Data Mining for Business Applications

KDD-2006 Workshop

Rayid Ghani
Accenture Technology Labs
161 N. Clark St
Chicago, IL 60601
rayid.ghani@accenture.com

Carlos Soares
LIACC/Fac. of Economics, University of Porto
Rua de Ceuta 118, 6 andar
4050-180 Porto, Portugal
csoares@liacc.up.pt

ABSTRACT

Even though data mining has been successful in becoming a major component of various business processes as well as in transferring innovations from academic research into the business world, the gap between the problems that the research community works on and real-world ones is still significant. We believe that it is essential for the business and the academic research communities to interact frequently. The goal of the KDD-2006 Workshop on Data Mining for Business Applications was to gather both researchers and business practitioners and talk about their different perspectives and to share their latest problems and idea. We wanted to not only bring them together at KDD but also to create relationships that would continue and grow after the event as well.

1. INTRODUCTION

Data Mining in various forms is becoming a major component of business operations. Almost every business process today involves some form of data mining. Customer Relationship Management, Supply Chain Optimization, Demand Forecasting, Assortment Optimization, Business Intelligence, and Knowledge Management are just some examples of business functions that haven been impacted by data mining techniques.

Even though data mining has become critical to businesses, most of the academic research in data mining is conducted on mostly publicly available data sources. This is mainly due to two reasons: 1) the difficulty academic researchers face in getting access to large, new, and interesting sources of data 2) limited access to domain experts who can provide a practical perspective on existing problems and provide a new set of research problems. Corporations are typically wary of releasing their internal data to academics and in most cases, there is limited interaction between industry practitioners and academic researchers working on related problems in similar domains.

The goals we defined for the Workshop on Data Mining for Business Applications were:

1. Bring together researchers (from both academia and industry) as well as practitioners from different fields to talk about their different perspectives and to share their latest problems and ideas.
2. Attract business professionals who have access to interesting sources of data and business problems but not the expertise in data mining to solve them effectively.

KDD is a unique venue for this purpose as it gathers researchers from both academia and industry who are involved in data mining. Therefore, we organized a workshop where business professionals could present and discuss their problems, views and ideas on the field, as well as pose research challenges, thus making it easier to attract this audience to participate in the conference as well as to interact with the research community.

In general our goals were achieved effectively. The audience included around 80 people, a broad and varied set of participants, with a lot of interaction not only during the workshop but also after its conclusion. Therefore, we believe that this workshop served as a bridge between the traditional KDD community and business professionals — two groups of participants that have a lot to learn from each other.

Some of the topics addressed in the workshop were:

- Novel business applications of data mining
- New classes of research problems motivated by real-world business problems.
- Data mining as a component of existing business processes
- Integration of data mining technologies with other kind of technologies that already exist inside corporations
- Lessons learned from practical experiences with applying data mining to business applications

The program included 15 presentations and two panel discussions. The theme of the first panel was "Bridging the Gap between Data Mining Research and Practical Business Applications" and the participants were Ronny Kohavi (Microsoft), Karl Rexer (Rexer Analytics) and Galit Shmueli (University of Maryland). The second panel, entitled "Deploying Data Mining Solutions: Stories, Challenges and Open Issues" was organized around four the presentations, namely the ones by Tyler Kohn (FortisForge), Ramin Mikaili (Accenture), Richard Boire (Boire Filler Group) and Françoise Soulié Fogelman (KXEN). The proceedings and presentations are available on the web at
http://labs.accenture.com/kdd2006_workshop/

The following sections report on the panels while Section 4 summarizes the content of the presentations. The last section presents some conclusions.

2. PANEL: BRIDGING THE GAP BETWEEN DATA MINING RESEARCH AND PRACTICAL BUSINESS APPLICATIONS

Both of the panel discussions were very lively and generated a lot of comments and questions from the audience. The first panel addressed Bridging the Gap Between Data Mining Research and Practical Business Applications (Ronny Kohavi, Karl Rexer, and Galit Shmueli). We asked the panelists to talk about 1) what they think of the current state of the gap between data mining research (in academia as well as industrial labs) and practical applications in the business world and 2) what steps can the two groups (researchers and business practitioners) take to bridge this gap.

