Data Min Knowl Disc (2008) 17:129–135
DOI 10.1007/s10618-008-0117-y

Guest editorial: special issue on utility-based data mining

**Gary M. Weiss · Bianca Zadrozny ·
Maytal Saar-Tsechansky**

Received: 8 July 2008 / Accepted: 17 July 2008 / Published online: 12 August 2008

1 Introduction

Data mining has increasingly been employed in a variety of data-rich domains. As is the case for many new fields of study, at its inception data mining focused on simple scenarios for which methods such as classification, clustering and association mining could provide satisfactory answers. However, the real world scenarios in which data-driven analysis can provide valuable insights are almost always more complex and entail different objectives than those commonly assumed by these data mining techniques. These complexities include opportunities to acquire additional data to improve induction or inference and to recommend decisions that optimize a domain-appropriate utility metric, such as profitability or return on investment. Indeed, as an applied field we should be concerned with how these complexities—and how the deficiencies of current methodologies in taking them into account—pose significant limitations to broadening the adoption of the field and undermine its impact in practice. Utility-Based Data Mining (UBDM) addresses this challenge by taking into account the complex economic environments in which data mining occurs. Our use of the term *utility* corresponds to its use in economics and in our specific context corresponds to the total measure of satisfaction, or expected satisfaction, associated with the *entire*

G. M. Weiss (✉)
Department of Computer and Information Science, Fordham University, Bronx, NY, USA
e-mail: gweiss@cis.fordham.edu

B. Zadrozny
Department of Computer Science, Federal Fluminense University, Niteroi, Brazil
e-mail: bianca@ic.uff.br

M. Saar-Tsechansky
Department of Information, Risk and Operations Management, McCombs School of Business,
University of Texas at Austin, Austin, TX, USA
e-mail: maytal@mail.utexas.edu

data mining process. Specifically, the economic factors relevant for maximizing the utility of this process include (1) the costs and benefits associated with obtaining data, (2) the costs associated with building the predictive model or extracting descriptive patterns using a data mining algorithm, and (3) the costs and benefits derived from utilizing the acquired knowledge.

Consider a direct marketing mail campaign as a motivating example. One of the goals of such a campaign is to optimize a utility measure such as response rate or return on investment, while budgetary constraints may also limit the total mailing costs. To achieve this goal, there are many decisions that must be made. For example, one must consider what historical data should be validated and cleaned, given the cost of such an effort, and what information about potential consumers should be purchased to enhance the objective of the campaign. Should additional small-scale campaigns be conducted before a large-scale campaign in order to obtain more labelled training data—and if so what customers should be targeted in these campaigns? What additional feature values, such as demographics or behavioral characteristics, should be acquired for these training data, and for which individual consumers? Also, when applying the induced model to a new (i.e., test) customer for which a decision is required, should additional feature values be acquired at that point? And, what information would be most cost-effective to acquire? Most of the work in data mining and machine learning has focused on devising better induction algorithms and it is important to consider what algorithms are most appropriate for a given problem. This determination, however, is largely dependent on the economic factors of the situation (i.e., the costs and benefits associated with each decision) and hence one cannot determine what algorithm is best without considering how well the algorithm factors in these economic considerations. Finally, the context in which data mining occurs may also place additional constraints on the data mining algorithm. For example, for some online targeting campaigns it may be beneficial to induce models frequently, and thus time-efficient algorithms and/or anytime induction techniques should be considered. All of these questions must be addressed in order to optimize the total utility of the campaign. The success or failure of a data mining application is heavily dependent on whether such utility factors are considered and optimized—perhaps more so than on whether the most advanced data mining algorithm is used. The articles in this special issue address challenges pertaining to the incorporation of such utility factors in real-world domains.

2 The special issue

In “Cost-sensitive Learning with Conditional Markov Networks,” Sen and Getoor (2008) address the need for taking utility considerations into account in domains where relational information is encoded in the data and the data can be naturally represented by a graph or network. As an example, data that models social relations between groups of individuals can be represented by a social network. Telecommunication companies, retail companies and social networking web sites sometimes construct social network models of their customers to improve the understanding of their behavior and to make predictions. As demonstrated by Sen and Getoor, given such a network it is beneficial to consider not only the costs of misclassifying individual examples, but also the costs associated with misclassifying groups of related examples—which they refer to as relational costs. The article proposes two cost-sensitive relational classifier learning methods. The first one is an extension of conditional Markov networks that estimates conditional probabilities and uses Bayes risk theory to assign each example to the class with the lowest expected cost; the second classifier directly incorporates misclassification costs into the Markov network learning and inference procedures. Sen and Getoor demonstrate that exploiting the correlations in the relational structure can help reduce misclassification costs and incorporating costs into the induction procedure is preferable to using Bayes risk theory ex-post.

