Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 187 (2008) 1429–1448
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Operations research and data mining

Sigurdur Olafsson *, Xiaonan Li, Shuning Wu

Department of Industrial and Manufacturing Systems Engineering, Iowa State University, 2019 Black Engineering, Ames, IA 50011, USA

Available online 15 November 2006

Abstract

With the rapid growth of databases in many modern enterprises data mining has become an increasingly important approach for data analysis. The operations research community has contributed significantly to this field, especially through the formulation and solution of numerous data mining problems as optimization problems, and several operations research applications can also be addressed using data mining methods. This paper provides a survey of the intersection of operations research and data mining. The primary goals of the paper are to illustrate the range of interactions between the two fields, present some detailed examples of important research work, and provide comprehensive references to other important work in the area. The paper thus looks at both the different optimization methods that can be used for data mining, as well as the data mining process itself and how operations research methods can be used in almost every step of this process. Promising directions for future research are also identified throughout the paper. Finally, the paper looks at some applications related to the area of management of electronic services, namely customer relationship management and personalization.

Keywords: Data mining; Optimization; Classification; Clustering; Mathematical programming; Heuristics

1. Introduction

In recent years, the field of data mining has seen an explosion of interest from both academia and industry. Driving this interest is the fact that data collection and storage has become easier and less expensive, so databases in modern enterprises are now often massive. This is particularly true in web-based systems and it is therefore not surprising that data mining has been found particularly useful in areas related to electronic services. These massive databases often contain a wealth of important data that traditional methods of analysis fail to transform into relevant knowledge. Specifically, meaningful knowledge is often hidden and unexpected, and hypothesis driven methods, such as on-line analytical processing (OLAP) and most statistical methods, will generally fail to uncover such knowledge. Inductive methods, which learn directly from the data without an a priori hypothesis, must therefore be used to uncover hidden patterns and knowledge.

We use the term data mining to refer to all aspects of an automated or semi-automated process for extracting previously unknown and potentially useful knowledge and patterns from large databases. This process consists of numerous steps such as integration of data from numerous databases,

* Corresponding author. Tel.: +1 515 294 8908; fax: +1 515 294 3524.
E-mail address: olafsson@iastate.edu (S. Olafsson).

0377-2217/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2006.09.023

preprocessing of the data, and induction of a model with a learning algorithm. The model is then used to identify and implement actions to take within the enterprise. Data mining traditionally draws heavily on both statistics and machine learning but numerous problems in data mining can also be formulated as optimization problems (Freed and Glover, 1986; Mangasarian, 1997; Bradley et al., 1999; Padmanabhan and Tuzhilin, 2003).

All data mining starts with a set of data called the training set that consists of instances describing the observed values of certain variables or attributes. These instances are then used to learn a given target concept or pattern and, depending upon the nature of this concept, different inductive learning algorithms are applied. The most common concepts learned in data mining are classification, data clustering, and association rule discovery, and of those will be discussed in detail in Section 3. In classification the training data is labeled, that is, each instance is identified as belonging to one of two or more classes, and an inductive learning algorithm is used to create a model that discriminates between those class values. The model can then be used to classify any new instances according to this class attribute. The primary objective is usually for the classification to be as accurate as possible, but accurate models are not necessarily useful or interesting and other measures such as simplicity and novelty are also important. In both data clustering and association rule discovery there is no class attribute and the data is thus unlabelled. For those two approaches patterns are learned along one of the two dimensions of the database, that is, the attribute dimension and the instance dimension. Specifically, data clustering involves identifying which data instances belong together in natural groups or clusters, whereas association rule discovery learns relationships among the attributes.

The operations research community has made significant contributions to the field of data mining and in particular to the design and analysis of data mining algorithms. Early contributions include the use of mathematical programming for both classification (Mangasarian, 1965), and clustering (Vinod, 1969; Rao, 1971), and the growing popularity of data mining has motivated a relatively recent increase of interest in this area (Bradley et al., 1999; Padmanabhan and Tuzhilin, 2003). Mathematical programming formulations now exist for a range of data mining problems, including attribute selection, classification, and data clustering. Metaheuristics have also been introduced to solve data mining problems. For example, attribute selection has been done using simulated annealing (Debuse and Rayward-Smith, 1997), genetic algorithms (Yang and Honavar, 1998) and the nested partitions method (Olafsson and Yang, 2004). However, the intersection of OR and data mining is not limited to algorithm design and data mining can play an important role in many OR applications. Vast amount of data is generated in both traditional application areas such as production scheduling (Li and Olafsson, 2005), as well as newer areas such as customer relationship management (Padmanabhan and Tuzhilin, 2003) and personalization (Murthi and Sarkar, 2003), and both data mining and traditional OR tools can be used to better address such problems.

In this paper, we present a survey of operations research and data mining, focusing on both of the abovementioned intersections. The discussion of the use of operations research techniques in data mining focuses on how numerous data mining problems can be formulated and solved as optimization problems. We do this using a range of optimization methodology, including both metaheuristics and mathematical programming. The application part of this survey focuses on a particular type of applications, namely two areas related to electronic services: customer relationship management and personalization. The intention of the paper is not to be a comprehensive survey, since the breadth of the topics would dictate a far lengthier paper. Furthermore, many excellent surveys already exist on specific data mining topics such as attribute selection, clustering, and support vector machine. The primary goals of this paper, on the other hand, are to illustrate the range of intersections of the two fields of OR and data mining, give some detailed examples of research that we believe illustrates the synergy well, provide references to other important work in the area, and finally suggest some directions for future research in the field.

2. Optimization methods for data mining

A key intersection of data mining and operations research is in the use of optimization algorithms, either directly applied as data mining algorithms, or used to tune parameters of other algorithms. The literature in this area goes back to the seminal work of Mangasarian (1965) where the problem of separating two classes of points was formulated as