UBDM 2006: Utility-Based Data Mining 2006 Workshop Report

Bianca Zadrozny
Federal Fluminense University
Rua Passo da Pátria, 156 – Bloco E
Niterói, RJ, CEP 24210-240, Brazil
bianca@ic.uff.br

Gary Weiss
Computer and Information Science Dept.
Fordham University
Bronx, NY 10458, USA
gweiss@cis.fordham.edu

Maytal Saar-Tsechansky
Red McCombs School of Business
University of Texas at Austin
Austin, TX 78712, USA
maytal@mail.utexas.edu

ABSTRACT

In this report we provide a summary of the Second International Workshop on Utility-Based Data Mining (UBDM-06) held in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The workshop was held on August 20, 2006 in Philadelphia, PA, USA. As was the case for the first UBDM workshop, this workshop brought together researchers who currently investigate how to incorporate economic utility aspects into the data mining process and practitioners who have real-world experience with how these factors influence data mining applications.

Keywords

Cost-sensitive learning, active learning, active information acquisition, utility-based data mining

1. INTRODUCTION

Early work in data mining rarely addressed the complex circumstances in which knowledge is extracted and applied. It was assumed that a fixed amount of training data was available and only simple objectives, such as predictive accuracy in the case of predictive models and support/confidence in the case of association rules, were considered. Over time, it became clear that these assumptions were unrealistic and that *economic utility* had to be considered and incorporated into the data mining process. This realization has led to research on *active information acquisition*, which focuses on methods for cost-effective acquisition of information for the training and test data, and research on *cost-sensitive learning*, which considers the costs and benefits associated with using the learned knowledge and how these costs and benefits should be factored into the data mining process. Other utility considerations that are relevant in the data mining context include the running time of the data mining algorithm as well as the costs and benefits associated with cleaning the data, transforming the data and constructing new features.

In most previous work, these different types of utility considerations have been studied in isolation, without much attention to how they interact. As had been the case for the UBDM-05 workshop, one goal of this workshop was to bring together researchers who currently consider different economic utility aspects in data mining to continue to promote an examination of the impact of economic utility throughout the entire data mining process. This workshop again encouraged the field to go beyond what has been accomplished individually in the areas of active information acquisition and cost-sensitive learning. The workshop organizers further suggested that all of utility-based data mining could be viewed using a common framework, with a key question being whether such a framework would be beneficial to real-world applications.

Past research which has addressed the role of economic utility in data mining has mostly focused on predictive classification tasks. An additional goal of this workshop was to explore methods for incorporating economic utility considerations into descriptive data mining tasks such as association rules mining.

2. WORKSHOP SUMMARY

The workshop received a strong response from the data mining community which was reflected in the quality and breadth of the accepted papers. Each submitted paper was reviewed by two or three members of the program committee. In total, nine regular papers were accepted for inclusion in the workshop. In addition, the workshop featured two invited talks and one panel discussion. All of the papers are available from the workshop web page at http://www.ic.uff.br/~bianca/ubdm-kdd06.html.

The nine accepted papers can be broadly categorized into three distinct topics: information acquisition, utility aspects of descriptive data mining, and cost-sensitive learning and its applications.

