Data Mining : As an imperative tool for Discovering Knowledge

Anoop K. Paharia
M.Tech.(CSE),final
U.I.T. B.U. Bhopal

Yachana Bhawsar
M.Tech.(CSE),final
U.I.T. B.U. Bhopal

Prof. Divakar Singh
HOD(CSE/IT)
U.I.T. B.U. Bhopal

Abstract

Data mining and knowledge discovery in databases have been attracting a foremost amount of research, industry, and media attention of late. What is all the excitement about? This paper provides an overview of this emerging field, clarifying how data mining and knowledge discovery in data-bases are related both to each other. The paper mentions particular realworld applications, specific data mining tech-niques, challenges involved in real world app-lications of knowledge discovery, and current and future research directions in the field.

Keywords

Data mining, KDD

1. Introduction

At an abstract level, the KDD field is concerned with the development of methods and techniques for making sense of data. The basic problem addressed by the KDD process is one of mapping low-level data (which are typically too voluminous to understand and digest easily) into other forms that might be more compact (for example, a Short report), more abstract (for example, a descriptive approximation or model of the process that generated the

data), or more useful (for example, a predictive model for estimating the value of future cases). At the core of the process is the application of specific data mining methods for pattern discovery and extraction

2. Key features of KDD

The traditional method of turning data into knowledge relies on manual analysis and interpretation[2]. For example, in the health-care industry, it is common for specialists to periodically analyze current trends and changing health-care data, say, on a quarterly basis. The specialists then provide a report detailing the analysis to the sponsoring health-care organiz-ation; this report becomes the basis for future decision making and planning for healthcare management. In a totally different type of appli-cation, planetary geologists sift through remotely sensed images of planets and asteroids, carefully locating and cataloging such geologic objects of interest as impact craters. Be it science, market-ing, finance, health care, retail, or any other field, the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products. Databases are increasing in size in two ways: (1) the number N of records or objects in the database and (2) the number d of fields or attributes to an object. Databases containing on the order of N = 109 objects are becoming increasingly common.

KDD is a proper subset of statistics. However, statisticians have not focused on considering issues related to large databases.

In addition, historically, the majority of the work has been primarily focused on hypothesis-verification as the primary mode of data analysis (which is certainly no longer true now). The de-coupling of data-base issues (storage and retrieval) from analysis issues is also a culprit. Furthermore, compared with techniques that data mining draws on from pattern recognition, machine learning, and neural networks, the traditional approaches in statistics perform little search over models and parameters (again with notable recent exceptions). KDD is concerned with formalizing and encoding aspects of the "art" of statistical analysis and making analysis methods easier to use by those who own the data, regardless of whether they have the pre-requisite knowledge of the techniques being used. We do not dismiss the dangers of blind mining and that it can easily deteriorate to *data dredging*. Hence KDD is an attempt to address a problem that the digital information era made a fact of life for all of us: *data overload*.

3. Need of KDD in today's World

One of the most important parts of a scientist's work is the discovery of patterns in data. Yet the databases of modern science are frequently so immense that they preclude direct human analysis. Inevitably, as their methods for gathering data have become automated, scientists have begun to search for ways to automate its analysis as well. Over the past five years, investigators in a new field called knowledge discovery and data mining have had notable successes in training computers to do what was once a unique activity of the human brain.

The study of climate change provides an excellent example of the difficulties of extracting useful information from modern data-bases[4]. In recent years, many questions about the earth's climate have moved into the realm of public policy: How rapidly are we destroying tropical rain forests? What might be the effects of global warming, and has it started already?

Be it a satellite orbiting our planet, a medical imaging device, a credit-card transaction verification system, or a super-market's checkout system, the human at the other end of the data gathering and storage machinery is faced with the same problem: *What to do with all this data?* Ignoring whatever we cannot analyze would be wasteful and unwise. Should one choose to ignore valuable information buried within the data, then one's competition may put them to good use; perhaps to one's detriment. In scientific endeavors, data represents observations carefully collect-ed about some phenomena under study, and the race is on for who can explain the observations best. In business endeavors, data captures information about the markets, competitors, and customers. In manu-facturing, data captures performance and optimi-zation opportunities, and keys to improving processes and troubleshooting problems[1].

Other real world application of both KDD and Data Mining are in marketing, investment, fraud detection, manufacturing, and telecommuni-cations and data cleaning.

4. Role of Data Mining in KDD

Data mining is just one part of the process of *knowledge discovery in data bases* (often abbreviated KDD)[5]. KDD is an iterative process with six stages: 1) develop an under-standing of the proposed application; 2) create a target data set; 3) remove or correct corrupted data; 4) apply data-reduction algorithms; 5) apply a data-mining algorithm and 6) interpret the mined patterns. Some steps may be skipped, and the process is not necessarily sequential often the results of one step cause the practitioner to back up to an earlier one. Although research tends to focus on step 5, the steps before and after data mining are equally important. In particular, it takes an expert in the application field, not a KDD expert, to decide whether the mined patterns are meaningful.

After defining a particular problem, the next step should be the creation of *training data*. This is a subset of the data

that trains the data-mining algorithm to interpret the rest of the data correctly just as a student learns a new subject by solving practice problems. Sometimes the KDD system itself can identify useful portions of the data for training; other times, a domain expert (human or not) performs this task.

