Report on UBDM-05: Workshop on Utility-Based Data Mining

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

Bianca Zadrozny
IBM T.J. Watson Research Center
1101 Kitchawan Road, Route 134
Yorktown Heights, NY 10598, USA
zadrozny@us.ibm.com

ABSTRACT

In this report we provide a summary of the First International Workshop on Utility-Based Data Mining (UBDM-05) held in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The workshop was held on August 21, 2005 in Chicago, IL, USA. This workshop was geared toward researchers with an interest in how economic utility factors affect data mining (e.g., researchers in cost-sensitive learning and active learning) and practitioners who have real-world experience with how these factors influence data mining applications.

Keywords

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

1. INTRODUCTION

Early work in predictive data mining did not address the complex circumstances in which models are built and applied. It was assumed that a fixed amount of training data were available and only simple objectives, namely predictive accuracy, were considered. Over time, it became clear that these assumptions were unrealistic and that *economic utility* had to be considered during the three main phases of data mining: 1) acquiring training data, 2) building a model and 3) applying the model in realistic environments. The machine learning and data mining communities responded with research on *active information acquisition*, which focused on methods for cost-effective acquisition of information for the training data (phase 1) and research on *cost-sensitive learning*, which considered the costs and benefits associated with using the learned knowledge and how these costs and benefits should be factored into the data mining process (phase 3). The utility considerations for phase 2 include the running time of the algorithm as well as the costs and benefits associated with cleaning the data, transforming the data and constructing new features.

Almost all work that considers the impact of economic utility on data mining focuses exclusively on one of the three stages in the data mining process. Thus, economic factors have been studied in isolation, without much attention to how they interact. One goal of this workshop was to begin to remedy this deficiency by bringing together researchers who currently consider different economic aspects in data mining, and by promoting an examination of the impact of economic utility throughout the entire data mining process. This workshop 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.

Past research which has addressed the role of economic utility in data mining has focused on predictive classification tasks. An additional goal of this workshop was to encourage researchers to explore methods for incorporating economic utility considerations into descriptive data mining tasks and other types of predictive data mining tasks.

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, 13 regular papers were accepted for inclusion in the workshop. In addition, the workshop featured 3 invited talks and one group/panel discussion. All of the papers are available from the workshop web page at http://storm.cis.fordham.edu/~gweiss/ubdm-kdd05.html.

The majority of the 13 accepted papers dealt with either active information acquisition or cost-sensitive learning and applications. The active information acquisition papers all focused on active feature-value acquisitions, while the cost-sensitive learning papers addressed a broad range of cost-sensitive learning issues.

2.1 INVITED TALKS

The workshop featured three invited talks, which were staggered throughout the day. The first invited talk, *Toward Economic Machine Learning and Utility-based Data Mining* was presented by Foster Provost from New York University's Stern School of Business. This talk introduced a general economic setting for utility-based data mining that includes as special cases cost-sensitive learning, traditional active learning, semi-supervised learning, active feature acquisition, progressive sampling and budgeted learning. Thus, this talk introduced a framework for utility-based data mining and demonstrated that a great deal of existing research can be viewed from a common perspective. It also clearly demonstrated that issues relating to economic utility arise throughout the entire data mining process and highlighted open questions that the community can address in future work.

Later in the morning Robert Holte from the University of Alberta presented the second invited talk, *Cost-Sensitive Classifier Evaluation*. This talk discussed how to evaluate classifier performance in a cost-sensitive setting when the misclassification costs and/or class distributions are unknown. It was argued that the classic technique for visualizing classifier performance in this situation—the ROC curve—does not allow several important experimental questions to be answered visually and that the cost curves, introduced by Drummond and Holte at KDD 2000, overcome these deficiencies. In particular, unlike ROC curves, cost curves address what is a given classifier's performance or expected cost given a set of misclassification costs and class distribution. The speaker also gave a live demonstration of a tool for drawing and analyzing cost curves given the output of different classifiers.

The third invited talk, *Machine Learning Paradigms for Utility-based Data Mining* was presented in the afternoon session by Naoki Abe from the IBM T.J. Watson Research Center. This talk described a number of existing machine learning paradigms that are relevant for utility-based data mining. Among these paradigms are active (or query) learning, active on-line learning, associative reinforcement learning and general reinforcement learning. The talk briefly reviewed learning theory techniques and results for each of these paradigms. Also, it showed some examples of real world applications for which some of these paradigms have proved to be satisfactory.

2.2 CONTRIBUTED PAPERS

Thirteen contributed papers were presented at the workshop. The talks are briefly described in this section. Those talks that address similar topics are discussed together (with few exceptions they were presented contiguously at the workshop).

2.2.1 Information Acquisition

Four of the talks presented at the workshop focused on the topic of active information acquisition and, in particular, focused on feature-value acquisition—where some feature values are unknown but can be acquired for a cost. The papers propose new policies for acquiring costly features so as to obtain the best model for a given cost.

