On Business-Oriented Knowledge Discovery and Data Mining

Wu Hao
Liuzhou Railway VocationalTechnology College
Liuzhou, China, 545007
E-mail: lywh88@126.com

***Abstract*—This paper will discuss issues in data mining and business processes including Marketing, Finance and Health. In turn, the use of KDD in the complex real-world databases in business and government will push the IT researchers to identify and solve cutting-edge problems in KDD modelling, techniques and processes. From IT perspectives, some issues in economic sciences consist of business modelling and mining, aberrant behavior detection, and health economics. Some issues in KDD include data mining for complex data structures and complex modelling. These novel strategies will be integrated to build a one-stop KDD system.**

Keywords-Business intelligence; data mining; behavior detection

I. INTRODUCTION

Knowledge Discovery and Data Mining is a subject that researches and develops methods and technology for efficiently discovering valuable knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. This valuable information can include patterns, associations, changes, anomalies and significant structures and models. KDD attempts to formulate, analyse and implement basic induction processes that facilitate the extraction of meaningful information and knowledge from data.

The potential impact on Australian industry of more widespread and effective use of KDD can be inferred from the results of a 2003 IDC study of 40 US and European companies that use predictive analytics KDD:

- projects using predictive analytics KDD yielded a median ROI of 145%;
- the projects used predictive analytics KDD to enhance business processes, and resulted in improved operational decisions;
- predictive analytics KDD was used to tackle problems of greater scope and complexity than those that did not employ KDD; and
- projects using predictive analytics KDD required higher investment levels and yielded higher overall returns over five years.

Whilst considerable progress has been made in the field of data mining, there are a number of very serious limitations to this work. They arise in part from the fact that data mining researchers have been drawn from Computer Science and Mathematics and have formulated problems arising from their conception of the nature of knowledge to be extracted and the types of data encountered. This has led largely to the development of algorithms and their relatively mechanical application in different domains.

What has been lacking is a true intellectual engagement by the data mining researchers and the data users and domain experts in the different fields in which data mining seeks to be applied. This is demonstrated by the fact that in several fields, particularly in Economic Sciences and Business, domain experts often develop the domain models without recourse to KDD techniques. Thus the Gauch model for derivatives relies on mixture density models while the Black–Scholes model relies on a statistical model. On the other hand, these fields by not engaging significantly with the data miners, are missing out on potential enhancements to understanding of their domains and evidence based enrichment of models and theories. Engagement between the data miners and the data users and domain experts will allow them to challenge and expand the precepts and theoretical underpinning in both KDD and business domain areas. This paper seeks to provide the environment to facilitate such engagements.

In this paper we have chosen Economic Sciences as the data users because of the richness of the data, different types of data, and complexity and continuing rapid evolution of the theories and models that are necessary to underpin the domain structure and behavior.

Important elements which have not previously been given enough attention in the generic KDD techniques include: (1) Nature of knowledge; (2) Nature of data, such as relational data, Text, Web, Stream, Multimedia, XML, Spatial-temporal; (3) Data are often incomplete, inconsistent, or conflicting; (4) Data often distributed across databases; (5) Data are in varying degrees of relevance to the central analytical task; (6) In real world, privacy, trust, security and anonymity of data are real issues.

The above limitations pose the new challenges for the data miners. These limitations are hardly overcome without domain experts' Involvements. To circumvent these critical issues, we study strategies for effective and efficient business data mining.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. In Section 3, semi-supervised subspace learning strategies are proposed for high dimensional data. Section 4 presents semi-supervised negative rule learning and outlier detection. Conclusion is given in Section 5.

Published by Atlantis Press, Paris, France.

II. Business Modeling and Mining

Typically Economic Science researchers have built analytical type models focusing on specific (considered important) aspects of agent behaviour (such as utility maximization, learning behaviour). Whilst this research agenda has provided important theoretical insights into the micro behaviour of agents, the connection to understanding macro behaviour still remains a distant goal with such traditional approaches. This section designs key research into the application of KDD methodologies and technologies to creating a kind of Economic Science laboratory environment in which a number of typical research issues can be addressed in a set of three projects. Each of the following three projects is designed to feature key characteristics of the gap between market processes and the machinery developed for mining marketing data.