The choice of training examples is both an important and an extremely challenging task. The sample must be large enough to justify the validity of the discovered know-ledge. Otherwise, like a student who has done too few practice problems, the data-mining program is likely to discover "rules" that don't work on other parts of the data. (Because of the low quality of the output, mining small data sets is sometimes called data dredging)

After selecting the training data, the next step is to clean the data and select or enhance the relevant features. For example, in a business application a person's income might be relevant to marketing strategies, but his or her Social Security number would not. Data reduction through feature enhancement is part-icularly important in image databases because of the sheer magnitude of the pixel data.

Next comes the choice of a data-mining algorithm. There are myriad methods available, and the choice depends strongly on the kind of data and the intended use for the mined knowledge. Is the model intended to be predictive or explanatory? Should the patterns discovered be understandable by people, or is reliability the most important consideration? Neural networks, for example, have been very popular in the machine learning community—they have been used to create machines that can recognize barcodes or learn to steer a car. But they are often less suitable for KDD, because they do not explain to a human user how they arrive at their predictions. If the goal of knowledge discovery is *human* knowledge, a "blackbox" oracle cannot be a desirable solution.

Decision-tree algorithm produces a classification structure that is particularly easy for a person to understand[4]. Decision trees have had many business and industrial applications; for example, in the financial industry, they are used to decide whether a person is a good or poor credit risk. We shall explain below what a decision tree is and how a decisiontree algorithm finds it.

Knowledge discovery does not end with the data mining algorithm. The remaining step is to interpret the meaning of the mined patterns and verify that they are accurate. The inter-pretation can often be assisted by visualization methods, another large field of research Accura-cy can be assessed by testing the results on a set of validation data. If the decision tree does not perform well enough, or if nothing of interest has been found, the investigator must go back to a previous step.

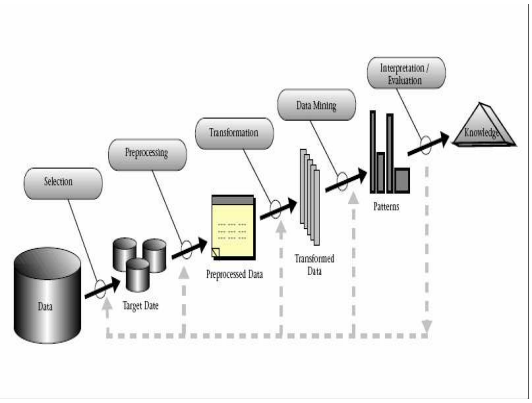

Figure 1: KDD Steps

5. Key elements of Data mining

There are three primary components in any data-mining algorithm: -
1) Model representation, 2) model-evolution,
3) Search [1].

5.1. Model representation is the lang-uage used to describe discoverable patterns. If the representation is too limited, then no amount of training time or examples can produce an accurate model for the data. It is important that a data analyst fully

comprehend the representtatio-nal assumptions that might be inherent in a particular method. It is equally important that an algorithm designer clearly state which represent-tational assumptions are being made by a parti-cular algorithm. Note that increased representat-ional power for models increases the danger of over fitting the training data, resulting in reduced prediction accuracy on unseen data.

5.2. Model-evaluation criteria are quantitative statements (or *fit functions*) of how well a particular pattern (a model and its parameters) meets the goals of the KDD process. For example, predictive models are often judged by the empirical prediction accuracy on some test set. Descriptive models can be evaluated along the dimensions of predictive accuracy, novelty, utility, and understandability of the fitted model.

5.3. Search method consists of two components: (1) parameter search and (2) model search. Once the model representation (or family of representations) and the model-evaluation criteria are fixed, then the data-mining problem has been reduced to purely an optimization task: Find the parameters and model from the selected family that optimizes the evaluation criteria. In parameter search, the algorithm must search for the parameters that optimize the model-evaluation criteria given observed data and a fixed model representation. Model search occurs as a loop over the parameter search method: The model representation is changed so that a family of models is considered.

There are some issues and challenges in data mining system they are-

1) Limited information, 2) Noise or missing data, 3) User interaction and prior knowledge, 4) Uncertainty, 5) size, updates and irrelevant fields [3].

6. Conclusion

Our primary aim was to clarify the relation between knowledge discovery and data mining. We provided an overview of the KDD process and basic data-mining methods. There are many data-mining techniques, particularly specialized methods for particular types of data and domain. This paper represents a step toward a common framework that we hope will ultimately provide a unifying vision of the common overall goals and methods used in KDD. We hope this will eventually lead to a better understanding of the relation between data mining and KDD. Knowledge Discovery has been proven to be a promising approach for enhancing the intelligence of software systems and services. Research in this area is broadly dispersed over various disciplines.

7. References

1) Jiawei han & Micheline kamber – "*data mining" –concept and technique*"
2) H liu and H Motorola, editors, *feature election for knowledge discovery and data mining*. Boston: Kluwer publishers, 1998.
3) Pujari Arun K -"*data mining techniques*"
4) FayyadU.M.Piatetsky-Shapiro G., Smyth P., Uthurusamy R (Eds): *Advances in Knowledge Discovery and Data Mining*. Menlo park, CA: AAAI Press/The MIT Press, 1996.
5) Heckerman D. Bayesian networks for data mining .*Data mining and Knowledge Discovery,* 1997.