Similar to the work by Sen and Getoor (2008), the second article in this issue addresses a problem in which performance is not based solely on the classification performance of individual examples. In "Quantifying Counts and Costs via Classification," Forman (2008) analyzes the quantification problem, where the performance of a classifier is based on how well it estimates global properties of the test data. For count classification the utility is based on how well the classifier estimates the class distribution of the data whereas for cost quantification the utility is based on how well it estimates the total cost associated with misclassifying examples of each class. Quantification problems occur in many domains and the motivating problem discussed by Forman involves estimating the prevalence of support problems in a product line (e.g., the number of cracked computer screens). As Forman effectively demonstrates, the quantification problem is fundamentally different than classification—an insurance company can accurately estimate the total number of accidents by its subscribers without being able to accurately predict which individual subscribers will have an accident. Forman empirically demonstrates that the most straightforward approach for handling quantification tasks, which involves using the most accurate classifier, does not perform very well. Rather, other approaches, which are free to select a decision threshold that is more stable, yield substantially more precise and less biased estimates, especially when the class distribution of the data is highly skewed and the class distribution may change over time. Given the prevalence of quantification problems in practice, there is a stark shortage of work on this subject in the data mining and machine learning literatures. That alone makes this work quite significant and opens up avenues for future research.

One general issue that arises in Utility-Based Data Mining, where the model evaluation metric is selected to best match the utility considerations of the problem domain, is how to ensure that the data mining algorithm optimizes for that metric—and whether such optimization is necessary at all. Forman (2008) shows that this optimization is necessary for the quantification problem, since a classification algorithm optimized for predictive accuracy does not perform as well as one tuned to the quantification problem. In "PRIE: A System for Generating Rulelists to Maximize ROC Performance," Fawcett (2008) shows that a separate-and-conquer rule learning classification algorithm, designed specifically to optimize ROC performance, outperforms standard classification algorithms which seek to optimize predictive accuracy. This work is significant because ROC analysis has virtually become the standard utility metric when misclassification cost information is unavailable; yet almost all research on ROC analysis relies on standard classification algorithms, which were not designed to optimize ROC performance. PRIE chooses rules that extend the convex hull in ROC space and improves the efficiency of the rule generation process by using geometric properties of ROC space to eliminate large numbers of rules that are unlikely to extend this convex hull. This paper is significant to UBDM because it proposes a principled method to *directly* build a rule-based classifier that optimizes ROC performance; the paper also lays the foundation for future advances in ROC-based learning.

Some UBDM-related problems prove quite challenging for data mining induction algorithms. One such issue, which has received a great deal of attention in recent years, concerns learning from imbalanced data sets, where one class is severely underrepresented in the data. Conventional data mining methods often perform poorly in this situation since the underrepresented class typically has a much higher misclassification rate but also higher misclassification costs. One of the most common strategies for addressing this is to resample the data in order to alter the class distribution of the training data—usually so that the data becomes less imbalanced. The simplest variations of this strategy involve undersampling the majority class or oversampling the minority class. However, one drawback is that it is difficult to determine a priori the best sampling strategy, or combination of strategies, and the best sampling rate. In “Automatically Countering Imbalance and its Empirical Relationship to Cost,” Chawla et al. (2008) propose a wrapper paradigm that discovers a good sampling strategy for a variety of common utility measures. The authors also present one of the most comprehensive empirical evaluations of sampling strategies to date. They analyze their results to gain insight into the class imbalance problem and show that their wrapper approach generally outperforms other cost-sensitive learning methods.

