Postprocessing in Machine Learning and Data Mining

Ivan Bruha

Dept. Computing & Software
McMaster University
Hamilton, Ont., Canada L8S 4L7

email:bruha@mcmaster.ca
http://www.cas.mcmaster.ca/~bruha

A. (Fazel) Famili

Institute for Information Technology
National Research Council of Canada
Ottawa, Ont., Canada K1A 0R6

email: Fazel.Famili@iit.nrc.ca
http://www.iit.nrc.ca/~fazel

ABSTRACT

This article surveys the contents of the workshop *Post-Processing in Machine Learning and Data Mining: Interpretation, Visualization, Integration, and Related Topics* within *KDD-2000: The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, USA, 20-23 August 2000. The corresponding web site is on

www.acm.org/sigkdd/kdd2000

First, this survey paper introduces the state of the art of the workshop topics, emphasizing that postprocessing forms a significant component in Knowledge Discovery in Databases (KDD). Next, the article brings up a report on the contents, analysis, discussion, and other aspects regarding this workshop. Afterwards, we survey all the workshop papers. They can be found at (and downloaded from)

www.cas.mcmaster.ca/~bruha/kdd2000/kddrep.html

The authors of this report worked as the organizers of the workshop; the programme committee was formed by additional three researches in this field.

1. POSTPROCESSING IS A SIGNIFICANT COMPONENT OF KDD

Knowledge Discovery in Databases (KDD) has become a very attractive discipline both for research and industry within the last few years. Its goal is to extract "pieces" of knowledge from usually very large databases. It portrays a robust sequence of procedures that have to be carried out so as to derive reasonable and understandable results [11], [16].

The data that are to be processed by a knowledge acquisition algorithm are usually noisy and often inconsistent [4]. Many steps must be performed before the actual data analysis starts. Therefore, certain *preprocessing* procedures have to precede the actual data analysis process. Next, a result of a knowledge acquisition algorithm, such as a decision tree, a set of decision rules, or weights and topology of a neural net, may not be appropriate from the view of custom or commercial applications. As a result, a concept description (model, knowledge base) produced by such an inductive process has to be usually postprocessed. *Postprocessing* procedures usually include various pruning routines, rule filtering, or even knowledge integration. All these procedures provide a kind of symbolic filter for noisy and imprecise knowledge derived by an inductive algorithm. Therefore, some preprocessing routines as well as postprocessing ones should fill up the entire chain of data processing.

Research in knowledge discovery is supposed to develop methods and techniques to process large databases in order to acquire knowledge (which is "hidden" in these databases) that is compact, more or less abstract, but understandable, and useful for further applications. The paper [12] defines knowledge discovery as a nontrivial process of identifying valid, novel, and ultimately understandable knowledge in data.

In our understanding, knowledge discovery refers to the overall process of determining useful knowledge from databases, i.e. extracting high-level knowledge from low-level data in the context of large databases. Knowledge discovery can be viewed as a multidisciplinary activity because it exploits several research disciplines of artificial intelligence such as machine learning, pattern recognition, expert systems, knowledge acquisition, as well as mathematical disciplines such as statistics, theory of information, uncertainty processing.

The entire chain of knowledge discovery consists of the following steps:

(1) *Selecting the problem area.* Prior to any processing, we first have to find and specify an application domain, and to identify the goal of the knowledge discovery process from the customer's viewpoint. Also, we need to choose a suitable representation for this goal.

(2) *Collecting the data.* Next, we have to choose the object representation, and collect data as formally represented objects. If a domain expert is available, then he/she could suggest what fields (attributes, features) are the most informative. If not, then the

simplest method is to measure everything available.

(3) *Preprocessing of the data.* A data set collected is not directly suitable for induction (knowledge acquisition); it comprises in most cases noise, missing values, the data are not consistent, the data set is too large, and so on. Therefore, we need to minimize the noise in data, choose a strategy for handling missing (unknown) attribute values (see e.g. [5], [7], [14]), use any suitable method for selecting and ordering attributes (features) according to their informativity (so-called attribute mining), discretize/ fuzzify numerical (continuous) attributes [3], [10], and eventually, process continuous classes.