2.1 INVITED TALKS

The workshop featured two invited talks. The first invited talk, which opened the workshop, was *Budgeted Learning of Probabilistic Classifiers* by Russell Greiner from the University of Alberta. The talk introduced and discussed challenges and several approaches for solving the budgeted learning problem. Budgeted learning involves induction where unknown feature values can be acquired individually for each training example at a cost, and where the total cost must not exceed a predefined budget. The goal is to acquire the feature values that maximize classifier performance given the budget. The speaker contrasted this problem with other related fields such as active learning, optimum experimental design and online learning. The speaker presented a Bayesian framework for this active model selection, analyzed the hardness of the problem, presented several heuristics strategies and evaluated their performance for a variety of setting. Empirical evaluations showed that while different policies perform better for different scenarios, one policy, Biased Robin, performs consistently well even though it does not employ information about the budget. The Biased Robin (BR) policy suggests to “play the winner”, i.e., to continue selecting feature values of a given attribute as long as it continues to improve the model’s performance. The speaker also discussed the budgeted learning problem for learning a Naïve Bayes classifier where it is necessary to determine not only the feature but also the class value for which a feature value is acquired. Empirical results on synthetic and UCI data sets show that SFL and BR successfully reduce the cost of obtaining a Naïve Bayes classifier with a given generalization performance, because they effectively identify discriminate features. SFL acquires values for the feature for which the expected regret is lowest when the entire budget is spent to acquire values of this features. Empirical results were also presented for learning a bounded active classifier in which the induction algorithm considers a classification budget for acquiring feature values at inference time. The procedure that performed significantly better was a Randomized SFL (RSFL). Rather than using the SFL scores of the expected loss from an acquisition deterministically, RSFL acquires features of a given class from a distribution, where the weight assigned to each acquisition is inversely proportional to its loss score. The speaker discussed various directions for future work. These included more general problems, where class labels as well as feature values are missing and more complex costs models, such as non-uniform costs, or the case where some feature values must be purchased in bundles. Algorithmic challenges, such as developing policies with guarantees on performance were also discussed.

The second invited talk, which opened the afternoon session, was *Reinforcement Learning and Utility-Based Decisions* by Michael Littman from Rutgers University. In this talk, the speaker drew a parallel between Reinforcement Learning (RL) and Utility-Based Data Mining. In RL the goal is to act so as to maximize the utility of behavior, while minimizing experience and computational costs. In UBDM, the goal is to act so as to maximize the utility of using the mined knowledge while minimizing the costs of acquiring and mining the data. Therefore, the two frameworks have a similar structure and, in particular, require a joint optimization which is computationally intractable to perform exactly. The speaker pointed out that, because of this intractability, recent trends in RL literature suggest the appropriateness of a PAC ("*probably approximately correct*") style of analysis for RL. In the case of RL, this can lead to algorithms able to obtain near-optimal utility with polynomial-bounded amounts of experience and computation. The speaker gave examples of PAC algorithms for RL beginning with algorithms for the simplest class of RL problems, the k-Armed Bandits problem. Then, he moved on to PAC algorithms for model-based RL, where the learner has access to a model of the environment. He presented the Model Based Interval Estimation (MBIE) algorithm and compared it with two previous approaches: E^3 and R_{MAX}. Finally, he discussed recent work that could potentially lead to PAC-based model-free RL algorithms. The speaker also showed two practical examples of the application of RL algorithms: a robotic example, where a car had to learn to maintain constant speed while going up a ramp, and a network repair example, where the goal was to recover from a corrupted network interface configuration. The speaker concluded the talk by saying that the PAC idea of keeping costs within bounds instead of performing a joint minimization could also lead to practical algorithms for UBDM.

2.2 CONTRIBUTED PAPERS

Nine contributed papers were presented at the workshop. These papers are briefly described in this section. The papers that address similar topics were presented contiguously at the workshop and are discussed together here.

2.2.1 Information Acquisition

Much of the work on data mining ignores information acquisition costs. In *Maximizing Classifier Utility when Training Data is Costly*, Gary Weiss and Ye Tian investigate the impact on the data mining process when training examples are not free, but rather have a fixed cost. They introduce a utility measure that factors in the cost of the training examples and the cost of misclassification errors and then use this measure to analyze the performance of ten data sets as the training set size is varied. This analysis is then used to identify the optimal training set size for each data set. A progressive sampling strategy is then evaluated and empirically shown to identify a near-optimal training set size—with near-optimal classifier utility. This is the first work to extend research on progressive sampling to take the cost of training data into account.

