www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 11 Nov 2015, Page No. 14966-14968

The Interdisciplinary Nature Of Knowledge Discovery Databases And Data Mining

T.Bharathi
Assistant Professor, Department of Computer Science
Arignar Anna Govt Arts College. Villupuram.
Email: mithulvarsha19@gmail.com

ABSTRACT: **Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. In the last few years, knowledge discovery and data mining tools have been used mainly in experimental and research environments, business user etc. A large degree of the current interest in KDD is the result of the media interest surrounding successful KDD applications, for example, the focus articles within the last two years in Business Week, Newsweek, Byte, PC Week, and other large-circulation periodicals. Unfortunately, it is not always easy to separate fact from media hype. Nonetheless, several well documented examples of successful systems can rightly be referred to as KDD applications and have been deployed in operational use on large-scale real-world problems in science and in business.**

*Keywords:***Datamining,Database,Data reduction,Regression**

I. Introduction

The rapid emergence of electronic data management methods has made us to enter into the era of so called "Information Age." Powerful database systems for collecting and managing are made used in virtually all large and mid-range companies -- there is hardly a transaction that does not generate a computer record

somewhere. Each year more operations are being computerized, all accumulate data on operations, activities and performance. All those data hold valuable information, e.g., trends and patterns, which could be used to improve business decisions and optimize success. However, today's database contains so much data that it becomes almost impossible to manually analyze them for valuable decision-making information. In many cases, hundreds of independent attributes need to be simultaneously considered in order to accurately model system behavior. Therefore, humans need assistance in their analysis capacity. This need for automated extraction of useful knowledge from huge amounts of data is widely recognized now, and leads to a rapidly developing market of automated analysis and discovery tools. Knowledge discovery and data mining are techniques to discover strategic information hidden in very large databases. Automated discovery tools have the capability to analyze the raw data and present the extracted high level information to the analyst or decision-maker, rather than having the analyst finds it for himself or herself. In the last few years, knowledge discovery and data mining tools have been used mainly in experimental and research environments[7]. Powerful database systems for collecting and managing data are in use now-a-days in virtually all large and mid-range companies. Each year more operations are being computerized, accumulating all data on operations, activities and performance. All these data hold valuable information, for e.g., trends and patterns, which could be used to improve business decisions and optimize success. Automated discovery tools have the capability to analyze the raw data and present the extracted high level information to the analyst or decision-maker, rather than having the analyst himself.

II. Knowledge Discovery Process

There is still some confusion about the terms Knowledge Discovery in Databases (KDD) and data mining. Often these two terms are used interchangeably. The term KDD is used to denote the overall process of turning low-level data into high-level knowledge. A simple definition of KDD is as follows: Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Hence data mining is just one step in the overall KDD process. The following Steps are involved in the KDD Process[10]

- Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end user with today's technology,
- 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. This involves considerations of homogeneity of data, any dynamics and change over time, sampling strategy (such as uniform random versus stratified), sufficiency of sample, degrees of freedom, and so forth.
- Data cleaning and preprocessing. Basic operations such as the removal of noise or "outliers," if appropriate; collecting the necessary information to model or

DOI: 10.18535/Ijecs/v4i11.16

accounting for noise; deciding on strategies for handling missing data fields; accounting for time sequence information, known changes, and appropriate normalization; and so forth are involved here.[6]

- Data reduction and transformation This involves finding useful features to represent the data, depending on the goal of the task' using dimensionally reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data; and projecting the data onto spaces in which a solution is likely to be easier to find.
- Choosing the data mining task.This involves deciding whether the goal of the KDD process is classification, regression, clustering, summarization, dependency modeling, or change and deviation detection[3]
- Choosing data mining algorithms - select the methods to be used for searching for patterns in or fitting models to the data.

III. Database Issues

Any realistic knowledge discovery process is not linear, but rather iterative and interactive. Any one step may result in changes in earlier steps, thus producing a variety of feedback loops. This motivates the development of tools that support the entire KDD process, rather than just the core data-mining step. Such tools require a tightly integrated database systems or data warehouse for data selection, preprocessing, integrating, transformation etc.[8]

Many tools currently available are generic in nature. Such tools usually operate separately from the data source, requiring a significant amount of time spent with data export and import, pre-and post-processing, and data transformation. However, a tight connection between the knowledge discovery tool and the analyzed database, utilizing the existing DBMS support, is clearly desirable. For the reviewed knowledge discovery tools, the following features are inspected: Ability to access a variety of data sources: In many cases, the data to be analyzed is scattered throughout the corporation, it has to be gathered, checked, and integrated before doing a meaningful analysis

IV. Data Mining Methods

At the core of the KDD process are the data mining methods for extracting patterns from data. These methods can have different goals, dependent on the intended outcome of the overall KDD process. It should also be noted that several methods with different goals may be applied successively to achieve a desired result. For example, to determine which customers are likely to buy a new product, a business analyst might need to first use clustering to segment the customer database. Then he/she may apply regression to predict buying behavior for each cluster. Most data mining goals fall under the following categories:

Data Processing: Depending on the goals and requirements of the KDD process, analysts may select, filter, aggregate, sample, clean and /or transform data. Automating some of the most typical data processing tasks and integrating them seamlessly in to the overall process may eliminate or at least greatly reduce the need for programming specialized routines and for data export/import, thus improving the analyst's productivity[2].

Prediction: Given a data item and a predictive model, predict the value for a specific attribute of the data item. For example, given a predictive model of credit card transactions, predict the likelihood that a specific transaction is fraudulent.

