ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computers & Operations Research ▮▮▮ (▮▮▮) ▮▮▮–▮▮▮

computers & operations research

www.elsevier.com/locate/dsw

Editorial

Introduction to operations research and data mining

Over the past several years, the field of data mining has seen an explosion of interest from both academia and industry. Data mining is an interdisciplinary field and draws heavily on both statistics and machine learning. In these two areas, such problems as learning how to classify data and finding natural clusters of data have been studied extensively for decades. Furthermore, the earliest mathematical programming formulations of both data classification [1] and data clustering [2] date back almost 40 years. Building on this early work, the recent growing interest in data mining has been paralleled by a similar growth of research in optimization for data mining [3,4], and the operations research community has the potential to continue to contribute significantly to this field.

Data mining involves inductively learning new and interesting patterns from examples or data instances contained in large databases. The methods applied to this task draw on a long history, but the resurgence of interest in such inductive methods for data analysis is largely driven by the fact that databases in modern enterprises have become very large. In such databases, what constitutes useful information is often unknown and hidden, and traditional means of analysis that are hypothesis driven, e.g. on-line analytical processing (OLAP) and many statistical methods, may fall short in transforming such data into relevant knowledge. This motivates the use of inductive data mining methods, which learn directly from the data without an a priori hypothesis, and can scale up to be applied to very large databases.

The process of data mining is not limited to inductive learning, and the data must be preprocessed before a learning algorithm is applied. This usually includes such steps as data cleaning and data reduction. Data cleaning may include accounting for missing values, and determining how to deal with noisy and inconsistent data. Data reduction may be done along both the attribute and instance dimensions, that is, either reducing the number of variables (attributes) or data points (instances). Once the data have been prepared, inductive learning most commonly involves learning one of the three concepts: classification, clustering, or association rule discovery. Classification involves learning a model that can discriminate between the values (classes) of a particular target attribute called the class attribute. This is a supervised learning task, since the training data from which the algorithm learns is labeled, that is, the class values for the target attributes are known. If there is no such class attribute, two types of unsupervised learning are most commonly used. Data-clustering algorithms aim to discover natural groupings, or clusters, of instances, whereas association rule discovery aims to discover relationships between the attributes. Once the patterns have been obtained, they must be validated and then the knowledge learned can be implemented.

0305-0548/$ - see front matter
doi:10.1016/j.cor.2005.01.019

The intersection between operations research and data mining is quite broad and this is illustrated by the papers in this focused issue. The first three papers use optimization methods for important exploratory and preprocessing steps in the data mining process. As in most other data analysis, visualization of the data plays an important role in data mining, but since the data to be mined is usually high dimensional, that is, there are many attributes, producing useful visualizations is a very challenging problem. An important approach is to map the data to two or three dimensions, which can then be easily visualized, in a manner that optimally preserves the structure of the data. In the first paper of this focused issue, Abbiw-Jackson et al. [5] show how the problem of mapping high-dimensional data to two or three dimensions for visualization can be formulated as a combinatorial optimization problem, namely a quadratic assignment problem, and present an effective heuristic for solving the problem.

High dimensionality poses difficulties not only for visualization, but also for the subsequent use of inductive algorithms for learning patterns. It is therefore common in data mining to preprocess the data by selecting which attributes or features should be used and which attributes should be discarded because they are irrelevant or redundant. As is discussed in [6], this is inherently a combinatorial optimization problem. Addressing it as such, the authors present a new metaheuristic for solving the problem, and evaluate its performance. A particularly novel aspect of their approach is the adaptive use of random sampling of instances, which makes the algorithm more scalable to large databases.

Another common preprocessing step in data mining is the discretization of data, that is, transforming continuously valued attributes into intervals of discrete categorical values. This process is important because many learning algorithms are better able to handle discrete data, and since the process also reduces the number of possible values for an attribute, this can be thought of as another form of data reduction. In [7], the authors formulate this problem as the shortest path optimization problem with a given discretization cost as objective function, and show that this optimization-based approach compares favorably to other well-known methods for discretization.

