Machine Learning, 57, 5–11, 2004

Editorial: Data Mining Lessons Learned

NADA LAVRAČ nada.lavrac@ijs.si
Institute Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia; Nova Gorica Polytechnic, Vipavska 13, 5000 Nova Gorica, Slovenia

HIROSHI MOTODA motoda@sanken.osaka-u.ac.jp
Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

TOM FAWCETT tom.fawcett@hp.com
Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304, USA

Introduction

Data mining is concerned with finding interesting patterns in data. Many techniques have emerged for analyzing and visualizing large volumes of data. What one finds in the technical literature are mostly success stories of these techniques. Researchers rarely report on steps leading to success, failed attempts, or critical representation choices made; and rarely do papers include expert evaluations of achieved results. An interesting point of investigation is also why some particular solutions, despite good performance, were never used in practice or required additional treatment before they could be used. Insightful analyses of successful and unsuccessful applications are crucial for increasing our understanding of machine learning techniques and their limitations.

The UCI Repository of Machine Learning Databases (Blake & Merz, 1998) has served the machine learning community for many years as a valuable resource. It has benefited the community by allowing researchers to compare algorithm performance on a common set of benchmark datasets, most taken from real-world domains. However, its existence has indirectly promoted a very narrow view of real-world data mining. Performance comparisons, which typically focus on classification accuracy, neglect important data mining issues such as data understanding, data preparation, selection of appropriate performance metrics, and expert evaluation of results. Furthermore, because the UCI repository has been used so extensively, some researchers have claimed that our algorithms may be "overfitting the UCI repository".

Challenge problems such as the KDD Cup, CoIL and PTE challenges have also become popular in recent years and have attracted numerous participants. Contrary to the "UCI challenge" of achieving best accuracy results in many different domains, these challenge problems usually involve a single difficult problem domain, and participants are evaluated by how well their entries satisfy a domain expert. Such challenges can be a very useful source of feedback to the research community, provided that thorough analysis of results has been performed (for example, Elkan, 2001, describes lessons from the CoIL Challenge 2000).

With this background in mind, a workshop on *Data Mining Lessons Learned* (DMLL-2002) was organized at the *Nineteenth International Conference on Machine Learning* (ICML-2002) in Sydney in July of 2002 in order to gather and extract lessons from data mining endeavors. This workshop featured three invited talks and ten contributing authors presenting the different aspects of practical data mining applications and data mining competitions, together with their lessons learned. These reports are available in the on-line DMLL-2002 proceedings (Lavrač, Motoda, & Fawcett, 2002).

In early 2003 a call for papers was issued for a special issue of the *Machine Learning* journal. In contrast to a previous special issue of *Machine Learning* (volume 30, issue 2–3) on applications, the main goal of this special issue is to focus on the lessons learned from the data mining process rather than on the applications themselves. The result is this issue containing this editorial, one introductory paper and six contributed papers. The aim of this special issue is to gather experiences gained from data mining applications and challenge competitions, in terms of the lessons learned both from successes and from failures, from the engineering of representations for practical problems, and from expert evaluations of solutions.

Lessons of the articles in this special issue

The introductory paper to this special issue (Lavrač et al., 2004a) presents lessons from data mining applications, including experience from science, business, and knowledge management in a collaborative data mining setting. The six contributed papers involve diverse domains and address various data mining problems encountered with their respective tasks. While some of the lessons are rather specific, there are common themes among the articles. We discuss the most prominent of these themes as manifested in the articles.

Importance of representation

One of the recurring lessons of data mining applications is the importance of instance representation; that is, the choice of the feature set used to express examples. We see this issue arise repeatedly in the papers in this issue. Real-world domains are often characterized by concepts dispersed over large numbers of features (the so-called curse of dimensionality). This problem is prominent in two papers of this issue: the paper on classifying brain scans (Mitchell et al., 2004) and the analysis of the CoIL Challenge 2000 (Putten & Someren, 2004).

