Editorial

Data mining and operational research: techniques and applications

Journal of the Operational Research Society (2009) **60**, 1043–1044. doi:10.1057/jors.2009.16

Data mining (DM) involves the use of a suite of techniques that aim to induce from data, models that meet particular objectives. DM algorithms are built on a range of techniques, including information theory, statistics, linear and non-linear models, AI, meta-heuristics. Within the context of data analysis methods, data mining can be considered to be an exploratory knowledge discovery approach to be contrasted with a confirmatory approach in which a hypothesis is specified and the validity of the hypothesis is tested against the data.

DM models can be designed to find new useful patterns within the data that can be exploited. For example, an insurance company will already be aware that young men in fast cars pose a high risk (as do older women). Such a pattern in the data will be of no interest to them. However, they may be interested if they were to find that young men in fast, classic cars are much less of a risk. They could then develop pricing and marketing strategies to exploit this newly discovered pattern.

In the context of contemporary organizations, the major motivating factors for the interest in, and application of, data mining techniques can be categorized as: changed business environment (eg pressure on traditional marketing techniques, shorter time to market, shorter product life cycles, and increased competition and business risks), drivers (eg customer, competition, and data asset), and enablers (eg data flood in organizations, new information technology solutions in data mining, data warehousing, and hardware).

Many organizations routinely collect large amounts of data on their clients, employees, and suppliers. Factories collect data on their machinery and productivity. Hospitals collect data on patients, doctors, medications, and procedures. These data are often untidy, incomplete, and sometimes erroneous and yet, if used properly, it can be a valuable asset for management and effective organizations can find ways to exploit it. Predicting fraudulent behaviour in a customer database, for example, will generally require highly accurate models even if they can only be applied to a small proportion of the database. On the other hand, if the organization is planning a targeted marketing campaign and wishes to predict which customers might buy a certain product, the accuracy of the model can be much lower.

Both operational research (OR) and DM are aimed at the development and application of models that can improve the quality of organizational decision-making process. While OR can be considered to be analysis-driven and DM can be considered to be data-driven, both approaches involve the development and use of algorithms that have their genesis in mathematics, statistics, and other techniques of mathematical and other systems sciences.

Over the last two to three decades, there has been increased interest in DM and its techniques by practitioners and researchers not only in OR but also in a wide range of other research and practitioner communities. It has become increasingly clear that OR techniques can be used to enhance the DM process and, more recently, there has been the recognition that DM techniques can be used to enhance OR algorithms and practice. This special issue of *JORS* followed a call for papers at the intersection on DM and OR. Its emphasis is on real, applied or applicable work rather than simply theory. Twenty-five papers were submitted and, following stringent refereeing, the best of these appear in this issue.

The papers in this special issue can be considered to fall into two major categories: (1) *OR to DM*: the use of insights and/or techniques from OR to design effective DM algorithms and/or processes; (2) *DM to OR*: the use of insights and/or techniques from DM to more effectively address what could be considered to be decision-making problems in the OR application domain.

The *OR to DM* category consists of six papers. Many DM problems, including feature (or variable) selection and clustering can be modelled as combinatorial optimization (CO) problems and this approach is used by several authors. Multi-criteria and other OR techniques are used by other authors.

- When databases have a large number of fields, the efficiency and effectiveness of learning algorithms can be enhanced if a suitable subset is selected before the learning phase. Too large a subset can result in too large a search space with resulting poor performance; too small a subset can lose vital information and lead to a poor quality classifier. Feature selection, which is one of the pre-processing activities in the data mining process, has the objective of selecting the best subset of the variables. This problem can be modelled as a CO problem; a heuristic search procedure for its solution using nested partitions is presented in the paper *Near-optimal feature selection for large databases* by Yang and Olafsson.
- Clustering is a very commercially important topic in DM wherein similar records are clustered together. When

applied to customer databases, clustering can help companies to better understand their market and thereby to formulate future policy. A partitional cluster of the database assigns every record to one and only one cluster. Providing a measure of a 'good' cluster is no easy task but, provided this can be achieved, finding the best cluster becomes a CO problem. One difficulty of using such an approach for large databases is that of scalability. Sampling is usually used to overcome the difficulties but this will inevitably introduce noise into the process. In the paper *An optimization approach to partitional data clustering* by Kim *et al*, an optimization-based clustering approach is described that uses algorithms specifically designed to handle such noise. Mastrogiannis *et al* in *CLE.KMODES: a modified k-modes clustering algorithm* also focus on the clustering problem but for data sets that have categorical attributes. Rather than using CO modelling, they adopt and adapt concepts and techniques of multicriteria outranking relations theory to the design of a more effective clustering algorithm for such data sets.

- Metaheuristics such as genetic algorithms (GAs), simulated annealing (SA), and tabu search (TS), widely used in CO, are also standard techniques in DM. Hybrid techniques found to be so effective in OR are now also becoming more common in DM. One of the more effective techniques in machine learning is that of support vector machines (SVM). In the paper *A two-stage classification procedure for near-infrared spectra based on multi-scale vertical energy wavelet thresholding and SVM-based gradient-recursive feature elimination* by Cho *et al*, SVM is applied to the classification of near-infrared spectra but only following a multi-scale vertical energy thresholding procedure. In the paper *support vector learning with an evolutionary engine* by Stoean *et al*, SVM learning is adopted within an evolutionary approach that enables the coefficients of the decision function to be acquired explicitly. Computational results suggest that the approach has exciting potential. Evolutionary approaches are also hybridized in a GA-neural network classifier. In Pendharkar's paper *Misclassification cost minimizing fitness functions for genetic algorithm based artificial neural network classifiers*, approaches are studied that can be used to formulate a misclassification cost minimizing GA fitness function in this context.

Our second category, *DM to OR*, comprises three papers. Two of the papers involve the utilization of an undirected learning technique (viz. association rule mining) while the

third involves the use of two directed learning techniques (viz. logistic regression and decision tree induction).

- In many countries, cost increases in the healthcare sector outstrip that of other areas. There is thus an interest in reducing cost while providing high-quality service to patients in an environment that often involves high uncertainty. With regard to outpatient treatment facilities, it is therefore important to match demand with capacity as a mismatch will affect either cost or the quality of service. In *A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling*, Glowacka *et al* utilize the association rule mining technique along with simulation to address an operational problem of outpatient no-shows that occur in the healthcare sector.
- Given the highly competitive marketplace in which modern business organizations operate, there has been an increasing interest in effective and efficient supply chain management. In the paper *Potential use of FP-growth algorithm for identifying competitive suppliers in SCM*, Lin presents a hybrid approach that involves the use of association rule mining to address the issue of identifying the primary and secondary suppliers in a supply chain management system.
- As is to be expected, business organizations are very interested in retaining value-adding customers. However, in a competitive marketplace, the interests of the business organization and its customer do not always coincide, and so churning may occur as some customers abandon the given business relationship to pursue others that are viewed as being more advantageous to them. Modern business organizations are thus interested in minimizing churning. In *Domain knowledge integration in data mining using decision tables: case studies in churn prediction*, Lima *et al* present an approach that involves the use of multiple data mining techniques (logistic regression and decision trees) to address the issue of churning within the context of customer relationship management.

As guest editors of this issue, we would like to express our thanks to our authors for some stimulating papers and our thanks to our referees for their helpful and timely responses.

Kweku-Muata Osei-Bryson
Virginia Commonwealth University, Richmond, VA
Vic J Rayward-Smith
University of East Anglia, UK