A. Marketing related research

Data marketers seek to identify and predict profit potential in customers so that the efficiency of markets can be enhanced. However, not all databases contain complete information on customer behaviors and characteristics, particularly in cases where customers can (and do) deal with multiple providers of products and services (e.g., banks) holding additional data about these customers. This paper suggests to investigate ways to analyze multiple datasets to develop more complete information on customer potentials. Other problems involve the need to specify models of behavior (e.g., choice models) that permit better segmentation and more efficient targeting of customers by including much richer sets of actionable customer characteristics and behaviors in models; this requires one to analyze complex datasets including transactional and experimental data with large numbers of individual measures to identify relationships that can be specified in behavioral models that form the foundation for agent-based simulations and optimization methods linking customer behaviors to firm performance and market efficiency.

B. The Modeling of Market Processes

This paper suggests to combine the power of KDD technologies to model economic and financial markets as adaptively evolving systems of interacting heterogeneous agents who can differ along a range of dimensions (risk, expectations etc). It will also be possible to incorporate the biological characteristics of markets as well. This leads to a framework that allows incorporation of many of the key sectors of market economies (production firms, capital markets, labor markets, the market for information and assurance on the information, the government); provisions for applying KDD technologies to calibrate the model to real world data, which can then be used as a kind of laboratory to various kinds of market experiments such as change in tax and other types of policy regimes, different ways of regulating the financial system, the restructuring of regulated monopolies (e.g. electricity suppliers, Telcos), the restructuring of audit markets in the face of enhanced regulation. These approaches can go beyond what is possible with traditional econometric modeling because of the handling of agent heterogeneity, local interactions and consistently linking so many sectors of the market.

C. Learning in Market Environments

A common element in the other projects of this program and more broadly in Economic Science research is that of learning and adaptation. In a social environment of interacting self-interested agents a very important consideration is the way the agents seek to learn about and adapt to this environment. Within the context of agents globally optimizing their own particular objective function (eg expected utility) a range of learning procedures has been developed, such as; neural networks, genetic algorithms; reinforcement learning. The social environments considered in this project will go beyond the individual agent optimizing viewpoint and consider learning schemes in which the strategies of the interacting agents jointly evolve to satisfy possibly several global objectives. This paper suggests to build provisions for incorporating the key characteristics of actual human decision making in the learning schemes that the agents evolve on the basis of their own locally perceived benefits; incorporating the ideas of evolutionary game theory so as to better capture the anomalies (from the point of view of the rational individual agent paradigm) uncovered in laboratory experiments in which human agents display behavior at odds with the predictions of standard paradigms; and calibrating the learning algorithms to empirical decision making data.

D. Detection Technology

Two years ago corporate fraud in the USA was quoted as costing US shareholders in excess of US $600 billion per annum. There is an urgent need to develop measurement strategies to enable governments and corporations to gain reliable measurements of these crimes so that they have a better understanding of the size and extent of these problems and their impact on revenue. This paper suggests solutions to assist governments and corporations to measure accurately fraud and abuse, improve detection rates, reduce miss rates and develop policies to minimise its occurrence.

This paper suggests to use KDD expertise to develop both discovery and detection techniques that will be able to identify new and established patterns of fraud and abuse in social security and tax data and to assist with gaining more accurate measurements of the base rates and financial costs of these crimes. This paper also suggest to draw upon both KDD and econometric technologies. KDD technologies allow management and analysis of huge data sets to discover and detect patterns that may signal fraudulent activity. The econometric methodologies will allow for effective analysis of tax audit cases to identify areas where tax audit policy may be improved. This leads to a better understanding of the various socio-economic characteristics

Published by Atlantis Press, Paris, France.

of individuals and businesses that are not complying with their tax obligations, which otherwise would not be possible with the use of only aggregated data. The benefits from both methodological approaches are very significant for public policy. Therefore, one can build provisions for reducing the current costs associated with detecting fraudulent activities; and strategies for establishing appropriate policy measures to curb such activities in future.

