European Journal of Operational Research 221 (2012) 469–479

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Invited Review

# Synergies between operations research and data mining: The emerging use of multi-objective approaches

David Corne [a], Clarisse Dhaenens [b,c], Laetitia Jourdan [b,c,*]

[a] Heriot-Watt University, Scotland, United Kingdom
[b] Université Lille 1, Laboratoire d'Informatique Fondamentale de Lille, UMR CNRS 8022, Cité Scientifique, Bâtiment M3, 59655 Villeneuve d'Ascq cedex, France
[c] INRIA Lille-Nord Europe, Parc Scientifique de la Haute Borne, 40 avenue Halley, 59650 Villeneuve d'Ascq, France

ARTICLE INFO

*Article history:*
Received 23 September 2011
Accepted 23 March 2012
Available online 30 March 2012

*Keywords:*
State-of-the-art
Operations research
Knowledge-based systems
Knowledge discovery
Multi-objective optimization

ABSTRACT

Operations research and data mining already have a long-established common history. Indeed, with the growing size of databases and the amount of data available, data mining has become crucial in modern science and industry. Data mining problems raise interesting challenges for several research domains, and in particular for operations research, as very large search spaces of solutions need to be explored. Hence, many operations research methods have been proposed to deal with such challenging problems. But the relationships between these two domains are not limited to these natural applications of operations research approaches. The counterpart is also important to consider, since data mining approaches have also been applied to improve operations research techniques. The aim of this article is to highlight the interplay between these two research disciplines. A particular emphasis will be placed on the emerging theme of applying multi-objective approaches in this context.



## 1. Introduction

Data mining (DM) has recently seen an explosion of interest in many fields of applications, owing to the increasing amount of data available, and the growing understanding that deeper analyzes are far more valuable than simple summary statistics. Data mining is an inductive (not deductive) process. Its aim is to infer knowledge that is generalized from the data in the database. This process is generally not supported by classical DataBase Management systems. Data mining problems raise interesting challenges for several research domains, such as statistics, information theory, databases, machine learning, data visualization, and also for operations research (OR), since very large search spaces of solutions need to be explored. Hence, for several years, numerous research efforts using operational research methods to solve data mining problems have been reported, and several reviews of such approaches have been published [68,91,92]. However, the synergy between operations research (OR) and data mining (DM) is not a one-way street; as described by Meisel and Mattfeld, three kinds of synergies may be achieved [82]: 1/OR can contribute to the efficiency of DM techniques, 2/DM can increase the number of problems in which OR can be applied by means of a less rigorous model building process, 3/finally, increased system performance can result from complementary uses of these two research domains.

In this article we will use a simpler categorization of the synergies between DM and OR, which emphasizes two types of interaction, in terms of how OR and DM can contribute to each other. Hence, the first point of view (similar to the first of Meisel and Mattfeld's synergies) is to analyze how OR can contribute to the efficiency of DM techniques. The second point of view looks at how DM can contribute to OR methods. In our view, the second synergy of Meisel and Mattfeld, concerned with using DM techniques to better capture the structure of the underlying system, may be merged into our second type of DM/OR interaction, since it yields the same overall result of enhancing OR via deployment of DM.

Our first point of interest is to analyze how OR can be useful in the challenges faced by applications of DM. In other words, how OR approaches can contribute in helping DM difficult problems. We will see in this review that there are several answers, using several approaches, which all tend to center on using OR to deal with one or other NP-hard optimization problem that arises in a DM task. In particular, metaheuristics have been widely used in this context, and several books dedicated to metaheuristics and data mining have been published [26,35]. Meanwhile, multi-objective metaheuristic approaches are increasingly also being proposed in this context [61,59]. Thus, this article will pay a particular attention to this multi-objective aspect and methods that have been proposed for that. Therefore, the notion of quality criterion related to the objective function, for example, will be discussed.

* Corresponding author at: Université Lille 1, Laboratoire d'Informatique Fondamentale de Lille, UMR CNRS 8022, Cité Scientifique, Bâtiment M3, 59655 Villeneuve d'Ascq cedex, France. Tel.: +33 0 3 59 57 78 81.
*E-mail addresses:* david.corne@gmail.com (D. Corne), clarisse.dhaenens@lifl.fr (C. Dhaenens), laetitia.jourdan@inria.fr (L. Jourdan).

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.03.039

The second fundamental question, when synergies between OR and DM are under analysis, is to understand how DM techniques can help OR methods. Even though this thread of research is less studied, significant such work is emerging [64]. The objectives of such a synergy may be for example, to either improve the quality of results obtained by OR approaches, or to speed up the execution of algorithms.

