IDEA GROUP PUBLISHING **ITB12936**

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

This paper appears in the publication, *Advanced Data Mining Technologies in Bioinformatics*
edited by Hui-Huang Hsu Reichgelt © 2006, Idea Group Inc.

Chapter I

Introduction to Data Mining in Bioinformatics

Hui-Huang Hsu, Tamkang University, Taipei, Taiwan

Abstract

Bioinformatics uses information technologies to facilitate the discovery of new knowledge in molecular biology. Among the information technologies, data mining is the core. This chapter first introduces the concept and the process of data mining, plus its relationship with bioinformatics. Tasks and techniques of data mining are then presented. At the end, selected bioinformatics problems related to data mining are discussed. Data mining aims at uncovering knowledge from a large amount of data. In molecular biology, advanced biotechnologies enable the generation of new data in a much faster pace. Data mining can assist the biologist in finding new knowledge from piles of biological data at the molecular level. This chapter provides an overview on the topic.

Introduction

Progress of information technologies has made the storage and distribution of data much easier in the past two decades. Huge amounts of data have been accumulated at a very fast pace. However, pure data are sometimes not-that-useful and meaningful because what people want is the knowledge/information hidden in the data. Knowledge/information can be seen as the patterns or characteristics of the data. It is much more valuable than data. Thus, a new technology field has emerged in the mid 1990's to deal with the discovery of knowledge/information from data. It is called *knowledge discovery in databases (KDD)* or simply *data mining (DM)* (Chen et al., 1996; Fayyad et al., 1996). Although knowledge and information can sometimes be distinguished, we will treat them as the same term in this chapter.

Data pile up like a mountain. But most of them are not that useful, just like earths and rocks in a mountain. The valuables are metals like gold, iron, or diamond. Just like the miner wants to dig out the valuables from the earth and rock, the data miner uncovers useful knowledge/information by processing a large amount of data. Formal definitions of KDD and DM have been given in different ways. Here are three examples: "Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad, 1996, p. 20). And, "Data mining is the process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions" (Simoudis, 1996, p. 26). Or, to be simpler, "data mining is finding hidden information in a database" (Dunhum, 2003, p. 3). Uncovering hidden information is the goal of data mining. But, the uncovered information must be:

1. **New:** Common sense or known facts are not what is searched for.
2. **Correct:** Inappropriate selection or representation of data will lead to incorrect results. The mined information needs to be carefully verified by domain experts.
3. **Meaningful:** The mined information should mean something and can be easily understood.
4. **Applicable:** The mined information should be able to be utilized in a certain problem domain.

Also, in Simoudis (1996), crucial business decision making is emphasized. That is because the cost of data mining is high and was first applied in business problems, e.g., customer relationship management, personalized advertising, and credit card fraud detection.

There is actually a slight difference between KDD and DM. Data mining is the algorithm/method used to find information from the data. Meanwhile, KDD is the whole process including data collection, data preprocessing, data mining, and information interpretation. So DM is the core of KDD. However, data preprocessing and information interpretation are also very important. Without proper preprocessing, the quality of data might be too bad to find meaningful information. Also, without correct interpretation, the mined information might be mistakenly used. Thus, the cooperation between data mining