Integration of Data Mining and Operations Research

Stephan Meisel
University of Braunschweig, Germany

Dirk C. Mattfeld
University of Braunschweig, Germany

INTRODUCTION

Basically, Data Mining (DM) and Operations Research (OR) are two paradigms independent of each other. OR aims at optimal solutions of decision problems with respect to a given goal. DM is concerned with secondary analysis of large amounts of data (Hand et al., 2001). However, there are some commonalities. Both paradigms are application focused (Wu et al., 2003; White, 1991). Many Data Mining approaches are within traditional OR domains like logistics, manufacturing, health care or finance. Further, both DM and OR are multidisciplinary. Since its origins, OR has been relying on fields such as mathematics, statistics, economics and computer science. In DM, most of the current textbooks show a strong bias towards one of its founding disciplines, like database management, machine learning or statistics.

Being multidisciplinary and application focused, it seems to be a natural step for both paradigms to gain synergies from integration. Thus, recently an increasing number of publications of successful approaches at the intersection of DM and OR can be observed. On the one hand, efficiency of the DM process is increased by use of advanced optimization models and methods originating from OR. On the other hand, effectiveness of decision making is increased by augmentation of traditional OR approaches with DM results. Meisel and Mattfeld (in press) provide a detailed discussion of the synergies of DM and OR.

BACKGROUND

The aim of DM is identification of models or patterns representing relationships in data. The DM process was shaped by two milestones. First, the discovery of relationships in data was established as a multi-step procedure (Fayyad et al., 1996). Secondly, a framework for the core DM step, including the definition of DM tasks was specified (Hand et al., 2001). One of the tasks is exploratory data analysis by means of interactive and visual techniques. Further, descriptive modeling aims at describing all of the data. Predictive modeling comprises classification and regression methods. The task of discovering patterns and rules focuses on particular aspects of the data instead of giving a description of the full data set at hand. Finally retrieval by content addresses the search for similar patterns in text and image datasets.

According to a DM task, a model or pattern structure is selected. The DM algorithm determines an instance of the structure best fitting a given set of target data. The target data set is tailored to the requirements of the DM algorithm in a preprocessing step modifying an initial set of collected data.

OR approaches require identification of decision variables and definition of a pursued objective. A decision model is developed, specifying the set of feasible values of the decision variables and an objective function. A search procedure is then applied in order to determine the optimal values for the decision variables. In case the structure of the decision model does not allow for efficient search methods, the decision model structure is often replaced by heuristic decision rules allowing for deliberate selection of a solution.

The use of optimization methods originating from OR is established since long for at least some of the DM tasks. Mangasarian (1997) discusses the relevance of mathematical programming for large-scale DM problems. A summary of early works in the field is given by Bradley et al. (1999). An application domain

specific article by Padmanabhan and Tuzhilin (2003) gives an overview on the use of optimization for DM in electronic customer relationship management.

However, the multitude of new developments at the intersection of DM and OR does not only comprise more advanced optimization models for DM. Rather, many bare improved OR approaches by integration of DM.

MAIN FOCUS

Regarding recent advances published in literature three types of synergies of DM and OR can be distinguished. On the one hand, application of optimization methods to increase DM efficiency. On the other hand, the use of DM to increase OR effectiveness either by improvement of a decision model structure or by improvement of decision model. Each of the three synergies is discussed below.

Increased Efficiency

Optimization problems may be encountered at several points in both of the major DM steps. Some of the works from literature focus on preprocessing operations. However, most papers are about the use of OR for efficient implementation of descriptive and predictive modeling, explorative data analysis as well as the discovery of patterns and rules.

1. *Preprocessing*—Preprocessing is split into a series of problems of different complexity. Some of these may not be met seriously without the application of OR methods. Examples are the feature subset selection problem, the discretization of a continuous domain of attribute values and de-duplication of information in databases.

Yang and Olafsson (2005) formulate the feature subset selection problem as combinatorial optimization problem. For solution they apply the nested partitions metaheuristic. Pendharkar (2006) considers feature subset selection as a constraint satisfaction optimization problem and proposes a hybrid heuristic based on simulated annealing and artificial neural networks. Meiri and Zahavi (2006) also apply simulated annealing to combinatorial feature subset selection and outperform the traditional stepwise regression method.

Janssens et al. (2006) model a discretization problem as shortest path network and solve it by integer programming.

OR-based de-duplication is considered by Spiliopoulos and Sofianopoulou (2007). They model the problem of calculating dissimilarity between data records as modified shortest path problem offering new insights into the structure of de-duplication problems.

2. *Descriptive Modeling*—The most common technique for the DM task of descriptive modeling is cluster analysis. Boginsiki et al. (2006) formulate the clustering problem as NP-hard clique partitioning problem and give a heuristic allowing for efficient solution.

Saglam et al. (2006) develop a heuristic for solving a mixed integer model for clustering. They show the procedure to outperform the well known k-means algorithm in terms of accuracy. Beliakov and King (2006) formulate the fuzzy c-means algorithm as a bi-level optimization problem and solve it by a discrete gradient method. The algorithm is capable of identifying non-convex overlapped d-dimensional clusters, a property present in only a few experimental methods before.

Kulkarni and Fathi (in press) apply a branch and cut algorithm to an integer programming model for clustering and find the quality of its LP relaxation to depend on the strength of the natural clusters present. Hence, they specify the conditions for an optimal solution to be expected by a branch and cut algorithm. Innis (2006) builds an integer program for seasonal clustering taking into account the time order of data.

3. *Predictive Modeling*—A number of recent publications offer elaborate approaches for the task of predictive modeling by use of OR-methods. Üney and Türkay (2006) present a multi-class data classification method based on mixed-integer programming. Exceeding traditional methods, they introduce the concept of hyperboxes for defining class boundaries increasing both classification accuracy and efficiency.

Jones et al. (2007) present a goal programming model allowing for flexible handling of the two class classification problem. The approach pursues both, maximization of the level of correct classifications and minimization of the level of misclassifications.