Integrating and streamlining data chain management

John Wang

Department of Management and Information Systems,
Montclair State University,
Montclair, NJ 07043, USA
Fax: +1-973-655-7678
E-mail: j.john.wang@gmail.com

Biographical notes: John Wang is a Full Professor at the Montclair State University. Having received a scholarship award, he came to the USA and completed his PhD in Operations Research from Temple University. He has published over 100 refereed papers and six books. He has also developed several computer software programs based on his research findings. He is the Editor of the *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (six-volume) and the Editor of the *Encyclopedia of Data Warehousing and Mining,* 1st (two-volume) and 2nd edition (four-volume). He is also the Editor-in-Chief of *International Journal of Data Analysis Techniques and Strategies*. His long-term research goal is on the synergy of data mining, operations research and cybernetics.

Facilitating the transformation process from data to information to knowledge is a paramount issue for every organisation. Many companies are being flooded with data and volumes of conflicting information, but with limited real usable knowledge. Statisticians focus on accuracy; operation researchers target optimisation; intelligent machine learners advocate black box solutions; database administrators emphasise completeness, to name a few. However, rarely should a process be looked at from limited angles or in parts. Current isolated islands of data mining, data modelling and data management (DMMM) should be connected; an integrated and systematic 'union' of DMMM is long overdue; the synergy of DMMM is strong enough; a single one-stop shop is much needed by the community.

IJDMMM concentrates on the following seven integrations:

- integration of DMMM
- integration of statistics, machine learning and databases
- integration of each element of data chain management (DCM)
- integration of different types of information
- integration of algorithms in software
- integration from data pre-processing to post-processing
- integration between theory and applications.

As an analogy, supply chain management is a major milestone in OR/MS field. In the past, each decision-making model, such as forecasting, inventory, transportation, waiting line, assignment, warehousing, maintenance, etc., were applied separately and individually. As a result, too much attention could be paid to algorithms from a small component of a whole supply chain, while at the same time totally discounting the integration and collaboration among different suppliers, partners and customers. The same is true here with DMMM. Since little attention is usually given to the entire process and sequence of what we call DCM, it is time to look at broader ranges, larger chains and bigger pictures.

There are six articles in this issue. ‘Multi-label large margin hierarchical perceptron’ is an excellent paper that introduces a new type of multi-label learning procedure. Woolam and Khan investigate data that has two unique properties: it is multi-label and the labels belong to a structured hierarchy. The authors extend the hierarchical perceptron algorithm and demonstrate good results. The hierarchical perceptron calculates a loss to perform an update. This loss is decided on a one-to-one basis between two labels. This algorithm builds on powerful large margin principals developed in recent years. Derivation of the algorithm is followed by rigorous analysis. A new loss function is created to perform updates on a many-to-many basis between groups of labels. Empirical results against two separate databases demonstrate great performance. Their thorough analysis utilises cutting edge hierarchical error metrics.

In the paper titled ‘Completing missing views for multiple sources of web media’, Subramanya, Wang, Li and Liu proposed a novel method for modelling the problem of inference with missing information for applications with multiple information sources. The information sources are called views and they typically provide complementary information for the underlying problem. This novel modelling naturally supports a class of applications where multiple sources of information are acquired and processed for integration, with each source consisting of its own set of features. The model enabled a systematic approach to completing the missing views in order to achieve optimal inference using all relevant information sources. The authors employed the canonical correlation analysis as the basic tool in designing their algorithm for missing view prediction, recognising that there is correlation among the multiple information sources of the underlying problem. Two interesting examples of practical significance were given to demonstrate and evaluate their approach: automatic web page classification and automatic photo tag recommendation. They presented empirical findings and insights obtained on challenging test datasets.

The contribution by Hühn and Hüllermeier addresses the problem of ordinal classification, which has recently received increasing attention in the machine learning and data mining field. In ordinal classification problems, the finite set of class labels is endowed with a total order. Learning algorithms trying to exploit this order information essentially rely on the assumption that the ordinal class structure is also reflected in the topology of the instance space. As suggested by the title of the paper, asking ‘Is an ordinal class structure useful in classifier learning?’ the authors critically question the validity of this arguably plausible yet non-evident assumption. Since it eventually legitimates research on dedicated learning algorithms, it is indeed surprising that this assumption has not been analysed more closely so far. Concretely, Hühn and Hüllermeier investigate, on an experimental basis, the question whether or not existing methods for ordinal classifier learning are truly able to exploit class order information. To this end, they propose a simple though elegant approach the basic idea of which is related to permutation tests. Provided the answer to the above question is affirmative, the performance of an algorithm should deteriorate when re-ordering the classes in a random way. The results of their study essentially confirm that existing learning methods are able to exploit an ordinal class structure. Moreover, they reveal several properties of ordinal classification methods that are mainly responsible for their efficacy.

