of all the data that is available
Data cleaning and preprocessing to remove noise, handling missing data and outliers
Data reduction and projection in order to focus on the features that are relevant to the problem
Match goals of process to a data mining method. Decide the purpose of the model such as summarization or classification.
Choose the data mining algorithms to match the purpose of the model (from step 5)
Data mining, i.e. run algorithms on data.
Interpretation of mined patterns to make them understandable by the user, such as summarization and visualization.
Acting on the discovered knowledge, such as reporting or making decisions.

Figure 1 KDD Process

C. *Data Cleaning and Preprocessing*

This stage includes data cleaning by handling missing values and removal of noise and outliers from available datum. If one suspects that a certain attribute are missing or insufficient reliable then collect necessary information to model or account for noise and also deciding the strategies to handle the missing data fields. It may involve complex statistical methods or data mining algorithm [2]. A prediction model for the attributes may be developed, and then missing attributes may be predicted.

D. *Data Transformation*

In this stage, the generation for the better data for the data mining is prepared and developed. It also includes finding useful features to represent the data relative to the goal. Method includes dimensionality reduction/transformation that reduces number of variables. This step can be crucial for the success of KDD process because if we do not use the right transformation in the beginning we may obtain surprising effects.

E. *Selection of appropriate Data Mining Tasks*

After completing the first four steps we decide what type of Data Mining algorithm is to use like summarization, classification, regression, clustering, etc. Regression is often used at the statistical method used for numeric prediction. Clustering is the method by which records are grouped together. This decision may depend on the goal of the KDD process and on the previous steps. There are two major goal of data mining: prediction and description [8]. Prediction refers to the supervised data mining. It involves predictions about future events. Description data mining include unsupervised data mining. It presents information to the user in a human understandable form. Most of the data mining techniques are based on inductive learning, where model is constructed explicitly or implicitly by generalizing the sufficient number of training sets.

F. *Selection of Data Mining Algorithm*

This stage includes decision of which models and parameters may be appropriate match method to goal of KDD process. Two classes of model are available i.e. logical models (purely deterministic) and statistical models (non-deterministic) [6], which are most widely used due to uncertainty in real world data. It also include the selecting the specific method to be used for searching patterns. Meta learning focuses on explaining what causes a data mining to be successful or not in a particular problem. Each algorithm has some parameters of learning [12]. This approach helps to understand the conditions under which a data mining algorithm is more appropriate.

G. *Data Mining*

Finally the implementation of data mining algorithm is reached and it's the main step of whole KDD process [7]. The data mining step of KDD process involves repeated iterative application of particular data mining methods. Data mining involves fitting model to determining pattern from the observed data.

H. *Interpretation and Visualization*

In this step, one interprets the mined patterns with respect to the goal of the KDD. We also visualize of extracted patterns and models and also visualize of the data given the extracted models. The discovered knowledge is also documented for the future use.

I. *Evaluation*

Actually, the success of this step determines the effectiveness of the KDD process. We check inconsistencies with other prior extracted or believed knowledge.

IV. KDD TECHNIQUES

Learning algorithms are an integral part of KDD. Learning techniques may be supervised or unsupervised. In general, supervised learning techniques enjoy a better success rate as defined in terms of usefulness of discovered knowledge. According to [1], learning algorithms are complex and generally considered the hardest part of any KDD technique. Machine discovery is one of the earliest fields that has contributed to KDD [5]. While machine discovery relies solely on an autonomous approach to information discovery, KDD typically combines automated approaches with human interaction to assure accurate, useful, and understandable results.

There are many different approaches that are classified as KDD techniques. There are quantitative approaches, such as the probabilistic [2] and statistical [8] approaches. Online analytical processing (OLAP) is an example of a statistically-oriented approach. Automated statistical tools are available in both commercial as well as in the public domain. There are approaches that utilize visualization techniques. There are classification approaches [10] such as Bayesian classification, inductive logic, data cleaning/pattern discovery [7], and decision tree analysis [11]. Other approaches include deviation and trend analysis [5], genetic algorithms, neural networks [12], and hybrid approaches that combine two or more techniques. For example, the Bayesian approach may be logically grouped with probabilistic approaches, classification approaches, or visualization approaches.

