INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

KNOWLEDGE DISCOVERY AND DATA MINING FRAMEWORK AND ITS APPLICATIONS

MISS. ABHILASHA A. DESHMUKH [2], DR. H. R. DESHMUKH[2]

1. Student of Master of Engineering in (CSE), IBSS college of Engineering and Technology, Amravati, India.
2. Head of the Department of (CSE), IBSS College of Engineering and Technology, Amravati, India

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: This paper presents a framework for Knowledge Discovery in Databases. We describe finks between data mining, knowledge discovery, and other related fields. We then define the KDD process and basic data mining algorithms, discuss application issues and conclude with an analysis of challenges facing practitioners in the field. . We look at the existing tools, describe some representative applications, and discuss the major issues and problems for building and deploying successful applications and their adoption by business users. Finally, we examine how to assess the potential of a knowledge discovery application

Keywords: Data Mining, Knowledge Discovery, Framework, Bayesian networks

Corresponding Author: MISS. ABHILASHA A. DESHMUKH

PAPER-QR CODE

Access Online On:

www.ijpret.com

How to Cite This Article:

Abhilasha A. Deshmukh, IJPRET, 2015; Volume 3 (9): 494-501

INTRODUCTION

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational techniques and tools to assist humans in extracting usefull information (knowledge) from the rapidly growing volumes of data, these techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD). This paper is an initial step towards a common framework that we hope will allow us to understand the variety of activities in this multidisciplinary field and how they fit together.

We view the knowledge discovery process as a *set* of various activities for making sense of data. At the core of this process is the application of *data mining* methods for pattern t discovery. We examine how data mining is used and outline some of its methods. Finally, we look at practical application issues of KDD and enumerate challenges for future research and development. Introduction patterns.

2. Data Mining Tools

Knowledge Discovery in Databases (KDD) is an umbrella term used to describe a large variety of activities for making sense of data. We will use the term knowledge discovery to describe the overall process of finding useful patterns in data, which includes not only the data mining step of running the discovery algorithms, but also pre- and post-processing, and various other activities Fields Historically the notion of finding useful patterns in data has been given a variety of name. The term *data mining* has been mostly used by statisticians, data analysts, and the ~Throughout this paper we use the term "pattern" to designate *pattern* or *model* extracted from the data management.

The practical view of tile KDD process interactive nature of the process. Here we broadly outline some of its basic steps:

1. Developing an understanding of the application domain and the relevant prior knowledge, and identifying the *goal* of the KDD process from the customer's view point.

2. Creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.

3. Data cleaning and preprocessing: basic operations such as the removal of noise if appropriate, collecting the necessary information to model or account for noise, deciding on

strategies for handling missing data fields, accounting for time sequence information and known changes.

4. Data reduction and projection: finding use fid features to represent tile data depending on tile goal of tile task. Using dimensionality reduction or transformation methods to reduce tile effective number of variables under consideration or to find invariant representations for the data.

5. Matching tile goals of tile KDD process to particular data mining *method:* e.g., summarization, classification, regression, clustering,

6. Choosing the data mining algorithm(s): selecting method(s) to be used for searching for patterns the data. This includes deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process

7. Data mining: searching for patterns of interest in a particular representational form or a set of such representations: classification rules or trees, regression, clustering, and so forth. The user can significantly aid the data mining method by correctly performing the preceding steps

8. Interpreting mined patterns, possibly return to any of steps 1-7 for further iteration. This step can also involve visualization of the extracted patterns/ models, or visualization of the data given the extracted models.

9. Consolidating discovered knowledge: incorporating this knowledge into another system for filrther action, or simply documenting it and reporting it to interested parties.

3. Data Mining Methods

Although the boundaries between prediction and description are not sharp the distinction is useful for understanding the overall discovery goal. This is in contrast to many machine learning and pattern recognition applications where prediction is often the primary goal. The goals of prediction and description are achieved via the following primary data mining methods.

- **Classification:** learning a function that maps (classifies) a data item into one of several predefined classes.
- **Regression:** learning a function which maps a data item to a real-valued prediction variable and the discovery of functional relationships between variables.

- **Clustering:** identifying a finite set of categories or clusters to describe the data. Closely related to clustering is the method of *probability density estimation* which consists of techniques for estimating from data the joint multi-variate probability density function of all of the variables/fields in the database.

- **Summarization:** finding a compact description for a subset of data, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques.

- **Dependency Modeling:** finding a model which describes significant *dependencies* between variables Change and Deviation Detection: discovering the most significant changes in the data from previously measured or normative values.

4. Data Mining Algorithms

There exist a wide variety of data mining algorithms. Popular model representations include decision trees and rules, nonlinear regression and classification, example-based methods (including nearest-neighbour and case-based reasoning methods), probabilistic graphical dependency models (including Bayesian networks), and relational learning models (including inductive logic programming). An important point is that each technique typically suits some problems better than others. For example, decision tree classifiers can be very useful for finding structure in high-dimensional spaces and are also useful in problems with mixed continuous and categorical data, However, classification trees may not be suitable for problems where the true decision boundaries between classes are described by a 2nd-order polynomial. Thus, there is no 'universal' data mining method and choosing a particular algorithm for a particular application is something of an art.

5. Representative Applications

Numerous knowledge discovery applications and prototypes have been developed for a wide variety of domains including marketing, finance, manufacturing, banking, and telecommunications. A majority of the applications have used predictive modeling approach, but there were also a few notable applications using other methods. Here we describe some of the representative examples.

5.1 Marketing

KDD tools to the rapidly growing sales and marketing databases. Because of a strong competitive pressure, the often saturated market potential and maturity of products, there is a shift from a quality to an information competition where detailed and comprehensive

knowledge on the behavior of customers and competitors is crucial. Market research companies collect data on special markets, analyze this data and sell data and analyses to their clients. The clients add their own data for further analyses. Medium sized datasets are captured when market research companies perform surveys (e.g. 2000 persons interviewed each month).