Many environments offer opportunities to acquire data that can be used to improve learning and inference. UBDM offers intriguing challenges to devise acquisition policies that enhance the utility derived from data mining. The next two articles in this special issue consider these opportunities. The first of these articles, “Maximizing Classifier Utility when there are Data Acquisition and Modeling Costs”, by Weiss and Tian (2008), analyzes a simple data acquisition setting, where one can decide how many labeled training examples to acquire, but cannot choose which examples to select. This data acquisition scenario may occur, for example, when data is purchased from a third-party data supplier. Weiss and Tian (2008) analyze the relationship between the total utility of the classifier and the number of training examples and provide a progressive sampling strategy for finding a classifier with near-optimal utility. They initially consider only the cost of acquiring training examples and the cost of misclassification errors, and then extend this analysis to include the costs associated with inducing the classification model. While the utility model that is presented is quite simple, it is perhaps the only example of UBDM research that takes into account utility information from all major stages of the data mining process: data acquisition, model induction, and model evaluation. Hopefully this article will stimulate other researchers to incorporate utility considerations in multiple stages of the data mining process.

A particular form of information acquisition, namely active learning (Cohn et al. 1994) in an online environment, is studied by Rokach et al. (2008) in “Pessimistic Cost-sensitive Active Learning of Decision Trees for Profit Maximizing Targeting Campaigns.” The setting they explore is common in practice and offers many intriguing challenges. Specifically, they consider a direct marketing campaign in which information about consumers is available (independent variables) but consumer responses to the offer being solicited are not known and must be acquired. In this setting, it is necessary to solicit offers before a satisfactory model to estimate the utility from each prospective solicitation is available. Thus there are opportunities to develop policies to suggest what solicitations can improve the desired objective the most. The challenge is primarily due to the tension between the costs of acquiring information and the uncertainty, at the time an acquisition decision is made, regarding the utilities that will result from different prospective acquisitions. In an online setting, there is also an interesting tension between the potential to inform future decisions via information acquisition and the desire to maximize the overall campaign profitability. This online learning scenario is common and presents a richer domain than is addressed by traditional active learning.

In the final article in the special issue, "Classification Trees and Decision-analytic Feedforward Control: A Case Study from the Video Game Industry," Brydon and Gemino (2008) demonstrate that the classification methods used to induce classification trees can be extended to provide assistance with decision analysis and, in particular, with reasoning about the probabilistic information embedded within the classification trees. This research demonstrates how existing data mining methods can be modified to provide decision analysis for a complex real world domain—video game development. Brydon and Gemino (2008) make use of extensive data from the video game industry to determine the development choices that are likely to yield the greatest possible expected value. This work is also notable in that it provides a realistic case study of a complex domain; the data mining literature can greatly benefit from similar efforts to offer insights into important real-world challenges.

3 Challenges and the future of utility-based data mining

As demonstrated by the articles in this special issue and those which appeared in the two recent workshops on Utility-Based Data Mining (Weiss et al. 2005; Zadrozny et al. 2006), there is significant ongoing research on UBDM. In this section we offer some general comments about the state-of-the-art and suggest specific areas for future research.

Most work in UBDM either aims at improving a model's predictive accuracy, or, in the case of cost-sensitive classification, reducing the cost of a model's misclassification errors. However, other related utility objectives are desirable in practice. For example, classification models are often used to estimate the probability of uncertain outcomes (classes), or to rank the likelihood of such outcomes; thus designing induction techniques to optimize these objectives is desirable. The papers by Fawcett (2008) and Forman (2008) in this special issue demonstrate the benefits of tuning existing induction algorithm to directly optimize two important evaluation metrics.

Furthermore, not only during model induction or inference should a wider variety of relevant utility objectives be considered, but also during other stages of the data mining process, such as the data acquisition stage. Most work on information acquisition for classifier induction considers improving a single model's classification accuracy. Recent work has begun to consider other related utilities, including improving a given model's class probability estimation (Saar-Tsechansky and Provost 2004; Melville et al. 2005) or the utility implications of decisions informed by predictive modeling (Saar-Tsechansky and Provost 2007; Rokach et al. 2008). However, more research is required to establish theoretical foundations for improving utilities of arbitrarily complex, real-world settings via information acquisition. These settings include *online* model induction as addressed by Rokach et al. (2008) and settings in which decision-making informed by the inference involves *multiple* uncertain outcomes (classes) or when *multiple models* are used to inform decisions.

Most existing work in UBDM relates to supervised learning, namely prediction tasks. However, utility considerations also impact descriptive data mining tasks, such as clustering and association rule mining. Although there has been some work in this area (Yao and Hamilton 2006; Shen et al. 2002), additional work is needed to ensure that the utility considerations associated with these tasks are not neglected.