V. ETHICS IN KDD

Main Ethics involved in KDD are:

- ***Large sources of data***
  There are huge volumes of data in the world. The data is too big, moves too fast, or does not fit the structures of existing KDD architectures. In KDD once the goal has been determined, we choose the dataset or subset of samples or variables on which to make discoveries. In the step of data discovery, we have to decide whether quality of data is satisfactory for the goal. If important attributes missing in the dataset the entire process may fail. It includes finding what data is available, obtaining additional data if required and then integrating all the data for the knowledge discovery into one data set. The whole process of KDD depends on the available data set but due to large source of data it is very difficult to choose the data set at which we can apply the KDD process.
- ***Multiple domains***
  The possible domains are direct marketing, health industry, e-commerce, customer relationship management (CRM), telecommunication industry and financial sector. One should select proper domain with full intention.
- ***Different data mining techniques***
  Data mining is available in various forms like text mining, web mining, audio & video data mining, pictorial data mining, relational databases, and social networks data mining.
- ***Need for privacy & security***
  Datasets which are heart of KDD should be too much secure so that one cannot misuse them.
- ***Different presentation formats***
  Data comes in different varieties i.e. in different formats. These include a long list of data such as documents, emails, social media text messages, video, still images, audio, graphs, and the output from all types of machine-generated data from sensors, cell phone GPS signals, DNA analysis devices, and more. This type of data is characterized as unstructured or semi-structured. This variety of unstructured data creates problems for storage, mining and analysing it.

VI. Conclusion And Future Directions

KDD is a rapidly expanding field with promise for great applicability. Knowledge discovery purports to be the new database technology for the coming years. The need for automated discovery tools had caused an explosion in the number and type of tools available commercially and in the public domain. Although KDD process seems to be systematic yet there is an ethic involved in each step, so success can only be achieved through wise decisions and human expertise.

It is anticipated that commercial database systems of the future will include KDD capabilities in the form of intelligent database interfaces. Some types of information retrieval may benefit from the use of KDD techniques. Due to the potential applicability of knowledge discovery in so many diverse areas there are growing research opportunities in this field. Many of these opportunities are discussed in [10].

VII. References

[1] Brachman, R.J., and Anand, T. The Process Of Knowledge Discovery In Databases: A Human-Centered Approach. In Advances In Knowledge Discovery And Data Mining , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 37-57.

[2] Buntine, W. Graphical Models For Discovering Knowledge. In Advances In Knowledge Discovery And Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 59-82.

[3] Buntine, W. "A Guide to The Literature On Learning Probabilistic Networks From Data." IEEE Transactions on Knowledge and Data Engineering 8, 2 (Apr. 1996), 195-210.

[4] Fayyad, U.M., Djorgovski, S.G., and Weir, N. Automating The Analysis And Cataloging Of Sky Surveys In Advances In Knowledge Discovery And Data Mining , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 472-493.

[5] Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. From Data Mining To Knowledge Discovery: An Overview. In Advances In Knowledge Discovery And Data Mining , eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 1-34.

[6] Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C. Knowledge Discovery In Databases: An Overview. In Knowledge Discovery In Databases, eds. G. Piatetsky-Shapiro, and W. J. Frawley , AAAI Press/MIT Press, Cambridge, MA., 1991, pp. 1-30.

[7] Guyon, I., Matic, N., and Vapnik, V. Discovering Informative Patterns And Data Cleaning. In Advances In Knowledge Discovery And Data Mining, eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAI Press/The MIT Press, Menlo Park, CA.1996, pp. 181-203.

[8] Hsu, C.N., and Knoblock, C.A. Using Inductive Learning To Generate Rules For Semantic Query Optimization. In Advances In Knowledge Discovery And Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA., 1996, pp. 425-445.

[9] Piatetsky-Shapiro, G. S*i*ftware: Tools For Data Mining And Knowledge Discovery. World Wide Web URL: http://www.kdnuggets.com/siftware.html

[10] Piatetsky-Shapiro, G., and Beddows, M. Knowledge Discovery Mine -- Data Mining And Knowledge Discovery Resources. World Wide Web URL:http://www.kdnuggets.com/.

[11] Kalavathy R, et al. KDD and data mining. Information and Communication Technology in Electrical Sciences (ICTES 2007), 2007. ICTES. IET-UK IEEE

[12] Ingre B, et al. Performance analysis of NSL-KDD dataset using ANN. International Conference on Signal Processing And Communication Engineering Systems (SPACES), 2015 IEEE