5.2 Investment

Many financial analysis applications employ predictive modeling techniques, such as statistical regression or neural networks, for tasks like portfolio creation and optimization and trading model creation. Such applications have been in use for several years. To maintain a competitive advantage, the users and developers of such applications rarely publicize their exact details and effectiveness. We can, however, mention a few examples. Fidelity Stock Selector fund is using neural network models to select investments and has performed quite well until recently

5.3 Fraud Detection

Not all the systems developed for this have been publicized, for obvious reasons, but several are worth mentioning. The HNC Falcon credit risk assessment system, developed using a neural network shell, is used by a large percentage of retail banks to detect suspicious credit card transactions. Falcon deployment was facilitated by the fact that credit card transaction data is captured by just a few companies. Even though each such company uses its own data format, every bank issuing credit cards uses one of these few formats. Therefore, an application that can work with even one format effectively can easily be adopted by a large number of banks.

6. Manufacturing and Production processes

The process is an application of KDD with a high potential profit. The goal is to discover process conditions that lead to good quality products. At present, large volumes of data generated during a production process are often only poorly exploited. Also, the relations between the control, process, and quality variables are not completely understood by the engineers. In addition, time and space constraints, which play an especially important role in manufacturing, are not well handled by most data mining tools. A typical example is a project which is run in a large chemical company in Europe to analyze a production process in a plant for polymeric plastics. Data includes control variables (e.g. quantities of raw material, the heating parameters), the process variables (temperatures, pressures, and chemical reaction times), and quality variables measured in a laboratory. Quality variables are determined several times a day, process and control variables nearly continuously.

6.1 Other Areas

Health care is an information-rich and high payoff area, ripe for data mining. The system performs an automatic drill-down through data along multiple dimensions to determine the most interesting deviations of specific quantitative measures relative to their previous and expected values. It explains "key" deviations through their relationship to other deviations in the data, and, where appropriate, generates recommendations for actions in response to these deviations. KEFIR uses a Web browser to present its findings in a hypertext report, using natural language and business graphics. Improving data quality is another important application area.

6.2 Discovery Agents

Finally, a novel and very important type of discovery system has appeared recently – Discovery Agents. Although the idea of active triggers has long been analyzed in the database field, really successful applications of this idea appeared only with the advent of the Internet. These systems ask the user to specify a profile of interest and search for related information among a wide variety of public domain and proprietary sources.

7. Assessing Benefits of KDD Applications

The domains suitable for data mining are those that are information rich, have a changing environment, do not already have existing models, require knowledge-based decisions, and provide high pay off for the right decisions. Given a suitable domain, we examine costs and benefits of by looking at the following factors.

- **Alternatives**: there should be no simpler alternative solutions.
- **Relevance:** relevant factors should be included.
- **Volume:** there should be a sufficient number of cases (several thousand at least). Extremely large databases may be a problem when the results are needed quickly.
- **Complexity:** the more variables (fields) there are the more data.
- **Quality**: Error rate should be relatively low.
- **Accessibility**: data should be easily accessible - accessing data or merging data from different sources increases the cost of an application.

- **Change:** although dealing with change is more difficult, it can also be more rewarding (the volatility benefit) since the application can be automatically and regularly "re-trained" on up to-date data.

- **Expertise**: The more expertise available, the easier is the project. It should be emphasized that expertise on the form and meaning of the data is just as important as knowledge of problem solving in the domain. Although the challenges are many and the difficulties are substantial, the future of data mining applications looks bright.

8. ACKNOWLEDGEMENTS

We thank to H. R. Deshmukh Sir, Nikhil Band Sir and Ankit Mune Sir. For helpful discussions and comments.

9. CONCLUSION

This paper presents a framework for Knowledge Discovery in Databases. We describe finks between data mining, knowledge discovery, and other related fields. Finally, we examine how to assess the potential of a knowledge discovery application. We further subdivide the Discovery goal into prediction, where the system finds patterns for the purpose of predicting the future behaviour of some entities; and description, where the sy5 tern finds patterns for the purpose of presenting them to a user in a human-understandable form.

REFERENCE:

1. Anand, T. and Kahn, G. 1992. SPOTLIGHT: A Data Explanation System. In Proceedings Eighth IEEE Conference on Applied AI, 2-8. Washington, D.C.: IEEE Press.

2. Anand, T. 1995. Opportunity Explorer: Navigating Large Databases Using Knowledge Discovery Templates. Journal of Intelligent Information Systems 4(1):27-38.

3. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. 1996. Fast Discovery of Association Rules. In AKDDM, Cambridge, MA: AAAI/MIT Press.

4. Agrawal, R., and Psaila, G. 1995. Active Data Mining. In Proceedings of KDD-95, 3-8, Menlo Park, CA: AAAI Press.

5. Brachman, R., et al. 1993. Integrated Support for Data Archaeology. In Proceedings of KDD-9J.

6. Workshop, Menlo Park, CA: AAAI Press. Brachman, R. and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human Manago, M. and Auriol, M. 1996. Mining for OR. ORMS Today, February, Special issue on Data Mining, 28-32.

7. Mannila, H., Toivonen, H., and Verkamo, A. 1995. Discovering Frequent Episodes in Sequences, In Proceedings of KDD-95, 210-215. Menlo Park, CA: AAAI Press.

8. Matheus, C., Piatetsky-Shapiro, G., and Mc-Neill, D. 1996. Selecting and reporting what is Interesting: The KEFIR Application to Healthcare Piatetsky-Shapiro, G. 1995.