Kohavi contrasted companies that have vastly different ways of doing research and deploying/transferring the results of the research. The extremes he mentioned ranged from having a traditional industrial research lab where researchers are hired to do just research, publish openly, and then work with product groups for technology transfer. The other extreme is a model where everyone performs software development and PhDs are hired as software developers. In the latter case, research is not a separate function but embedded in software development/engineering. This process makes technology transfer implicit but the research outcomes do not get distributed outside the company and publishing is typically not encouraged. This often makes a difference in what kind of talent can be attracted and retained.

Galit Shmueli proposed that in addition to looking at Statistics and Computer Science graduates, MBA students should be viewed as candidates for data mining jobs since they are increasingly taking more courses in analytics and are well-prepared for this role. She also discussed her experiences with academia and industry working together on funding proposals and consulting projects and pointed out the key to effective interaction being the alignment of incentives for both sides.

Karl Rexer pointed out many large companies do not have a dedicated analytics group and need external companies to fulfil this role for them. He emphasized the need to focus training on how to use data mining in specific business situations instead of teaching data mining as a technical discipline without regard to specific business problems.

3. PANEL: DEPLOYING DATA MINING SOLUTIONS: STORIES, CHALLENGES AND OPEN ISSUES

The second panel discussion, Deploying Data Mining Solutions: Stories, Challenges, and Open Issues (Tyler Kohn, Ramin Mikaili, Richard Boire, and Franoise Fogelman) gave the audience different perspectives of deploying data mining solutions. Tyler Kohn discussed modifications to the CRISP-DM model that facilitate better collaborations between academics and businesses. He pointed out several problems with the collaboration process today including misalignment of goals, power dynamics, intellectual property issues,and resource allocation. Ramin Mikaili presented a business-driven analytics framework developed by Accenture for telecommunications clients. He described several projects where this framework was used with extremely positive results. Richard Boire discussed a few situations where applying data mining techniques without the business context and knowledge resulted in incorrect results. Fogelman argued for the need to automate data mining by creating a large number of data mining models without much human intervention in order to make the technology accessible to the business masses. She pointed out that most companies today rely on an analytics group to build models for every department. If data mining is to become integral to every business, data mining tools would have to automate the process of data handling and model construction to reduce lag time as well as allow business users to focus on the business implications instead of worrying about understanding the statistical analysis.

4. PRESENTATIONS

The first presentation was by Galit Shmueli on *Forecasting Online Auctions using Dynamic Models*, a joint work with Wolfgang Jank and Shanshan Wang (University of Maryland). The goal is to forecast the final price of an item during the auction. The solution proposed consists of a dynamic forecasting model based on functional data analysis that takes price dynamics into account. The applications of such a system include enabling the user to choose the auction which is expected to achieve the lowest price or the sellers to use an insure-it-now option.

The second paper was presented by Germán Creamer (Columbia University), describing joint work with Yoav Freund (University of California, San Diego) on *A boosting approach for automated trading*. In electronic markets, the order book, which provides a very detailed view of the state of the market, is made available to all traders. The authors propose a boosting approach that makes use of this more detailed view to address the problem of short-term trading. The method was applied to the Penn-Lehman Automated Trading (PLAT) competition, obtaining the second best results in its group.

Next, Peter van der Putten discussed *A Decision Management Approach to Basel II Compliant Credit Risk Management*, a joint work with Arnold Koudijs and Rob Walker (Chordiant Software). They discuss the requirements defined by the Basel II Accord for credit risk management. Based on these new requirements, a number of opportunities for data mining are identified. However, the authors argue that data mining should be integrated with a knowledge-based approach for successful results to be achieved and give an overview of such a system.

The last paper of the first session was presented by Ronnie Alves, describing joint work with Pedro Ferreira, Orlando Belo, João Lopes, Joel Ribeiro (University of Minho), Luís Cortesão (Portugal Telecom Inovação) and Filipe Martins (Telbit) on *Discovering Telecom Fraud Situations through Mining Anomalous Behavior Patterns*. This work addressed the problem of superimposed fraud, consisting of illegitimate use of an account. Their approach is based on the notion of *signatures*, which consist of a set of features characterizing the normal behavior associated with an account. Two dynamic methods are proposed, one based on deviation detection and the second on cluster analysis. The methods are tested on data from Portugal Telecom and several anomalous situations were detected that were regarded as valuable by the domain experts.