(4) *Data mining: Extracting pieces of knowledge.* We reach the stage of selecting a paradigm for extracting pieces of knowledge (e.g., statistical methods, neural net approach, symbolic/logical learning, genetic algorithms). First, we have to realize that there is no optimal algorithm which would be able to process correctly any database. Second, we are to follow the criteria of the end-user; e.g., he/she might be more interested in understanding the model extracted rather than its predictive capabilities. Afterwards, we apply the selected algorithm and derive (extract) new knowledge.

(5) *Postprocessing of the knowledge derived.* The pieces of knowledge extracted in the previous step could be further processed. One option is to simplify the extracted knowledge. Also, we can evaluate the extracted knowledge, visualize it, or merely document it for the end user. They are various techniques to do that. Next, we may interpret the knowledge and incorporate it into an existing system, and check for potential conflicts with previously induced knowledge.

Most research work has been done in the step 4. However, the other steps are also important for the successful application of knowledge discovery in practice.

Postprocessing as an important component of KDD consists of many various procedures and methods that can be categorized into the following groups.

(a) Knowledge filtering: Rule truncation and postpruning. If the training data is noisy then the inductive algorithm generates leaves of a decision tree or decision rules that cover a very small number of training objects. This happens because the inductive (learning) algorithm tries to split subsets of training objects to even smaller subsets that would be genuinely consistent. To overcome this problem a tree or a decision set of rules must be shrunk, by either postpruning (decision trees) or truncation (decision rules); see e.g. [17].

(b) Interpretation and explanation. Now, we may use the acquired knowledge directly for prediction or in an expert system shell as a knowledge base. If the knowledge discovery process is performed for an end-user, we usually document the derived results. Another possibility is to visualize the knowledge [9], or to transform it to an understandable form for the user-end. Also, we may check the new knowledge for potential conflicts with previously induced knowledge. In this step, we can also summarize the rules and combine them with a domain-specific knowledge provided for the given task.

(c) Evaluation. After a learning system induces concept hypotheses (models) from the training set, their evaluation (or testing) should take place. There are several widely used criteria for this purpose: classification accuracy, comprehensibility, computational complexity, and so on.

(d) Knowledge integration. The traditional decision-making systems have been dependant on a single technique, strategy, model. New sophisticated decision-supporting systems combine or refine results obtained from several models, produced usually by different methods. This process increases accuracy and the likelihood of success.

2. WORKSHOP REPORT

This workshop was addressing an important aspect related to the Data Mining (DM) and Machine Learning (ML) in postprocessing and analyzing knowledge bases induced from real-world databases.

Results of a genuine ML algorithm, such as a decision tree or a set of decision rules, need not be perfect from the view of custom or commercial applications. It is quite known that a concept description (knowledge base, model) discovered by an inductive (knowledge acquisition) process has to be usually processed by a postpruning procedure. Most existing procedures evaluate the extracted knowledge, visualize it, or merely document it for the end user. Also, they may interpret the knowledge and incorporate it into an existing system, and check it for potential conflicts with previously derived knowledge (models). Postprocessing procedures thus provide a kind of "symbolic filter" for noisy, imprecise, or "non-user-friendly" knowledge derived by an inductive algorithm.

Consequently, the postprocessing tools are complementary to the DM algorithms and always help the DM algorithms to refine the acquired knowledge. Usually, these tools exploit techniques that are not genuinely logical, e.g., statistics, neural nets, and others.

The presentation and discussion within this workshop revealed the following:

- Four papers (i.e., half of accepted ones) were dealing with the association rules and their postprocessing. It indicates that the above topic is under immerse research.
- Nevertheless, also the other disciplines of postprocessing were presented. Two papers discussed evaluation of knowledge bases induced (rule qualities and interestingness measures). One paper brought up the knowledge revision; one the knowledge combination; one the visualization. Some papers also dealt with knowledge filtering.
- There is a need in commercial applications for more robust postprocessing methods since not only the databases but also the knowledge bases (models, rule sets) can reach extremely large sizes.

As for the workshop itself, there was one invited talk (A. Famili: "Post-processing: The real challenge") that provided an overview of postprocessing. It discussed some typical applications of these techniques to real-world data and explained why we need and where we use the results of postprocessing. Some examples from his past experience were given, too.

Fourteen research papers were submitted to this workshop. Each paper was reviewed by three members of the programme committee. After reviewing, eight of them were selected for publication, i.e. the acceptance rate was 57%.