By demonstrating how data mining-based approaches prove useful for discovering relevant knowledge from event history data describing life courses, ‘Mining event histories: a social science perspective’ opens social science as a promising application domain for data mining. Based on a long experience in individual longitudinal data analysis, Gilbert Ritschard and his collaborators provide a unique synthetic view of the multiple ways social scientists deal with time stamped sequential data. Their double expertise in event histories analysis and data mining permits them to clearly pinpoint where data mining-based methods may fruitfully complement classical statistical and data analysis techniques. Their argument is convincingly illustrated on real data from the 2002 biographical survey conducted by the Swiss Household Panel. The methods discussed range from survival trees to the mining of discriminating subsequence’s and for each of them the authors perspicaciously single out further developments necessary to cope with the specific expectations of the social scientist. This easily accessible paper provides a real breakthrough in sequential data analysis, both for social scientists and for data mining experts. By reading this paper, the former will discover how they can significantly enrich their analyses while the latter will find new perspectives for the data mining of sequential data.

In their paper titled ‘Combining multiple classifiers for wrapper feature selection’, Chrysostomou, Chen and Liu focus on the idea of combining more than one classifier for wrapper feature selection. This idea is rather novel since existing wrapper feature selection methods only use a single classifier to select relevant features. Based on this novel idea, the authors introduce the wrapper-based decision tree method (WDT), which combines multiple classifiers to select mutually agreed and unbiased relevant features and also visualise relationships among selected features by using decision trees. Since WDT has the novel ability of combining multiple classifiers for feature selection, the authors explore the effects of using different numbers of classifiers and classifiers of different nature on feature selection results. Exploring these two issues revealed interesting results. The results showed that few classifiers selected many relevant features whereas many classifiers selected fewer relevant features. In addition, it was found that decision tree classifiers selected higher number of features and features that generated accuracy levels much higher than other classifiers that were used. Overall, the authors of this paper make two significant contributions to the area of feature selection. First, they develop the WDT method, which can overcome the limitation of existing wrappers by combining multiple classifiers for feature selection. Second, they provide useful insight into the effects of using different number and nature of classifiers, which can be utilised by experts in the field for improving feature selection results.

Spatial data mining is related to the extraction of interesting and useful but implicit spatial patterns from spatial data. It has received considerable attention in the recent years but most of the works in this area are simple adaptations of conventional data mining tools and techniques, which do not recognise the uniqueness of the spatial dimension. The position paper on ‘A relational perspective on spatial data mining’ by Donato Malerba carefully examines the issues related to discovering knowledge from spatial data and advocates a multi-relational data mining approach in order to face these issues systematically. The paper reports an extensive list of significant challenges for researchers interested in applying multi-relational data mining methods to spatial data.

The author’s position is well-founded in the literature and based on his long-term experience with both spatial and relational data mining.

DCM is a newly integrated area. However, it needs to be continuously updated by means of real time practice. The entire system also requires substantial improvement. Hence, DMMM should be interlinked and interconnected with utmost utility of data towards optimum system effectiveness.

IJDMMM aims to provide a professional forum for formulating, discussing and disseminating these solutions, which relate to the design, development, deployment, management, measurement and adjustment of data warehousing, DMMM and other data analysis techniques. They should form a common ground on which a DCM system can be built, shared and supported by professionals from different disciplines.

IJDMMM provides a communication channel between practitioners and academics to discuss problems, challenges and opportunities in all aspects of data mining, data modelling, data analysis and data management. The process of knowledge creation can include multiple components, including data acquisition/collection, data accumulation, data maturation, data selection and refining, data storage and retrieval, data pre-processing, data analysis and validation, data maintenance and data presentation, data warehousing, data mining and/or modelling and information extraction. Therefore, DCM cannot be isolated, separated, broken, or ignored. It is an integrated and interconnected process.

IJDMMM publishes research papers, innovative ideas, reviews, surveys, debates, reports, case studies, position notes, practice comments, book reviews, commentaries and news. Special issues devoted to important topics in data mining, modelling and management will occasionally be published.

Hopefully, *IJDMMM* and *IJDATS* will be able to share a manager’s burdens, meet a practitioner’s challenges, explore an executive’s opportunities, and realise an entrepreneur’s dreams.

Together, let us celebrate the ‘birth’ of *IJDMMM,* nurture its ‘growth’, contribute to its ‘strength’ and protect its ‘health’.