III. HEALTH ECONOMICS

A. Variations in Health Care Use

This paper suggests to identify and quantify the observed variations in health care provisions and the contributing factors. The challenge is not to rely on means and identify outliers but to understand the underlying behaviors and how they result in particular patterns of use and health outcomes. For example, the British GP Harold Shipman, probably the most prolific serial killer, was not identified as an outlier on routinely available mortality data. Therefore, one can carry out general practice for the reasons for falling GP attendances, changes for differential uptake of new MBS items, changes in charging practices, by considering factors associated with the medical practitioners and/or populations served which lead to different patterns of use spatially and over time; pharmaceuticals for variations in prescribing behavior and the extent of substitutability between pharmaceuticals and other services; hospitals and aged care facilities for supply and utilisation of beds by considering factors associated with increasing use of private beds, day only admissions, and spatial differences in bed supply; and insurance for the relationship between insurance status and hospital use by considering the extent of adverse selection, cream skimming and/or healthy selection, and moral hazard.

B. Health Workforce

This paper suggests to quantify the observed variations in health workforce supply and to predict changes in supply. The challenge is not to rely on crude extrapolations from current means but to understand the underlying behaviors and how they affect workforce participation. Therefore, one can perform location for private practitioners, what factors influence choice of practice location (including rural/urban), and for employed staff, the choice of location and type of facility (private or public); choice of specialty for what factors attract and retain professionals to different areas of specializations; and retention and turnover for factors that enhance retention in the nursing workforce, the extent to which medical practitioners respond to incentives for rural practice and whether they stay there.

IV. DATA MINING IN COMPLEX DATA STRUCTURES

This paper tackles data in complex structureas follows.

A. Data Preprocessing

In practice, it has been generally found that data preprocessing takes approximately 80% of the total data engineering effort. Data preprocessing is, therefore, a crucial research topic. This project develops theories and technologies for generating quality data for data mining applications. We should carry out, such as discrepancy detection, ontology based data integration, and logics for resolving data conflicts. This paper suggests to focus on model logics for removing false knowledge and dispelling inconsistencies that follow epistemic properties: veridicality, introspection and consistency; and majority (or consensus, or arbitrary) based logics for resolving conflict that have the property of obeying the weighted majority principle in case of conflicts. Therefore, one can build provision for recovering incomplete data, purifying data, and resolving data conflicts; machinery for using domain ontology to correct inaccuracies, remove anomalies, eliminate duplicate records, fill holes in the data and check entries for consistency; and logics for resolving conflicts and removing non-veridicality data. This paper suggests to utilize external data collected from the Web or other media.

On the other hand, real datasets are simply huge. However, typical KDD problems only a small portion of the available features are relevant for prediction. In addition, many data mining applications require identification of a relatively small number of interpretable features. Post-1990 Statistics literature has seen enormous strides made in automatic feature selection and interpretable models with good predictive power; with tools such as Markov Chain Monte Carlo and additive models. However this work is, in general, not geared towards large mining applications. This paper suggests to selectively combine methods from both sets of literatures and build upon them with KDD problems in mind. One can build provision for sampling, partitioning, feature/attribute selection, instance selection, dimensionality reduction, information filtering, ensemble and subspace methods and techniques.

B. Mining Mixed Data

This subsection discusses issues in mining mixed data sets. Such data sets are common in the area of multimedia, web and text mining. Key strategies include analysis of large sets of complex unstructured or semi-structured data; discovering relationships between data items or segments within video sequences, based on their content; extracting patterns from sounds and relating them to patterns in image (video frame) sequences; categorizing speech and music; recognizing and tracking objects in large amount of video stream data; and discovering and interpreting relations between different multimedia components (for example, network of texts and images). Outcomes include Smart Financial Adviser system; Clinical Knowledge Discoverer system; Multimedia Stego Discoverer; and Clinical Knowledge Discoverer system.