The talk *Budgeted Learning of Bounded Active Classifiers* by Aloak Kapoor and Russell Greiner considers the "budgeted learning" problem in which a learner spends a fixed budget to acquire feature-values so as to produce the most accurate "active classifier". Given a classification model, an active classifier spends a fixed budget to acquire feature-values for each test instance. The paper presents a framework for the budgeted learning problem and proposes several feasible policies that a learner can employ. Empirical evaluation of the policies demonstrated that two of the policies are often superior to a benchmark policy in which values of every feature for each instance are acquired sequentially until the budget is exhausted. This talk was followed up by a second talk by the same authors, *Reinforcement Learning for Active Model Selection*, in which reinforcement learning techniques were used to learn an effective spending policy for acquiring feature values. The paper concludes that reinforcement learning is not likely to improve upon simpler policies.

The third talk on feature-value acquisition, *Economical Active Feature-value Acquisition through Expected Utility Estimation* by Prem Melville, Maytal Saar-Tsechansky, Foster Provost and Raymond Mooney examined two policies for acquiring feature values based on an estimation of the expected improvement in model accuracy per unit cost. The results showed that the author's Sampled Expected Utility policy is a promising strategy that reduces the cost of producing a model of a desired accuracy and also exhibits a consistent performance across domains.

The final paper on active feature-value acquisition, *Learning Policies for Sequential Time and Cost Sensitive Classification* by Andrew Arnt and Shlomo Zilberstein pointed out that the utility of acquiring a feature values depends not only on the dollar cost of measuring the feature, but on the time necessary to obtain the measurement. The problem is further complicated by the fact that the time spent measuring one attribute may impact the cost of future instances. The authors address the problem of how to select the feature values for measurement by using a decision theoretic approach, modeling the problem as a Markov Decision Process.

The paper *Noisy Information Value in Utility-based Decision Making* by Clayton Morrison and Paul Cohen addresses the question, "How much is information worth?" In the context of decisions, the value of information is generally the expected increase in utility of the decision as a result of having the information, but this becomes more complicated when the information cost is related to the quality of potentially noisy information. This talk presented an analytical decision model which incorporates the value of information when the information may be noisy.

2.2.2 Cost-Sensitive Learning

A number of the workshop talks related to cost-sensitive learning. The term "cost-sensitive learning" has traditionally been used in the predictive classification context where it generally refers to learning in domains where different classes have different misclassification costs.

The talk *One-Benefit Learning: Cost-Sensitive Learning with Restricted Cost Information* by Bianca Zadrozny presented a new formulation for cost-sensitive learning called One-Benefit learning. Instead of having the correct label for each training example, as in the standard classifier learning formulation, there is one possible label for each example (which may not be the correct label) and the cost or benefit associated with that label. The talk described how the One-Benefit learning problem could be reduced to the standard classifier learning problem so that any existing error-minimizing classifier learner could be used to maximize the expected benefit by correctly weighting the training examples.

Two of the talks on cost-sensitive learning concerned learning from data sets with unbalanced class distributions. This topic falls under cost-sensitive learning because in this situation the cost of misclassifying the rare class is typically much greater than that of misclassifying the common class (if this were not true then one would almost never predict the rare class). The talk *Does Cost Sensitive Learning Beat Sampling for Classifying Rare Classes?* by Kate McCarthy, Bibi Zabar and Gary Weiss analyzed the effectiveness of two methods for dealing with unbalanced class distributions when cost information is known. It compared the effectiveness of cost-sensitive learning with sampling strategies that reduce the class imbalance in the training set. The results indicated that cost-sensitive learning, under-sampling of the majority class and over-sampling of the minority class all perform well on some data sets and poorly on others; with no consistent winner. One question that arises when using sampling to deal with unbalanced data sets is what the sampling rate should be. The talk *Wrapper-based Computation and Evaluation of Sampling Methods for Imbalanced Datasets* by Nitesh Chawla, Lawrence Hall and Ajay Joshi discussed this issue. The authors described a wrapper approach for computing the amount of under-sampling and synthetic generation of minority class examples (a method more sophisticated than over-sampling) to be done in order to best improve classifier performance. The wrapper approach works by doing a guided search of the parameter space using cross-validation to obtain estimates of classifier performance. Experimental results show that this approach significantly outperforms the baselines both in the true positive rate and the average cost per test example.

Improving Classifier Utility by Altering the Misclassification Cost Ratio, by Michelle Ciraco, Michael Rogalewski and Gary Weiss showed that, for two cost-sensitive classifier learning programs, improved performance could be achieved by altering the cost information passed to the learner so that it does *not* match the

actual cost information upon which the classifier is evaluated. These counterintuitive results indicate that existing cost-sensitive learners may contain biases that lead to sub-optimal results.