Finally, as an applied field, UBDM is in an acute need for comprehensive application papers, such as by Brydon and Gemino (2008), where utility considerations play a central role in the research. Such papers are essential both to provide insight into the complexities associated with real-world problems and to offer new research questions. An important and related concern is the need for publicly available data and information about the associated problem settings, such as costs and utility objectives. Discussions with data mining researchers and practitioners in leading companies during two recent workshops on Utility Based Data Mining (Weiss et al. 2005; Zadrozny et al. 2006) brought up the need to break the primacy of predictive accuracy as an evaluation metric and shift the emphasis toward applied objective measures. While objectives such as profitability are often a function of accuracy, there are many practical settings, including information acquisition and induction from imbalanced class problems, where a focus on accuracy is less appropriate. Data from real problem settings, along with the associated utility objectives, would aid the development of new theoretical foundations and methodologies to address these issues.

Data mining has made great strides over the past two decades, including recent work on UBDM which attempts to better align data mining with the utilities of complex, real-work decisions. Continued progress in the field requires significant and consistent attention to these challenges, and we hope that this special issue contributes to this cause. We hope that you find the articles insightful and that they will inspire you to address these and other important utility-based data mining challenges.

Acknowledgements We would like to thank all the authors who submitted papers for consideration to the special issue. We would especially like to thank the reviewers for their time and thoughtful reviews. Finally, we would like to thank the Editor-in-Chief, Geoffrey Webb, for his valuable advice and encouragement, and the editorial staff for their support in the production of this special issue.

References

Brydon M, Gemino A (2008) Classification trees and decision-analytic feedforward control: a case study from the video game industry. Data Min Knowl Discov 17(2). doi:10.1007/s10618-007-0086-6

Chawla NV, Cieslak DA, Hall LO, Joshi A (2008) Automatically countering imbalance and its empirical relationship to cost. Data Min Knowl Discov 17(2). doi:10.1007/s10618-008-0087-0

Cohn D, Atlas L, Ladner R (1994) Improving generalization with active learning. Mach Learn 15(2): 201–221

Fawcett T (2008) PRIE: a system for generating rulelists to maximize ROC performance. Data Min Knowl Discov 17(2). doi:10.1007/s10618-008-0089-7

Forman G (2008) Quantifying counts and costs via classification. Data Min Knowl Discov 17(2). doi:10.1007/s10618-008-0097-y

Melville P, Yang SM, Saar-Tsechansky M, Mooney RJ (2005) Active learning for probability estimation using Jensen-Shannon divergence. In: Proceedings of the 16th European conference on machine learning (ECML), Porto, Portugal

Rokach L, Naamani L, Shmilovici A (2008) Pessimistic cost-sensitive active learning of decision trees for profit maximizing targeting campaigns. Data Min Knowl Discov 17(2). doi:10.1007/s10618-008-0105-2

Saar-Tsechansky M, Provost F (2004) Active sampling for class probability estimation and ranking. Mach Learn 54(2):153–178. doi:10.1023/B:MACH.0000011806.12374.c3

Saar-Tsechansky M, Provost F (2007) Decision-centric active learning of binary-outcome models. Inf Syst Res 18(1):1–19. doi:10.1287/isre.1070.0111

Sen P, Getoor L (2008) Cost-sensitive learning with conditional Markov networks. Data Min Knowl Discov 17(2). doi:10.1007/s10618-008-0090-5

Shen Y, Zhang Z, Yang Q (2002) Objective-oriented utility-based association mining. In: Proceedings of the 2002 IEEE international conference on data mining, Maebashi City, Japan, pp 426–433

Weiss GM, Tian Y (2008) Maximizing classifier utility when there are data acquisition and modeling costs. Data Min Knowl Discov 17(2). doi:10.1007/s10618-007-0082-x

Weiss GM, Saar-Tsechansky M, Zadrozny B (eds) (2005) Proceedings of the first international workshop on utility-based data mining, August 2005. ACM Press, Chicago, IL

Yao H, Hamilton HJ (2006) Mining itemset utilities from transaction databases. Data Knowl Eng 59(3): 603–626. doi:10.1016/j.datak.2005.10.004

Zadrozny B, Weiss GM, Saar-Tsechansky M (eds) (2006) Proceedings of the second international workshop on utility-based data mining, August 2006. ACM Press, Philadelphia, PA