For multiple data source mining, one can design a group of pattern discovery systems to deal with some issues in, mainly including logics for enhancing external data, logics for solving conflicts within external and internal data or patterns, a pattern discovery system and a post-analysis system. This paper suggests to mine local patterns at different data sources and forward the local patterns (rather than the original raw data) to a centralized place for global pattern analysis.

Also, this paper suggests to formulate an XML-enabled pattern discovery framework for tackling unstructured or semi-structured data. Unlike traditional well-structured data whose schema is known in advance, XML data (unstructured or semi-structured data) does not have a fixed schema, elements in XML data have contextual positions which carry the order notion, and the structure of data can be incomplete or irregular. One can propose formalism for describing the semantics of XML data; transformation of this formalism into XML schema for document design; pattern representation for XML data; and template model for guiding pattern discovery. And then, one can build provisions for extracting XML data structures, XML data modeling, finding embedded subtrees, and ontology mapping; and a template-guided declarative language for mining XML-enabled patterns by extension of an XML query language XQuery. The ultimate objective is to facilitate the utilization of XML data broadly existing in business data repositories.

C. *Data Mining for Complex Modeling*

This subsection explores mining techniques and methods for complex modelling. Core theoretical issues include: interest based pattern discovery; powerful pruning; high dimensional data classification and clustering; and data security-enhancing. Each project is also designed to support market data analysis applications.

Clustering High Dimensional Data: This paper suggests to establish strategies for dimensionality reduction; subspace clustering of high dimensional data; and subspace ranking. One can build provision for partitioning approaches for keeping the number of cells from exponentially increasing; neighbor formation for avoiding considering the exponential number of neighboring cells; density compensations for improving the clustering accuracy and quality; and methods for discovering clusters in subspaces effectively.

Strategies for Identifying Interesting Patterns: This paper suggests to investigate efficient and effective methodologies and techniques for identifying useful interactions within data, mainly including the measurement of the interestingness or potential value of a pattern; efficiently exploring the exponential space of potential patterns to identify interesting and valuable patterns; and communicating patterns and their value to users. One can build provisions for identifying unexpected patterns, exceptional patterns, high-value patterns, contrast sets, rare but significant events, and outliers.

V. CONCLUSION

Social science researchers have long been interested in analyzing a whole host of phenomena that arise in decentralized market economies. This paper has chosen Economic Sciences as the data users because of the richness of the data and continuing rapid evolution of the theories and models that are necessary to underpin the domain structure and behavior. These are hard problems that take both the economic sciences and KDD beyond what is currently being looked internationally. This paper has initiated research into issues through some strategies/suggestions, where some are focused on economic science and business issues and others on information technology issues of KDD. In particular, some strategies have been proposed to address these business data mining.

VI. ACKNOWLEDGEMENT

This work was supported in part by China NSF (61170131) and Guangxi NSF (2012GXNSFGA060004).

REFERENCES

[1] Bay, S. D. and Pazzani, M. J. Detecting Change in Categorical Data: Mining Contrast Sets. In: KDD 1999, pp.302-306.

[2] YS Qin, SC Zhang, et al. Semi-parametric Optimization for Missing Data Imputation. Applied Intelligence, 27(1): 79-88 (2007).

[3] Shichao Zhang, et al. CIBuilder: A System Prototype for Building Confidence Intervals for Different Groups. KDD 2006 Demo.

[4] Shichao Zhang. Detecting Differences between Contrast Groups. IEEE Transactions on Information Technology in Biomedicine, 12(6): 739-745 (2008).

[5] Shichao Zhang: Nearest neighbor selection for iteratively kNN imputation. Journal of Systems and Software, 85(11): 2541-2552 (2012)