One of the goals of the workshop was to broaden the scope of cost-sensitive learning, to cover tasks other than predictive classification tasks. Two papers in the workshop addressed this goal. *Utility Based Data Mining for Time Series Analysis—Cost-sensitive Learning for Neural Network Predictors* by Sven Crone, Stefan Lessmann and Robert Stahlbock described how economic utility can be factored into data mining methods for regression and time-series analysis. These methods have typically used symmetric error metrics even though errors in forecasting are often asymmetric. The authors address this issue by developing an asymmetric cost function which is then used as the objective function for training a neural network. The second paper *A Fast High Utility Itemsets Mining Algorithm* by Ying Liu, Weikeng Liao and Alok Choudhary applies the notion of utility to a descriptive data mining task, association rule mining. This work takes into account the fact that each item in a frequent itemset is not of equal value (i.e., utility). The authors consider different values of individual items as utilities and use *utility mining* to identify itemsets with high utilities. Since the downward closure property used in association rule mining does not hold for utility mining, the authors develop a two-phase algorithm to efficiently prune down the number of candidates.

The workshop also included a talk based on a position paper entitled *Contextual Recommender Problems* by Omid Madani and Dennis DeCoste. This paper describes a real-world problem that can potentially be solved using utility-based data mining methods. The problem described is that of placing ads or related links on web pages based on the context information (e.g. contents of the page or information on the current user) such as to maximize some utility function (e.g. the ad click rate). The paper shows the relationship between this problem and multi-armed bandit problems, while highlighting two significant differences: the fact that contexts/arms may not be enumerable and the non-stationarity of user preferences.

2.3 Other

There was one paper that did not fit into the active information acquisition or cost-sensitive learning categories. This paper addressed the issue of utility for the second phase of data mining—the phase associated with building the model. *Interruptible Anytime Algorithms for Iterative Improvement of Decision Trees* by Saher Esmeir and Shaul Markovitch described two interruptible algorithms for inducing decision trees. An anytime algorithm is one that can trade additional resources for improved quality of the output (in this case an improved decision tree). An anytime algorithm is interruptible if the required resources need not be specified up front and if the algorithm can be interrupted at any time and produce a result. One of the interruptible anytime algorithms developed by the authors operates by repeatedly selecting a subtree "whose reconstruction is estimated to yield the highest marginal utility and rebuilding it with higher resource allocation." In this work it is assumed that the algorithm will be interrupted as soon as the decision tree is needed, but the stopping criterion could be modified to incorporate a more global notion of utility. That is, the algorithm should be allowed to run until the cost of allowing it to continue running (based on the need to use the learned model) exceeds the benefit derived from allowing it to run (i.e., improved classification performance).

2.4 Group/Panel Discussion

Authors, speakers and participants contributed to an open discussion. Practitioners commented about the significance and relevance of the problems addressed by the papers presented in the workshop to the data mining industry. It was noted that the industry struggles to satisfy economic constraints but often no appropriate scientific frameworks are available. Participants identified a variety of specific industries and problem domains to which the problems addressed in the papers can be mapped, such as induction from skewed class distributions, costly information acquisitions, and decision costs. It was noted that the workshop had highlighted the range and complexity of issues that the research community currently addresses.

An important discussion addressed the need to focus research efforts on a well-defined research agenda. It was noted that various contributions to the utility-based data mining literature address different objectives, assume different problem settings, and evaluate the work on different and sometimes proprietary data sets. This makes it difficult to compare different solutions and to advance the field, more generally. It was subsequently proposed that it would be important for the community to formulate several generic problems and evaluation measures. It may also be useful to compile relevant data sets for each of the problems. Having a common problem setting and objectives would facilitate comparisons among solutions and allow contributions to build upon earlier work.

3. CONCLUSION

Overall, the UBDM-05 workshop was a success. The workshop published a set of diverse papers, which address many of the goals set out by the workshop organizers. In particular, the papers show how economic utility impacts the three phases of the data mining process outlined in Section 1. In addition, the papers do not focus exclusively on predictive classification tasks, which have dominated work on utility-based data mining.

The workshop was attended by participants from both academia and industry. In particular, the industry participants made it clear that the issues the workshop focused on are issues they struggle with every day. The group/panel discussion also demonstrated that there was a lot of active interest in the topic of utility-based data mining. Due to a tight schedule the program did not leave substantial amounts of time for discussion.

One area that did not receive a lot of attention at the workshop concerned the notion of maximizing utility *throughout* the entire data mining process, as opposed to maximizing utility at just one phase of the process. This is not surprising since this was the first workshop on utility-based data mining, but we hope that in the future researchers and practitioners will begin to address this issue, which will, by necessity, bring together the active learning and cost-sensitive learning communities.

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. We thank the authors, guest speakers, program committee members and attendees for their contributions towards this end. We are indebted to Mohammad Zaki, the KDD-05 Workshop Chair, and to the SIGKDD for organizational and funding assistance. We also thank Foster Provost and our anonymous workshop proposal reviewers for their insightful suggestions and encouragement.