The aim of this review is to provide interesting pointers to how OR and DM can enrich each other. The remainder is organized as follows: the second section is designed to present to the OR community a short introduction to 'knowledge discovery', in order to help define the scope of this very general term and to make this article be self-content. It will describe the main data mining tasks and the principal and classical algorithms in this field. Section 3 will then deal with the first question: how operations research can help data mining. Section 4 is dedicated to the other side of the coin: how data mining may be useful for operations research techniques. In both Sections 3 and 4, a particular emphasis will be given to multi-objective models and methods. Section 5 will conclude the review and will suggest some interesting research perspectives for both communities.

## 2. Knowledge discovery: a brief introduction

'Knowledge discovery and data mining' (KDD) is a phrase that describes a large area of research concerned with discovering and exploiting the considerable amount of potentially useful knowledge that is often 'hidden' in databases. Such knowledge is regarded as hidden, since standard statistical techniques simply fail to find it, and the discovery of interesting rules or associations tends to be infeasible with exact algorithms owing to the size of the database. Data mining is at the heart of the KDD process (see Fig. 1). It allows us to extract useful information from large data sets or databases in reasonable time, usually by employing approximate algorithms. This discipline lays at the intersection of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and similar domains. The reader interested in knowledge discovery in general, or datamining in particular, can refer to some of the following books [46,67,80]. In this section we focus on data mining tasks and algorithms (see Fig. 2). Each of these tasks are briefly described in this section, presenting their goals, and the algorithms most commonly associated with them.

### 2.1. (Supervised) classification

The aim of this task is to build a model that predicts the value of one variable from the known values of other variables. In classification, the variable being predicted, called the 'class', is categorical, and the task becomes regression when the predicted variable is numerical. Several approaches have been proposed. We will expose here briefly some of them, but to have a more general view of these methods, the reader may, for example, refer to [73].

#### 2.1.1. K-nearest-neighbor

K-nearest-neighbor (K-NN) classification is one of the most fundamental and simple classification methods [20]. It is suitable for a classification study when there is little or no prior knowledge about the distribution of the data. The N-nearest-neighbor algorithm relies on the distances between examples in the feature space: an object is assigned to the most common class shared by its K nearest neighbors. It can be useful to weight the contribution of the neighbors, so that nearer neighbors contribute more to the average than more distant ones. If $K = 1$, then the object is simply assigned to the class of its nearest neighbor. The neighborhood is defined by the distance metric used, which is commonly the Euclidean distance.

#### 2.1.2. Decision trees

Decision trees (or classification trees) are very popular for classification, since they are simple to understand and to interpret. A decision tree is built through a process known as *binary recursive partitioning*. This process recursively splits (or 'partitions') the data into groups. At each stage, the splitting is realized in a way that maximizes a score function for the split. The score function is chosen so that it favors the degree to which each individual group contains datapoints that are all of the same class. The main difference between different decision tree construction algorithms is the score function that is used to guide the splitting process. For example, *Information gain* is used by the ID3, C4.5 [101] and C5.0 tree generation algorithms, and is based on the concept of entropy used in information theory. *Gini impurity* is used by the CART algorithm [13] and measures how often a randomly chosen datapoint from the group would be incorrectly labeled if it was randomly labeled according to the distribution of labels within the group.

#### 2.1.3. Naive Bayes

The Naive-Bayes classifier uses a probabilistic approach based on applying Bayes' theorem with strong (Naive) independence assumptions. To assign the class to a sample, it computes the conditional probabilities of different classes given the values of the features, and predicts the class with the highest conditional probability. Naive-Bayes is simple and can be applied to multi-class classification problems, but it assumes independence between variables, which is typically untrue (i.e. it is a naive assumption). In spite of its simplified assumptions, Naive-Bayes classifiers often work very well in many complex real-world situations.

#### 2.1.4. Neural networks

Artificial neural networks (ANN) are widely used for classification and are a promising alternative to various conventional classification methods [124]. An artificial neural network is essentially a

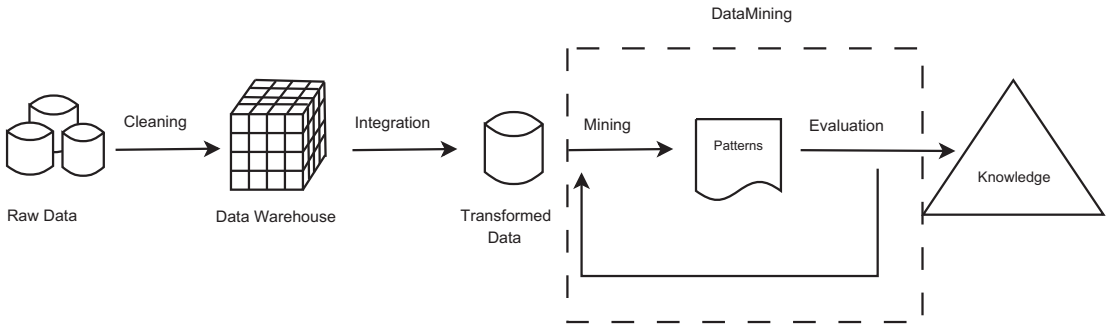


**Fig. 1.** An overview of the KDD process.