Mitchell et al. (2004) faced difficult representation problems in their work on classifying human cognitive states from brain activity. Their data were very high dimensional (10^5 features), sparse (fewer than a dozen training examples per class) and noisy. Their task is to classify the cognitive state of a human subject (for example, to determine whether the subject is looking at a picture or a sentence) based on functional Magnetic Resonance Imaging (fMRI) scans. Such scans produce a great amount of data: an fMRI scan produces a series of three dimensional brain images, each containing approximately 15,000 volume elements called voxels. These scans are performed once per second, yielding tens of millions

of voxels. Thus the number of features at one time point is large. In contrast, the number of sessions is less than a dozen.

The researchers accommodated this sparseness in two ways. They adapted their training to perform k-fold cross validation, where each fold includes one example per class. Recognizing that feature selection was critical they investigated the interactions of classifier types with different selection strategies. Prior work has shown that when the number of training examples is small relative to the number of features, generative classifiers such as Gaussian Naive Bayes (GNB) often outperform discriminative classifiers such as Support Vector Machines (SVM). The results of Mitchell et al. are consistent with this observation. The SVM performed more poorly than GNB when no feature selection was used but outperformed GNB when using feature selection. Further, these two outperformed K-NN consistently, especially when no feature selection was attempted.

The researchers also had to deal with the anatomical variability of their human subjects, which made analysis of fMRI much more difficult. However, appropriate abstraction makes it possible to map one brain into another, i.e., by averaging the voxels in a particular region of interest into a "supervoxel". With this data preparation it is indeed possible to train a classifier to capture significant subject-independent regularities in brain activity that are sufficiently strong to detect single interval cognitive states in human subject who are not part of the training set.

Finally, Mitchell et al. (2004) report an interesting set of feature selection experiments. The task in this domain is characterized as having a *zero-signal* learning setting: there is a class corresponding to a state where there is no activation signal corresponding to the target class. Feature selection can exploit this property. Feature selection in this case is equivalent to selecting informative voxels. Extensive experiments with feature selection showed that discrimination-based feature selection overfits the data and activity-based feature selection outperforms it in zero-signal classification problems, especially with increasing data dimension, noise, and sparsity, and as the proportion of truly relevant features decreases.

Van der Putten and van Someren (2004) discuss a very different but similarly challenging domain: that of determining who would be inclined to buy a specific insurance product. This was the task of the CoIL Challenge 2000, and van der Putten and van Someren analyzed the contestants' submissions. The domain is characterized by noisy, correlated, redundant and high dimensional data with a weak relation between input and target. The quality of the reported results varied substantially across participants. Van der Putten and van Someren base their analysis on a bias-variance decomposition of the error. They find that a key problem is to avoid overfitting, and they estimate the effect of feature construction, feature selection and method selection on the solution accuracy. In general, improved modeling reduces the bias component in the error but tends to increase the variance component. They conclude that feature selection in such a domain is even more important than the choice of a learning method.

A final lesson in the importance of representation is offered by Krogel and Scheffer (2004), who address the task of learning from microarray experiments. Among the problems they faced was that of making effective use of the interaction between genes, which was given as relational data spread over multiple tables. Their solution to this is to propositionalize the relational data, by first counting joins of tables and then collapsing the joins

by aggregation functions (e.g., count, max and min). Aggregation solves the problem of the combinatorial explosion that results from joins and scales well for difficult problems. The resulting attributes contain relevant information despite some loss of information incurred by the aggregation.

Importance of context and external criteria

Data mining is often distinguished from machine learning by its concern with the entire process of data analysis and induction. This process includes the context of the problem and the intended use of the results. Such concerns are prominent in the articles of this issue.

A common requirement of the data mining process is that the results be understandable to, and actionable by, experts in the domain. Sometimes researchers can forget this goal in their pursuit of superior prediction performance. Domain experts tend to place a high value on simplicity and comprehensibility of results, more so than do most data mining researchers (or their algorithms).