The other paper related to information acquisition was a position paper by Omid Madani entitled *Prediction Games in Infinitely Rich Worlds*. The paper introduces the challenges and advantages of learning in infinitely information-rich worlds such as the World Wide Web, the visual/physical world, and people's actions. Madani proposes the learning process of playing prediction games toward making powerful massive unsupervised learning possible. The games are played by a system that takes its sequence of inputs from the world, or actively searches and acquires these experiences to generate learning episodes. The utility of such systems can be defined in terms of the operationality of the prediction system: coverage, depth, accuracy, and speed. Madani views the prediction system as consisting of a number of interacting components driven by multiple algorithms. The system learns, adapts, self-organizes and grows in its functionality over time. He outlines the desiderata for such a system which include (a) use of online algorithm that are time and memory efficient to enable prediction games with unbounded data that are processed and discarded online, (b) handling large number of features, (c) handling large numbers of categories, and (d) robustness to imperfections, uncertainty, and variety. The paper discusses several directions for future research including to research to efficiently learn and recognize myriad categories and the need to better understand natural systems that perform similar tasks such as the animal brain which uses experiences for massive learning that is not explicitly supervised.

2.2.2 Utility Aspects of Descriptive Data Mining

Four of the nine papers included in the workshop considered utility in the context of descriptive data mining tasks. The first two papers are concerned with finding high utility itemsets. This work is an extension of the work on finding frequent itemsets, which is the critical first step in association rule mining. High utility itemsets differ from frequent itemsets in that the utility of the items in each itemset are taken into account (e.g., the profit associated with an item may be considered).

The first paper related to descriptive data mining, *Efficient Mining of Temporal High Utility Itemsets from Data Streams* by Vincent Tseng, Chun-Jung Chu and Tyne Liang, focused on the problems associated with finding temporal high-utility itemsets in data streams. This is the first work to tackle this specific problem. The algorithm developed by the authors is notable because it generates relatively few temporal high-utility 2-itemsets and thus limits the memory space and CPU I/O time that is needed, meeting the critical time and space efficiency requirements for mining data streams.

The second paper, A *Unified Framework for Utility Based Measures for Mining Itemsets*, by Hong Yao, Howard Hamilton and Liqiang Geng, focuses on the measures used for utility-based itemset mining. It begins be reviewing the various utility-based measures available for itemset mining and describes ten such measures. These measures must factor in the interestingness and frequency of the itemset. The authors formalize the semantic significance of utility measures and classify existing measures into one of three categories: item level, transaction level and cell level. A unified framework is then proposed for incorporating utility-based measures into the data mining process via a unified utility function. Three mathematical properties (anti-monotone, convertible and upper-bound) of utility-based measures, which impact the time and space costs of itemset mining were then identified and each of the ten utility measures were then analyzed with respect to these properties. This work provides an excellent introduction to utility-based measures for itemset mining and a coherent way of organizing and comparing these measures.

The topic of utility measures, which was addressed in the previous paper, is also addressed in the third paper on descriptive data mining, *Assessing the Interestingness of Discovered Knowledge Using a Principled Objective Approach* by Robert Hilderman. In this case, however, the primary focus is on interestingness (which can be viewed as a type of utility) and the goal is to rank the patterns, or summaries, generated by data mining in order to aid the user by highlighting the most interesting patterns. This work theoretically and empirically evaluates twelve diversity measures as heuristic measures of interestingness. Five principles are identified that a measure of interestingness must satisfy to be useful for ranking summaries, and seven of the twelve interestingness measures are shown to satisfy all of these properties. This work thus addresses the important task of evaluating utility measures for the purpose of ranking the results from descriptive data mining.

Privacy is a very serious concern associated with data mining, especially when microdata—raw data that has not been summarized—is involved. The fourth paper, *Utility-Based Anonymization for Privacy Preservation with Less Information Loss* by Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi and Ada Wai-Chee Fu addresses this issue. A common approach for ensuring privacy involves generalizing or suppressing attributes that can be used to identify specific individuals. This work extends recent research by considering the utility of the attributes during the anonymization process, such that the amount of useful information preserved in the data is maximized. Two simple but efficient heuristic local recoding methods for utility-based anonymization are described and evaluated on real and synthetic data sets and shown to boost the quality of analysis using anonymized data, when compared to existing methods. Note that because this preprocessing step is concerned with how best to modify the descriptions of the data, we associate this work with descriptive data mining.

