Chapter 4
Data Mining and Knowledge Discovery

In Chapters 2 and 3, the knowledge-based system and Multi-Agent system were illustrated. These are significant methods and theories of Manufacturing Intelligence (MI). Data Mining (DM) and Knowledge Discovery (KD) are at the foundation of MI. Humans are immersed in data, but are thirsty for knowledge. With the wider application of database technology, a dilemma has arisen whereby people are ‘rich in data, poor in knowledge’. The explosion of knowledge and information has brought great benefit to mankind, but has also carried with it certain drawbacks, since it has resulted in knowledge and information ‘pollution’. Facing a vast but polluted ocean of data, a technical means to discard the bad and retain the good was sought. Data Mining and Knowledge Discovery (DMKD) was therefore proposed against the background of rapidly expanding data and databases. It is also the result of the development and fusion of database technology, Artificial Intelligence (AI), statistical techniques and visualization technology (Fayyad U., 1998). DMKD has become a research focus and cutting-edge technology in the field of computer information processing (Jef Woksem, 2001). The development background, conception, working process, classification and general application of DM and KD are firstly introduced in this chapter. Secondly, basic functions and assignment such as prediction, description, data clustering, data classification, conception description and visualization processing are discussed. Then the methods and tools for DM are presented, such as the association rule, decision tree, genetic algorithm, rough set and support vector machine. Finally, the application of DMKD in intelligent manufacturing is summarized.

INTRODUCTION

Background and Conception

DMKD originated from Knowledge Discovery in Database (KDD). It first appeared in August 1989 at a meeting of the 11th International Joint Conference of Artificial Intelligence. KDD has been defined as the non-trivial process of identifying hidden, previously unknown and potentially useful information.
There are many similar ways of describing KDD, such as Data Mining (DM), Knowledge Extraction, Information Discovery, Information Harvesting, and Data Archaeology. Although definitions differ, the essence of extracting hidden, interesting and high-level models from the data is uniform. Of these names, KDD and DM are two commonly-used terms. Scholars in the field of statistics, data analysis and information systems regular employ ‘DM’, but experts in the field of AI and machine learning use ‘KDD’. DM, etc.

In order to unify understanding, Fayyad offered a new definition of KDD and DM and made a distinction between them in a field paper (Fayyad U. & Piatetsky-Shapiro G. et al., 1996). The new definition for KDD is this: KDD is a process of identifying effective, innovative, potentially useful and ultimately understandable models. DM is defined as: DM is a step that generates a specific model through a specific algorithm in an acceptable computational efficiency. KDD is an entire process including the steps of data choice, data preprocessing, data transformation, DM, and pattern evaluation, which eventually lead to obtaining knowledge. DM is one of the key steps. Despite the distinction, scholars often equate KDD with DM for convenience. When referring to DM’ alone, it is taken to also mean the data preprocessing and the results of evaluation. Consequently, people tend to use these two names together, so DMKD occurs.

In simple terms, DM is extracting or mining knowledge, and can be defined from the point of statistics, databases, machine learning and so on. ‘Mining’ first appeared in statistics, from which perspective, DM is analyzing the data sets observed in order to find the unknown relationship between data, and providing understandable, innovative and useful summarized data to the data owner (Han J.W. & Kamber M., 2002). From the perspective of databases, DM is a process of finding interesting knowledge from large amounts of data stored in databases, data warehouse and other information storage (Han J.W. & Kamber M., 2001). From the perspective of machine learning, DM is defined as extracting the connotative, obviously unknown and potentially useful information from data (Witten I. H. & Frank E., 2000).

From 1989 to date, the DMKD international conferences sponsored by the American Association for Artificial Intelligence have been held many times. The conference was a biennial symposium before 1993, but since 1995 it has been held once a year and has developed into the international academic conference named KDD. The attendance increased from dozens to hundreds, and the ratio of accepted papers rose from 2:1 to 6:1. In 1997, the first Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) was held in the Asia-Pacific region and has since been held annually. The DMKD academic conferences have also been held once a year in Europe since 1997 under the name of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). In recent years, relevant international conferences such as Very Large Data Bases (VLDB) and Special Interest Group on Management of Data (SIGMOD) have also attracted a large number of DMKD papers, reports and corresponding topics. A DMKD special has been published in the IEEE Journal of Knowledge and Data Engineering, Intelligent Systems, Computational Intelligence and other journals in recent years. The international DMKD journal ‘Data Mining and Knowledge Discovery’ was first published in January 1997.

So far, much progress has been achieved in the research of DM and KD in relational databases and service databases. DMKD technology was initially application-oriented, because it originated against a background of strong application demand. In the international community, DMKD technology has been used in marketing, finance and banking, insurance, telecommunications, transportation, and other fields. In recent years, the spatial database,