For example, van der Putten and van Someren's (2004) analysis of the CoIL Challenge 2000 included an expert evaluation of the challenge entries. This was the so-called "descriptive" portion of the challenge, in which the goal was to convey the learned knowledge to a domain expert in a form that was both intelligible and actionable; in other words, the expert should be able to understand the knowledge and should be able to put it to use. However, the expert concluded that "Almost all entries lack a good description in words: participants seem to forget that most marketeers find it difficult to read statistics (and understand them)". So while the entries may have been well researched and quantitatively evaluated, they fell short of being useful to the people for whom the results were intended.

The need for intelligible results also arises in the work of Lavrač et al. (2004) on subgroup discovery. They are concerned with finding simple rules describing significant subpopulations of instances. One requirement was that domain experts understand the resulting subgroup descriptions. Lavrač et al. note that the quantitatively best-performing rules are not necessarily the ones most interesting to experts. The ROC convex hull method was used to evaluate rules, and the researchers observed that rules preferred by the expert were usually not on the hull. They also mention that when describing subgroups with a set of supporting factors, it is important to be able to generalize them into a metaphoric description. Kohavi et al. (2004) agree with the need to produce comprehensible models. Efficient algorithms whose output is comprehensible for business insight, and which can handle multiple data types (cyclic data, hierarchical attributes, different granularity data) are yet to be designed. Kohavi et al. pose this as a challenge for data mining.

External context is also necessary in order to identify an actionable target class for the data. Kohavi et al. (2004) phrase this as the ability to transform business questions into a worthwhile data mining tasks, and they identify it as a challenge of mining retail e-commerce data. The same issue arises in the work of Lavrač et al. (2004). Choosing instances to categorize in the target class requires deep understanding of the task to solve. One of their domains involved targeting potential consumers of a natural non-alcoholic sparkling drink. This task might be approached naïvely as a task to identify potential consumers of this drink from among all consumers. However, guidance from a marketing expert helped to refine

the target population in two ways. First, people who were not regular drinkers of sparkling beverages were excluded on the basis that they would not be inclined to use the product generally. Second, consumers of a strongly positioned competing brand were excluded on the basis that they were very unlikely to change. The resulting target population, with these two un-promising sub-populations removed, is likely to produce more useful results from data mining.

Shortcomings of conventional algorithms

The demands of real-world data mining sometimes require modifications and extensions to conventional algorithms. We observe this in several papers.

Hsu, Chung, and Huang (2004) propose a new method for personalized shopping recommendation after having tried various existing approaches. The analysis of transaction data of two retail companies revealed that the data are very skewed and sparse. The number of items in these stores is about 1000, but each customer purchases only 4 to 6 items at a time. The sales are also skewed and concentrated on a very small portion of products (this phenomenon is commonly characterized as the "80–20 rule" in business management, i.e., 20% of people own 80% of accumulated wealth). It is critical to identify potential customers for the products in the tail of this skewed product curve.

With such skewed and sparse transactions, association rule mining performs poorly even after proper preprocessing. Content-based filtering could perform well but it cannot be applied easily to shopping recommendation because products are usually heterogeneous and incomparable. Collaborative filtering techniques are applicable but require rating information from customers, which is not available from transaction records. Faced with these difficulties, the authors devised a probabilistic model that solved the problem. Existing collaborative filtering recommenders can be cast as instances of the three-node probabilistic graphical model (user, rating, their relation; or user, preference, their relation). A new model can be built within this framework by combining user cluster model and aspect model with two latent variables. This model allows recommendations to be generated for products that a customer has not purchased.

Similarly, Lavrač et al. (2004) faced difficulties using conventional rule learning algorithms for subgroup discovery. The task of subgroup discovery differs from that of classification and requires a rule learner that generates characteristic, rather than discriminative, rules. Rule learning algorithms are able to generate such rules, but the use of standard covering algorithms is not suitable. Only the first few rules generated by the covering algorithm may produce interesting subgroup descriptions with sufficient coverage, while subsequent rules are induced from smaller and strongly biased example sets, excluding the positive examples covered by previously induced rules. This bias hinders the covering algorithm from inducing descriptions uncovering significant subgroup properties of the entire population.