The next two papers deal with actual inductive learning methods for data mining, that is, methods that take the (usually preprocessed) data and learn a pattern. Specifically, the two papers deal with how to classify the data according to the possible values of a specific attribute called the class attribute. In [8], the author formulates the classification problem as an optimization problem where the objective is to maximize the number of correctly classified instances. The optimization problem is then solved using a simulated annealing algorithm. Also dealing with classification of data, Kros et al. [9], consider the use of neural networks for this task. In particular, they focus on the issue of missing and noisy data, a very common issue in many data mining applications, and evaluate what effect this has on the performance of the neural network.

The first five papers illustrate how operations research-related methodology is applied to solve data mining problems. The last three papers focus on the other side of the intersection of operations research and data mining, namely the application of data mining to a variety of problems. In [10], the authors show how data mining can be used to mine bills-of-materials in engineering design. Another application that has been well studied in the OR literature, is addressed by Wu [11]. In this paper, the author uses frequent itemset mining, which is part of the association rule discovery in data mining, to identify small subsets of items in a warehouse that satisfied majority of orders. This is important because with this knowledge, efficiency can be improved by assigning the discovered subset of items to a single zone within the warehouse. In the final paper, Boginski et al. [12] use a graph theoretic approach to mine-stock market data. In this paper, the stock market data is represented as a network and the authors study characteristics of this graph, and in particular, how it evolves over time.

The papers in this focused issue provide a very nice collection of articles illustrating the many ways in which operations research and data mining intersect, and I hope that they serve to further motivate members of the operations research community to contribute to this important field.

I would like to thank all of the authors who submitted their work for consideration in this focused issue, as well as the many referees who provided constructive comments and suggestions, and helped improve the quality of the eight papers that were eventually selected. Finally, I would like to thank the editor, Gilbert Laporte, for his support and encouragement of this focused issue.

References

[1] Mangasarian OL. Linear and nonlinear separation of patterns by linear programming. Operations Research 1965;13: 444–52.
[2] Rao MR. Cluster analysis and mathematical programming. Journal of the American Statistical Association 1971;66: 622–6.
[3] Bradley PS, Fayyad UM, Mangasarian OL. Mathematical programming for data mining: formulations and challenges. INFORMS Journal on Computing 1999;11:217–38.
[4] Padmanabhan B, Tuzhilin A. On the use of optimization for data mining: Theoretical interactions and eCRM opportunities. Management Science 2003;49(10):1327–43.
[5] Abbiw-Jackson R, Golden B, Raghavan S, Wasil E. A divide-and-conquer local search heuristics for data visualization. Computers and Operations Research 2005; in this issue.
[6] Yang J, Ólafsson S. Optimization-based feature selection with adaptive instance sampling. Computers and Operations Research 2005; in this issue.
[7] Janssens D, Brijs T, Vanhoof K, Wets G. Evaluating the performance of cost-based discretization versus entropy- and error-based discretization. Computers and Operations Research 2005; in this issue.
[8] Pendharkar PC. A data mining constraint satisfaction optimization problem for cost effective classification. Computers and Operations Research 2005; in this issue.
[9] Kros JF, Brown ML, Lin M. Effects of the neural network s-sigmoid function on KDD in the presence of imprecise data. Computers and Operations Research 2005; in this issue.
[10] Romanowski CJ, Nagi R, Sudit M. Data mining in an engineering design environment: OR applications from graph matching. Computers and Operations Research 2005; in this issue.
[11] Wu C. Application of frequent itemset mining to identify a small subset of items that can satisfy a large percentage of orders in a warehouse. Computers and Operations Research 2005; in this issue.
[12] Boginski V, Butenko S, Pardalos PM. Mining market data: a network approach. Computers and Operations Research 2005; in this issue.

S. Ólafsson
Department of Industrial Engineering, Iowa State University, Ames, IA 50010, USA
E-mail address: olafsson@iastate.edu.