The second session started with a presentation entitled *Interactivity Closes the Gap: Lessons Learned in an Automo-*

tive Industry Applications by Axel Blumenstock (University of Ulm), describing work carried out with Jochen Hipp, Steffen Kempe, Carsten Lanquillon and Rüdiger Wirth (DaimlerChrysler Group Research). In this paper, lessons learned from an application involving the early detection and explanation of problems with vehicles are discussed. The business experts in this project wanted to be actively involved in the data mining process. Given this need, it was hard to apply off-the-shelf techniques and the authors ended up developing new methods that emphasized simplicity and interactivity. The latter is particularly important as it enables the user to explore the data as well as to manipulate the model.

The following paper was presented by Tilmann Bruckhaus on *Customer Validation of Commercial Predictive Models*, representing joint work with William Guthrie (Numetrics). This talk addressed the well-known but largely unsolved problem of the gap between the metrics used for evaluation of data mining results in the research community the the metrics that are typically required in business settings. Based on their experience in the semiconductor industry, the authors present a mapping between typical questions that are raised by business users and research metrics typically used in the academic community.

The afternoon started with a talk by Claudia Perlich on *Quantile Trees for Marketing*, work carried out together with Saharon Rosset (IBM Research). The goal is to predict the IT *wallet* of IBM customers, i.e. the amount that they spend on IT. Accurate predictions are useful for the marketing department to estimate the potential of growth for those customers and to focus future marketing resources. The approach proposed has the advantage that it can be implemented as a wrapper around any regression trees algorithm.

The next talk by Akhil Kumar (Penn State University) was on *Mining and Querying Business Process Logs*. The author proposes a new distance-based algorithm for mining logs of processes. The underlying distance structure is also used to support the execution of queries concerning the corresponding process model.

The paper on *Using Data Mining in Procurement Business Transformation Outsourcing* by Moninder Singh and Jayant Kalagnanam (IBM Research) was presented by the first author. The problem addressed is the aggregation of different databases containing *spend information*, i.e. information about products and services acquired by organizations. This problem is relevant for Business Transformation Outsourcing service providers, among others. By aggregating the databases of their clients, they are able to make larger orders to their suppliers and, thus, negotiate better prices. The authors identify many issues and problems that arise in this task and the solution they have developed.

The following presentation was by Alex Kass, on the *Business Event Advisor: Mining the Net for Business Insight with Semantic Models, Lightweight NLP and Conceptual Inference*, which is joint work with Christopher Cowell-Shah (Accenture Technology Labs). The motivation for this work is illustrated by an episode describing how Bill Gates realized that Google could be serious competition for Microsoft while browsing the web. The authors propose a *corporate radar kit* that monitors the web for external events that are relevant for a given company, extracts events, builds short descriptions of those events and ranks them according to their expected impact using a domain model for that industry. The challenges identified in the process of developing a solution are discussed and a prototype in the automotive industry was also shown.

The last paper before the afternoon break, with the curious title of *Zen and the Art of Data Mining*, was by T. Dasu, E. Koutsofios and J. Wright (AT&T Labs Research). They present their views on the requirements for a successful data mining project, in the context of two monitoring applications. The first one is concerned with the analysis of the process of feeding billing data into the computers that process it. In the second application, the activity of servers and routers that support e-commerce sites is monitored. Besides identifying essential ingredients for success, the authors discuss some of the related open issues. In particular, they emphasize the difficulty in integrating data mining in the business processes, which are highly volatile.

5. CONCLUSIONS

We were able to solicit an interesting set of papers and speakers and are extremely thankful to all the authors, presenters, panelists and the program committee for their efforts in making this a successful workshop. We were successful at bringing together people from a variety of backgrounds and facilitating discussion among them. Researchers from both academia and industry, data mining consultants, statisticians, computer scientists, and business practitioners were all represented at the workshop. In order for data mining as a field to continue being successful, the research community needs to be in continuous dialog with the business practitioners and use these discussions to motivate new research that will be relevant for businesses in the future. We believe that this workshop will help us move towards this goal and improve the future of data mining.

6. WEBSITE

The papers and presentations from the workshop are on the Web at http://labs.accenture.com/kdd2006_workshop/