The authors of this report worked as the organizers of the workshop; the programme committee was also formed by additional three researches in this field:

Petr Berka, Laboratory of Intelligent Systems, University of Economics, Prague, Czech Republic
email: berka@vse.cz
http://lisp.vse.cz/~berka

Marko Bohanec, Institute Jozef Stephan, Jamova 37, Ljubljana, Slovenia
email: marko.bohanec@ijs.si
http://www-ai.ijs.si/MarkoBohanec/mare.html

W.F.S. (Skip) Poehlman, McMaster University, Hamilton, Canada
email: skip@church.cas.mcmaster.ca

3. SURVEY OF WORKSHOP PAPERS

3.1 B. Baesens, S. Viaene, J. Vanthienen: Post-processing of Association Rules

The authors of this paper explain the motivation for a post-processing phase to the association rule mining algorithm when plugged into the knowledge discovery in databases process. They focus on processing of large sets of association rules.

The technique of association rules allows one to discover intra-transactional records [1]. Over the last couple of years, one could see a surge in research on improving the algorithmic performance of the original algorithms, among them the author selected the Apriori algorithm [2] as a starting point.

A strong element of the association rule mining is its ability to discover all associations that exist in the transaction database. Unfortunately, this leads to sets of very large number of rules that are hard to understand. To overcome this drawback, the authors exploit the postprocessing and provide a basic rationale for post-processing the patterns generated by an association rule mining process.

3.2 F. Chung, C. Lui: A post-analysis Framework for Mining Generalized Association Rules with Multiple Minimum Supports

Chung and Lui also work in the field of postprocessing of association rules. They discuss the problem of mining association rules with multiple minimum support. Their algorithm is applied in such a way that the low-level rules have enough minimum support while the high-level rules are prevented from combinatorial explosion.

The authors utilize the generalized association rules [15] and multiple-level association rules [13]. They developed a postprocessing framework for finding frequent itemsets with multiple minimum supports. The explanation is accompanied by many graphs, tables, and illustrative examples.

3.3 J.P. Feng: Meta-CN4 for Unknown Attribute Values Processing Via Combiner and Stack Generalization

This paper introduces two meta-learning methods of combiner and stacked generalizer [8] in the inductive algorithm CN4 [6] with six routines for unknown attribute values processing.

In order to improve the performance of learning algorithms the idea of multistrategy (meta-strategy) learning was initiated. The principle of the combiner and stack generalizer consists of combining the decisions of several classifiers by a meta-classifier (a ‘supervisor’ classifier). The experiments proved that such a knowledge combination exhibits better performance than that of single classifiers.

The author exploits the above knowledge combination mechanism for processing of unknown (missing) attribute values. It is known that no routine for unknown attribute values processing is the best for all potential databases. One possible solution to this problem is to try experimentally which routine fits a given database. Another solution was proposed by the author. A database is processed by all six routines (that are available in the covering algorithm CN4). As a result we get six classifiers; the meta-classifier (combiner or stack generalizer) combines the decisions of these classifiers to get the final decision.

3.4 F. Franek, I. Bruha: Post-processing of Qualities of Decision Rules Within a Testing Phase

This paper introduces a new strategy that allows one to modify (refine) rule qualities during the classification of unseen objects.

If a classifier uses an unordered set of decision rules a problem arises concerning what to do if the classification of an unseen object ‘fires’ rules of different classes. One possible solution consists in calculating a numerical factor that explicitly indicates a quality (predictive power) of each rule, giving thus a higher priority to the rule(s) with a higher quality. In existing models, the rule qualities are calculated by a learning (data mining) algorithm and remain constant during the phase of classification.

The refinement is carried out in a feed-back loop so that it can be viewed as a postprocessing procedure.

3.5 Y. Ma, C.K. Wong, B. Liu: Effective Browsing of the Discovered Association Rules Using the Web

This is another paper bringing up the association rules. Interpreting the discovered knowledge to gain a good understanding of the domain is one of the important phases of KDD postprocessing. To expound a set of association rules is not a trivial task since the size of the complete set of these rules is usually very large.

The authors describe their system DS-Web that assists users in interpreting a set of association rules. They firstly summarize the set of rules in order to build a hierarchical structure for easy browsing of the complete set of rules. Then they propose to publish this hierarchy of rules via multiple web pages connected by hypertext links.