Regression: Given a set of data items, regression is the analysis of the dependency of some attribute values upon the values of other attributes in the same item, and the automatic production of a model that can predict these attribute values for new records. For example, given a data set of credit card transactions, build a model that can predict the likelihood of fraudulence for new transactions[6].

Classification: Given a set of predefined categorical classes, determine to which of these classes a specific data item belongs. For example, given classes of patients that correspond to medical treatment responses, identify the form of treatment to which a newpatient is most likely to respond.

Clustering: Given a set of data items, partition this set into a set of classes such that items with similar characteristics are grouped together. Clustering is best used for finding groups of items that are similar.[2]

Model Visualization: Visualization plays an important role in making the discovered knowledge understandable and interpretable by humans. Besides, the human eye-brain system

still remains the best pattern-recognition device known. Visualization techniques may range from simple scatter plots and histogram plots over parallel coordinates to 3D movies.

V. Framework - Data mining models

Data mining also known as knowledge discovery in databases, which analysis and exploration a large amount of data in the database , data warehouse or other information storeroom according to business objectives, reveals the hidden, unknown, non-trivial and potentially value of application regularity, and establish model. Data mining is a data analysis tool, which has unparalleled advantages compared to data analysis tools such as statistical analysis, on-line transaction processing (OLTP) and online analytical processing (OLAP). Knowledge discovery in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. In the last few years, knowledge discovery and data mining tools have been used mainly in customer relationship management system .The application of data mining technique in CRM is an emerging trend in business to improve the profitability of their interaction customers. In present decade, customer orientation has been one of the major concerns of commercial companies. Discovering the needs of customers and scheduling to meet those needs, or in other words customer relationship management (CRM), is an undeniable fact that can have an influence in customers' attraction and changing them into regular customers, which consequently leads to an increase in the company's interests. Data mining is one of the ways of discovering the Customers" needs in business for data analysis . This paper presents a comprehensive review of literature related to application of data mining

Within the context of CRM, data mining can be seen as a business driven process aimed at the discovery and consistent use of profitable knowledge from organizational data . It can be used to predict how likely an existing customer is to take his business to a competitor. Each of the CRM elements can be supported by different data mining models, which generally include association, classification, clustering, forecasting, regression, sequence discovery ,visualization.

(i) Association: Association aims to establishing relationships between items which exist together in a given record. Market basket analysis and cross selling programs are typical examples for which association modeling is usually adopted. Common tools for association modeling is apriori algorithms.

(ii) Classification: Classification is one of the most common learning models in data mining It aims at building a model to predict future customer behaviors through classifying
database records into a number of predefined classes based on certain criteria Common tools used for classification are neural networks, decision trees and if then-else rules.[5]

(iii) Clustering: Clustering is the task of segmenting a heterogeneous population into a number of more homogenous clusters.It is different to classification in that clusters are unknown at the time the algorithm starts.In other words, there are no predefined clusters. Common tools for clustering include neural networks and discrimination analysis.

(iv) Forecasting: Forecasting estimates the future value based on a record's patterns. It deals with continuously valued outcomes. It relates to
modeling and the logical relationships of the model at some time in the future. Demand forecast is a typical example of a forecasting model. Common tools for forecasting include neural networks and survival analysis.

VI. Conclusions

In this paper, we presented categorizes, compares, summarizes of the different data mining model and application of data mining,KDD. The application of KDD using Data mining system was quite general. This review paper is useful for applying data mining models on the existing database and generating new association rules to my future research by using data mining the experienced data. Successful Knowledge Discovery and Data Mining applications play an important role in data that have clearly grown to surpass raw human processing abilities. Developing new mining algorithms for classification, clustering, dependency analysis, and change and deviation detection that scales to large databases. Developing new mining and search algorithms capable of extracting more complex relationships between fields and able to account for structure over the fields (hierarchies, sparse relations).

References

[1] Xindong Wu. "Data Mining: artificial intelligence in data analysis". Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2004.pp.7..

[2] R. Evans, and D. Fisher, "Overcoming Process Delays with Decision Tree Induction," IEEE Expert, Vol. 9, No. 1,1994, pp. 60–66.

[3] S. Brin et al., "Dynamic Itemset Counting and Implication Rules for Market Basket Data," Proceedings of ACM SIGMOD Int'l Conf. Management of Data, 1997, pp.255–264.

[4] M. Pazzani, S. Mani, andW.R. Shankle, "ComprehensibleKnowledge-Discovery in Databases," Proceeding of 19thAnnual Conf. Cognitive Science Soc.1997, pp. 596–601.

[5] R. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, Industrial Applications of DataMining and Knowledge Discovery, Communzcatzons of ACM, vol. 39, no. 11.1996. Mining, vol. 39, no. 11.

[6] Xindong Wu. "Data Mining: artificial intelligence in data analysis". Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2004.pp.7..

[7] R. Evans, and D. Fisher, "Overcoming Process Delays with Decision Tree Induction," IEEE Expert, Vol. 9, No. 1, 1994, pp. 60–66.

[8] S. Brin et al., "Dynamic Itemset Counting and Implication Rules for Market Basket Data," Proceedings of ACM SIGMOD Int'l Conf. Management of Data, 1997, pp 255–264.

[9] M. Pazzani, S. Mani, andW.R. Shankle, "Comprehensible Knowledge-Discovery in Databases," Proceeding of 19th Annual Conf. Cognitive Science Soc.1997, pp. 596–601.

[10] R. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, Industrial Applications of Data Mining and Knowledge Discovery, Communzcatzons of ACM, vol. 39, no. 11.1996.