2.2.3 Cost-Sensitive Learning and its Applications

Three of the nine papers presented at the workshop considered the costs and benefits associated with using a predictive model and how these costs and benefits should be factored into the data mining process of learning the model. The first paper, entitled *Beyond Classification and Ranking: Constrained Optimization of the ROI* by Lian Yan and Patrick Baldasare, proposes an algorithm for learning a classifier that directly optimizes the return on investment (ROI), a common measure used in financial services applications such as predicting collectability of accounts receivable and predicting defection of mutual funds accounts. This problem is different from standard cost-sensitive learning or standard ranking problems because there is a budget constraint as well as example-dependent misclassification costs. The paper formulates this problem as a constrained optimization problem, converts into an unconstrained optimization problem and solves it using a gradient descent method. Experiments on two financial datasets comparing the proposed method to standard classification, weighted classification, ranking and regression show that the method leads to a substantial improvement of financial impact.

The second paper on cost-sensitive learning, entitled *Pricing Based Framework for Benefit Scoring* by Nitesh Chawla and Xiangning Li, also addresses a financial application of predictive data mining. The paper presents a method for deriving a pricing scheme that maximizes the total profit that can be obtained from using a probabilistic classifier. They illustrate the use of the method for a loan practice based on credit scoring models and a benefit matrix. The framework enables pricing the loan for each individual customer based on the loan amount and the corresponding interest rate, as well as the probability of defection predicted by the credit scoring model. Preliminary experiments on UCI Machine Learning repository datasets show that there is a strong relationship between the quality of the probability estimates and the resulting profits from the model.

The third and last paper on cost-sensitive learning, entitled *Maximum Profit Mining and Its Application in Software Development* by Charles Ling, Victor Sheng, Tilmann Bruckhaus and Nazim Madhavji, proposes a data mining solution to the problem of predicting escalation risks of software defects to assist human experts in the review process of software. Like in many business applications, the ultimate goal of software defect escalation prediction is to maximize the net profit, i.e., the difference in the profit before and after introducing the data mining solution. Note that this in general depends on the current policy for choosing which software defects to correct. The paper presents a novel and simple method for converting the maximum net profit problem into a cost-sensitive classification problem. This is done by expressing the net profit as a linear formula of false positives, false negatives, true positives and true negatives. After this transformation, it is possible to apply any existing cost-sensitive learning method to this problem. The paper compares a number of different cost-sensitive learning methods (Undersampling, Costing, MetaCost, Weighting and CSTree) on a real dataset from the software defect escalation domain and show that many of the methods can achieve large positive net profits. In particular, the results show that CSTree can produce a large positive net profit while providing comprehensible results, which is important for deploying data mining solutions in industry.

2.3 PANEL DISCUSSION

The workshop was concluded with a panel discussion. The participants were Russell Greiner of the University of Alberta, Nitesh Chawla of the University of Notre Dame, Gerald Fahner, Analytic Science Director at Fair Isaac Inc., and Dragos D. Margineantu of The Boeing Company. There was wide agreement among panel members and workshop participants that UBDM work as reflected in the research contributions in the field includes a set of challenging theoretical problems of significant practical implications. Participants and panelists noted that it is useful to pursue future UBDM meetings. It was noted that UBDM is related to some work in Experimental Design and Game Theory. Thus researchers were encouraged to examine work in these fields for relevant references and for inspiration.