Krogel and Scheffer (2004) studied the effectiveness of transduction and co-training for exploiting unlabeled data to analyze the benefit of text classification and information extraction for utilizing a collection of scientific abstracts. The results were contradictory to what are normally expected. Use of unlabeled data is commonly believed to

be beneficial. The results of using transductive SVM and co-training were disappointing and concluded to be less generally applicable. Taking unlabeled data into account is expected at least not to hurt performance and be beneficial when there are not enough labeled examples. Neither of these two hold for microarray data. Why this happened is not well understood.

Contributions

Each lesson contributed by a "lessons learned" article may be placed into one of three general categories.

1. A lesson may point out a problem, shortcoming or violated assumption of an existing method. In this role, a "lessons learned" paper contributes to the field by guiding research into new areas and by providing feedback on how well researchers' assumptions hold in the real world. For example, numerous applications papers have mentioned that skewed class distributions, unequal error costs and very sparse data are challenges for existing methods. Each of these problems has motivated papers, workshops and Ph.D. theses, and each remains an active topic in data mining.
2. The lessons may contribute to a general understanding of the importance of a topic or issue. For example, some papers in this issue emphasize the importance of representation. They serve to document the effort required to devise a useful feature set enabling a problem to be solved. Unlike the lessons in category one, such lessons may not be ready for immediate scientific investigation but they can motivate exploratory research. For example, an extensive investigation into how people generate useful representations or integrate business goals with data mining goals would be an ambitious undertaking but the results could be very valuable.
3. Finally, a "lessons learned" paper contributes to the practice of data mining by serving as a well-documented case study. It can provide advice and guidance to people working on related problems. Both positive and negative lessons can be useful in directing the effort of others.

We believe the articles in this special issue contribute to the field in these ways, and we hope they lead to the advancement of methods and practice of data mining.

Acknowledgments

We thank the organizers of the Nineteenth International Conference on Machine Learning (ICML-2002) for their help in organizing the workshop on *Data Mining Lessons Learned* (DMLL-2002) in Sydney in July 2002. We thank the reviewers of papers submitted to this special issue; and to Rob Holte, former Editor-in-Chief of the *Machine Learning* journal, for many valuable suggestions that helped us compose this special issue.

References

Blake, C. L., & Merz, C. J. (1998). *UCI Repository of machine learning databases.* Irvine, CA: University of California, Department of Information and Computer Science. Available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Elkan, C. (2001). Magical thinking in data mining: Lessons from CoIL Challenge 2000. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 426–431).

Hsu, C. N., Chung, H. H., & Huang, H. S. (2004). Mining skewed and sparse transaction data for personalized shopping recommendation. *Machine Learning, 57:1/2*, 35–59.

Kohavi, R., Mason, L., Parekh, R., & Zheng, Z. (2004). Lessons and challenges from mining retail e-commerce data. *Machine Learning, 57:1/2*, 83–113.

Krogel, M. A., & Scheffer, T. (2004). Multi-relational learning, text mining, and semi-supervised learning for functional genomics. *Machine Learning, 57:1/2*, 61–81.

Lavrač, N., Motoda, H., & Fawcett, T. (eds.) (2002). *Proceedings of the First International Workshop on Data Mining Lessons Learned*, DMLL-2002, held in conjunction with ICML-2002, Sydney, July 2002. Available: `http://www.hpl.hp.com/personal/Tom_Fawcett/DMLL-2002/Proceedings.html`

Lavrač, N., Cestnik, B., Gamberger, D., & Flach, P. (2004a). Decision support through subgroup discovery: Three case studies and the lessons learned. *Machine Learning, 57:1/2*, 115–143.

Lavrač, N., Motoda, H. Fawcett, T., Holte, R., Langley, P., & Adriaans, P. (2004). Introduction: Lessons learned from data mining applications and collaborative problem solving. *Machine Learning, 57:1/2*, 13–34.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., & Wang, X. (2004). Learning to decode cognitive states from brain images. *Machine Learning, 57:1/2*, 145–175.

van der Putten, P., & van Someren, M. (2004). A bias-variance analysis of a real world learning problem: The CoIL challenge 2000. *Machine Learning, 57:1/2*, 177–195.