3.6 A.E. Prieditis: VizLearn: Visualizing Machine Learning Models and Spacial Data

This paper introduces VizLearn, a visually-interactive machine learning system. This exploratory system can visualize machine learning models and data. It treats data as if it were from a geographical source by augmenting the original model with so-called fields. The system visualizes certain patterns at-a-glance that would otherwise be difficult to grasp by using non-visual methods.

VizLearn uses Bayesian networks for the knowledge representation because it permits flexibility in queries for classification. Also, it can handle both discrete and real-value data. It can be used to process unknown (missing) values. The author is currently extending VizLearn by probabilistic data brushing and abstraction.

The paper comprises quite a few figures that illustrate the characteristics of the author's system.

3.7 J. Smid, P. Svacek, J. Smid: Processing User Data for Intelligent Tutoring Models

Intelligent tutoring systems are based on a user and diagnostic models. These models must be validated by using a database of user test results. Consequently, the entire model is to be learned from a database of the user test data.

The authors propose a new model for intelligent tutorial system and discuss how to obtain data that specify this model, including their refinement.

3.8 P. Tan, V. Kumar: Interestingness Measures for Association Patterns: A Perspective

Another paper dealing with the association rules. The authors realized that the size of a set of association rules is usually extremely large. Therefore, there exists a need to prune the discovered rules according to their degree of interestingness. The interestingness is, in fact, equivalent to the idea of the rule qualities discussed in another paper of this workshop.

The authors in this paper present and compare various interestingness measures for association patterns that are proposed in statistics, machine learning, and data mining. They also introduce a new metric and show that it is highly linear with respect to the correlation coefficient for many interesting association patterns.

4. REFERENCES

[1] Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in massive databases. Proc. ACM SIGMOD International Conference Management of Data, Washington, D.C. (1993).

[2] Agrawal, R., and Srikant. R. Fast algorithms for mining association rules. Proc. International Conference Very Large Databases, Santiago, Chile (1994).

[3] Berka, P., and Bruha, I. Empirical comparison of various discretization procedures. International Journal of Pattern Recognition and Artificial Intelligence, 12, 7 (1998), 1017-1032.

[4] Brazdil, P., and Clark, P. Learning from imperfect data. In: Brazdil, P., and Konolige, K. (eds.), Machine Learning, Meta-Reasoning, and Logics, Kluwer, (1990), 207-232.

[5] Bruha, I., and Franek, F. Comparison of various routines for unknown attribute value processing: Covering paradigm. International Journal of Pattern Recognition and Artificial Intelligence, 10, 8 (1996), 939-955.

[6] Bruha, I., and Kockova, S. A support for decision making: Cost-sensitive learning system. Artificial Intelligence in Medicine, 6 (1994), 67-82.

[7] Cestnik, B., Kononenko, I., and Bratko, I. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In: Bratko, I., and Lavrac, N. (eds.), Progress in Machine Learning: EWSL-87, Sigma Press (1987).

[8] Chan, P., and Stolfo, S. Experiments on multistrategy learning by meta-learning. Proc. 2nd International Conference Information and Knowledge Management, (1993), 314-323.

[9] Cox, K., Eick, S., and Wills, G. Visual data mining: recognizing telephone calling fraud. Data Mining and Knowledge Discovery, 1, 2 (1997).

[10] Fayyad, U., and Irani, K.B. On the handling of continuous-valued attributes on decision tree generation. Machine Learning, 8 (1992), 86-102.

[11] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. Artificial Intelligence Magazine, (1996), 37-53.

[12] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurasamy, R. Advances in knowledge discovery and data mining. MIT Press, Cambridge, MA, USA, (1996).

[13] Han, J., and Fu, Y. Discovery of multiple-level association rules from large databases. Proc. International Conference Very Large Databases, Zurich, (1995).

[14] Quinlan, J.R. Unknown attribute values in ID3. International Conference ML, Morgan Kaufmann, (1989), 164-168.

[15] Srikant, R., and Agrawal, R. Mining generalized association rules. Proc. International Conference Very Large Databases, Zurich, (1995), 407-419.

[16] Stolorz, P. et al. Fast spatio-temporal data mining of large geophysical datasets. 1st International Conference on Knowledge Discovery and Data Mining, Menlo Park, Calif., (1995), 300-305.

[17] Toivonen, H .et al. Pruning and grouping of discovered association rules. ECML-99, Workshop Statistics, Machine Learning, and Discovery in Databases, Heraklion, Greece (1995).