In light of the potential practical implications of Utility-Based Data Mining research, panelists noted the striking lack of real-world, publicly available data which include real costs and utility information. In order to fill in the gap, researchers often generate random costs and benefits for publicly available data sets. This is not an ideal approach because it makes it difficult to compare methods and because it is never a good idea for a researcher to generate the evaluation data for his/her own work. To promote and further the impact of UBDM research, the panelists agreed that it would be beneficial to devise a set of well defined tasks and corresponding data sets (with cost/benefit information). The panelists discussed potential venues for obtaining relevant data as well as the creation of a simulated environment in which different tasks can be defined and from which data can be generated. It was noted that the flexibility of such a platform would support coordinated efforts to advance research progress in a variety of emerging practical tasks. It was noted that the platform could also help devise competitions similar to the Trading Agent Competition to help advance work in the field. This discussion concerning the need for cost and benefit information, and the recognition that this information will not always be available, led to a discussion about the need to develop robust methods that exhibit consistent performance for different utilities and data sets.

3. CONCLUSION

We feel that the workshop was a success and achieved our goals. The majority of attendees stayed for the full-day workshop and participated very actively in the discussions, demonstrating a growing interest among data mining researchers and practitioners in Utility-Based Data Mining. The discussions were lively and suggested interesting areas for future research as well as ideas for how to better coordinate and promote research efforts in the field.

One of our goals for this workshop was to broaden the scope of utility-based data mining to include descriptive data mining tasks. We believe this goal was successfully reached, since four of the nine papers addressed utility considerations in descriptive data mining tasks. This is notable because most of the previous work in utility-based data mining has focused on predictive tasks.

We have also seen the community move away from addressing pure cost-sensitive learning or information retrieval tasks to address basic issues that can help establish the foundations of UBDM. For example, two of the papers, *A Unified Framework for Utility Based Measures for Mining Itemsets* and *Assessing the Interestingness of Discovered Knowledge Using a Principled Objective Approach* are concerned with providing a general framework for describing and analyzing utility measures, and each of the papers reviews a large number of existing utility measures in this context. The work described in these papers can clearly aid researchers who introduce new utility measures.

We aimed for the workshop to facilitate interaction among researchers and practitioners and to promote the transfer of ideas among problems previously addressed in isolation and in this regard we think we were successful. Part of this success is due to the fact that we allocated significantly more time than last year to discussion. We thank the authors, guest speakers, program committee members and attendees for their contributions towards this end. We are indebted to Sunita Sarawagi, the KDD-06 Workshop Chair, and to the SIGKDD for organizational and funding assistance. We also thank our anonymous workshop proposal reviewers for their insightful suggestions and encouragement.

About the Authors:

Bianca Zadrozny is an Assistant Professor in the Computer Science Department of Federal Fluminense University in Brazil. Her research interests are in the areas of applied machine learning and data mining. She received her B.Sc. in Computer Engineering from the Pontifical Catholic University in Rio de Janeiro, Brazil, and her M.Sc. and Ph.D. in Computer Science from the University of California at San Diego. She has also worked as a research staff member in the data analytics research group at IBM T.J. Watson Research Center. (http://www.ic.uff.br/~bianca)

Gary Weiss is an Assistant Professor in the Computer and Information Science Department at Fordham University. His research interests include machine learning and data mining and the fundamental issues that arise when tackling complex, real-world problems. Specific topics he has worked on include utility-based data mining, learning from rare classes and cases, event prediction and data mining applications in the telecommunications industry. He received his B.S. from Cornell University, his M.S. from Stanford University and his Ph.D. in Computer Science from Rutgers University. He has worked at Bell Labs and AT&T Labs, including five years in a marketing analysis group where he applied data mining methods to complex business problems.
(http://storm.cis.fordham.edu/~gweiss)

Maytal Saar-Tsechansky is an Assistant Professor in the McCombs School of Business, The University of Texas at Austin. She received her Ph.D. from New York University and obtained her B.S and M.S from Ben Gurion University, Israel. Her research focuses on economic machine learning and data mining methods for data-driven business intelligence.
(www.mccombs.utexas.edu/faculty/Maytal.Saar-Tsechansky)