Comparison between Data Mining Algorithms Implementation

Yas A. Alsultanny

College of Graduate Studies-Arabian Gulf University
Kingdom of Bahrain
alsultanny@hotmail.com

Abstract. Data Mining (DM) is the science of extracting useful information from the huge amounts of data. Data mining methods such as Na¨ıve Bayes, Nearest Neighbor and Decision Tree are tested. The implementation of the three algorithms showed that Na¨ıve Bayes algorithm is effectively used when the data attributes are categorized, and it can be used successfully in machine learning. The Nearest Neighbor is most suitable when the data attributes are continuous or categorized. The last algorithm tested is the Decision Tree, it is a simple predictive algorithm implemented by using simple rule methods in data classification. Each of the three algorithms can be implemented successfully and efficiently after studying the nature of database according to their; size, attributes, continuity and repetition. The success of data mining implementation depends on the completeness of database, that represented by data warehouse, that must be organized by using the important characteristics of data warehouse.

Keywords: Data mining, knowledge discovery, Nearest Neighbour, Na¨ıve Bayes, Decision.

1 Introduction

Data mining is one of the important methods used in decision making by transforming data from different resources such as from data warehouse to knowledge extraction, in this section an overview of data mining will be introduced to show the basic principles of data mining.

Data mining is a process that is used to identify hider, unexpected pattern or relationships in large quantities of data. Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term data mining has mostly been used by statisticians, data analysts and the Management Information Systems (MIS) communities. It has also gained popularity in the database field. The phrase knowledge discovery in databases was coined at the first KDD to emphasize that knowledge is the end product of a data-driven discovery. It has been popularized in the artificial intelligence and machine-learning fields. Fig. 1 shows an overview of the data mining and KDD process [1].

H. Cherifi, J.M. Zain, and E. El-Qawasmeh (Eds.): DICTAP 2011, Part II, CCIS 167, pp. 629–641, 2011.

Data mining predicts future trends and behaviors, allowing business to make proactive, knowledge driven decisions. It moves beyond the analysis of past events provided by retrospective tools typical of decision support systems, answering questions that traditionally where too time consuming to resolve. Data mining scours databases for hidden patterns, finding predictive information that experts might overlook because it falls outside their expectations.

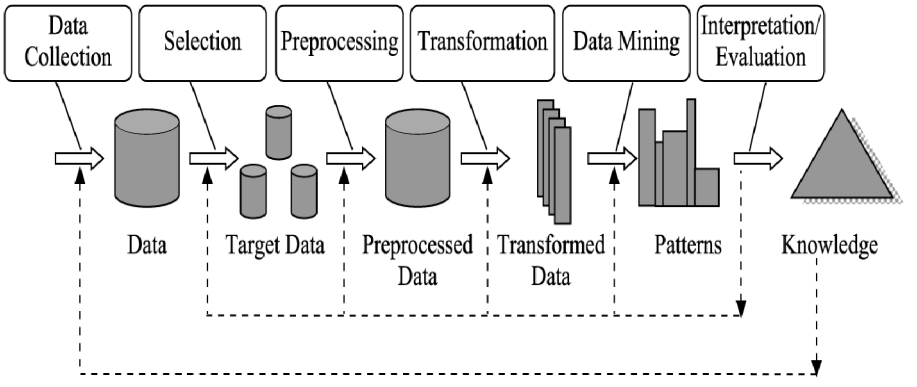

Fig. 1. An overview of the data mining and KDD process

The two high-level primary goals of data mining in practice tend to be prediction and description. The prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. Although the boundaries between prediction and description are not sharp (some of the predictive models can be descriptive, to the degree that they are understandable, and vice versa), the distinction is useful for understanding the overall discovery goal. The goals of prediction and description can be achieved using a variety of particular data-mining methods.

- Classification
- Regression
- Clustering
- Summarization
- Dependency modeling

2 Data Warehouse

Data warehouse is a repository of subject-oriented historical data that is organized to be accessible in a form readily acceptable for analytical processing activities (such as data mining, decision support querying, and other applications) [2].
The major benefits of a data warehouse are:

- The ability to reach data quickly, since they are located in one place.
- The ability to reach data easily and frequently by end users with Web browsers.