Multiple Criteria Optimization in Data Mining

Gang Kou
University of Electronic Science and Technology of China, China

Yi Peng
University of Electronic Science and Technology of China, China

Yong Shi
CAS Research Center on Fictitious Economy and Data Sciences, China & University of Nebraska at Omaha, USA

INTRODUCTION

Multiple criteria optimization seeks to simultaneously optimize two or more objective functions under a set of constraints. It has a great variety of applications, ranging from financial management, energy planning, sustainable development, to aircraft design. Data mining is aim at extracting hidden and useful knowledge from large databases. Major contributors of data mining include machine learning, statistics, pattern recognition, algorithms, and database technology (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). In recent years, the multiple criteria optimization research community has actively involved in the field of data mining (See, for example: Yu 1985; Bhattacharyya 2000; Francisci & Collard, 2003; Kou, Liu, Peng, Shi, Wise, & Xu, 2003; Freitas 2004; Shi, Peng, Kou, & Chen, 2005; Kou, Peng, Shi, Wise, & Xu, 2005; Kou, Peng, Shi, & Chen, 2006; Shi, Peng, Kou, & Chen, 2007).

Many data mining tasks, such as classification, prediction, clustering, and model selection, can be formulated as multi-criteria optimization problems. Depending upon the nature of problems and the characteristics of datasets, different multi-criteria models can be built. Utilizing methodologies and approaches from mathematical programming, multiple criteria optimization is able to provide effective solutions to large-scale data mining problems. An additional advantage of multi-criteria programming is that it assumes no deterministic relationships between variables (Hand & Henley, 1997).

BACKGROUND

The goal of data mining is to identify hidden, interesting, and useful structures from large databases (Fayyad & Uthurusamy, 2002). Methods and techniques from multiple disciplines, such as machine learning, statistics, pattern recognition, and database technology, have been applied extensively in data mining to extract patterns from data. Recently, the multi-criteria or multi-objective optimization-based methods have been proposed as another option for data mining tasks. For instance, Bhattacharyya proposed a multi-objective model for direct marketing (2000); Francisci and Collard addressed the interestingness measure of dependency rules by formulating the scenario as a multi-criteria problem (2003); and Kou, Peng, Shi, and Chen built a multi-criteria convex quadratic programming model for credit portfolio management (2006).

If a data mining task can be modeled as optimization problems with multiple objective functions, it can be cast into the multi-criteria optimization framework. Many data mining functionalities, such as classification, prediction, and interestingness measure, can be formulated as multi-criteria optimization problems. For example, in multi-criteria optimization context, the classification problem can be stated as one of simultaneously minimizing misclassified points and maximizing correctly classified points. The established methodologies and procedures for solving multi-criteria optimization problems and incorporating the results into the business decision process by the discipline of multi-criteria decision making (MCDM) can be applied to these data mining tasks.

MAIN FOCUS

Currently, the main focuses of multiple criteria optimization in data mining include: model construction, algorithm design, and results interpretation and application.

Model Construction

Model construction refers to the process of establishing mathematical models for multi-criteria data mining problems, which exist in many data mining tasks. For example, in network intrusion detection, the goal is to build classifiers that can achieve not only high classification accuracy, but also low false alarm rate. Although multiple objectives can be modeled separately, they normally can not provide optimal solutions to the overall problem (Fonseca & Fleming, 1995). Furthermore, a model may perform well on one objective, but poorly on other objectives. In this kind of scenario, multiple criteria optimization can be used to build models that can optimize two or more objectives simultaneously and find solutions to satisfy users' preferences.

Algorithm Design

Algorithm design is a set of steps that takes raw data as input and generates solutions as output. Specifically, algorithm design normally includes data preparation, optimization approach, and model assessment.

1. **Data preparation:** Raw data are selected and cleaned according to the requirements of data mining tasks. In addition, data need to be formatted into appropriate forms. Since multiple criteria optimization models can handle only numeric inputs, categorical attributes need to be transformed into numeric types.
2. **Optimization approach:** There are three main approaches to multiple criteria optimization (Freitas, 2004): (i) convert multiple objectives into a single-criterion problem using weight vectors; (ii) prioritize objectives and concentrate on objectives with high priorities, which is called the lexicographical approach; (iii) find a set of non-dominated solutions and allow business users to pick their desired solutions, which is also known as the Pareto approach. Each approach has its advantages and disadvantages. The first approach, reformatting multi-criteria as a single-objective problem, is by far the most popular one in data mining field due to its simplicity and efficiency.
3. **Model assessment:** Results of models, such as accuracy and generality, are assessed according to predefined criteria.

Results Interpretation and Application

Depending on application domains and user preferences, results should be interpreted differently. Take health insurance fraud detection as an example. Multiple criteria optimization can provide class labels and probability scores. Health insurance companies need to process large volumes of records and have limited resources to manually investigate potential fraud records (Peng et al., 2007). In this case, class labels alone, which distinguish normal records from fraud records, are not as useful as the combination of class labels and probability scores that can rank potential fraud records. Data miners should discuss with business users to determine which forms of results can better satisfy business objectives (Chapman, Clinton, Khabaza, Reinartz, & Wirth, 1999).

FUTURE TRENDS

The application of multiple criteria optimization techniques can be expanded to more data mining tasks. So far classification task is the most studied problem in the data mining literature. Multiple criteria optimization could be applied to many other data mining issues, such as data preprocessing, clustering, model selection, and outlier detection (Mangasarian, 1996). Another direction is to examine the applicability of the lexicographical and Pareto approaches in large scale data mining problems.

CONCLUSION

Many data mining tasks involve simultaneously satisfy two or more objectives and thus multiple criteria optimization is appropriate for these tasks in nature. The main components of multiple criteria optimization in data mining include model construction, algorithm design, and results interpretation and application. Model construction builds multi-criteria models; algorithm