Chapter I
Introduction to Data Mining Techniques via Multiple Criteria Optimization Approaches and Applications

Yong Shi
University of the Chinese Academy of Sciences, China
and University of Nebraska at Omaha, USA

Yi Peng
University of Nebraska at Omaha, USA

Gang Kou
University of Nebraska at Omaha, USA

Zhengxin Chen
University of Nebraska at Omaha, USA

ABSTRACT

This chapter provides an overview of a series of multiple criteria optimization-based data mining methods, which utilize multiple criteria programming (MCP) to solve data mining problems, and outlines some research challenges and opportunities for the data mining community. To achieve these goals, this chapter first introduces the basic notions and mathematical formulations for multiple criteria optimization-based classification models, including the multiple criteria linear programming model, multiple criteria quadratic programming model, and multiple criteria fuzzy linear programming model. Then it presents the real-life applications of these models in credit card scoring management, HIV-1 associated dementia (HAD) neuronal damage and dropout, and network intrusion detection. Finally, the chapter discusses research challenges and opportunities.

INTRODUCTION

Data mining has become a powerful information technology tool in today's competitive business world. As the sizes and varieties of electronic datasets grow, the interest in data mining is increasing rapidly. Data mining is established on the basis of many disciplines, such as machine learning, databases, statistics, computer science, and operations research. Each field comprehends data mining from its own perspective and makes its distinct contributions. It is this multidisciplinary nature that brings vitality to data mining. One of the application roots of data mining can be regarded as statistical data analysis in the pharmaceutical industry. Nowadays the financial industry, including commercial banks, has benefited from the use of data mining. In addition to statistics, decision trees, neural networks, rough sets, fuzzy sets, and vector support machines have gradually become popular data mining methods over the last 10 years. Due to the difficulty of accessing the accuracy of hidden data and increasing the predicting rate in a complex large-scale database, researchers and practitioners have always desired to seek new or alternative data mining techniques. This is a key motivation for the proposed multiple criteria optimization-based data mining methods.

The objective of this chapter is to provide an overview of a series of multiple criteria optimization-based methods, which utilize the multiple criteria programming (MCP) to solve classification problems. In addition to giving an overview, this chapter lists some data mining research challenges and opportunities for the data mining community. To achieve these goals, the next section introduces the basic notions and mathematical formulations for three multiple criteria optimization-based classification models: the multiple criteria linear programming model, multiple criteria quadratic programming model, and multiple criteria fuzzy linear programming model. The third section presents some real-life applications of these models, including credit card scoring management, classifications on HIV-1 associated dementia (HAD) neuronal damage and dropout, and network intrusion detection. The chapter then outlines research challenges and opportunities, and the conclusion is presented.

MULTIPLE CRITERIA OPTIMIZATION-BASED CLASSIFICATION MODELS

This section explores solving classification problems, one of the major areas of data mining, through the use of multiple criteria mathematical programming-based methods (Shi, Wise, Luo, & Lin, 2001; Shi, Peng, Kou, & Chen, 2005). Such methods have shown its strong applicability in solving a variety of classification problems (e.g., Kou et al., 2005; Zheng et al., 2004).

Classification

Although the definition of classification in data mining varies, the basic idea of classification can be generally described as to "predicate the most likely state of a categorical variable (the class) given the values of other variables" (Bradley, Fayyad, & Mangasarian, 1999, p. 6). Classification is a two-step process. The first step constructs a predictive model based on training dataset. The second step applies the predictive model constructed from the first step to testing dataset. If the classification accuracy of testing dataset is acceptable, the model can be used to predicate unknown data (Han & Kamber, 2000; Olson & Shi, 2005).

Using the multiple criteria programming, the classification task can be defined as follows: *for a given set of variables in the database, the boundaries between the classes are represented by scalars in the constraint availabilities.* Then, the standards of classification are measured by minimizing the total overlapping of data and maximizing the distances of every data to its class boundary