A DESCRIPTION OF A STRATEGY FOR BUILDING ROUGH SET CLASSIFIERS USING PERFORMANCE FILTERING OF REDUCTS

Staal Vinterbo[1], Lucila Ohno-Machado[2], Hamish Fraser[3]

[1] Research fellow, Knowledge Systems Group, Norwegian University of Science and Technology, 7034 Trondheim, Norway, staal@idi.ntnu.no, 73594480, 73594466 (FAX)

[2] Instructor in Radiology, Decision Systems Group, Harvard Medical School, Boston, USA, machado@dsg.harvard.edu, (617) 732-8543, (617) 739-3672 (FAX)

[3] Research fellow, Clinical Decision Making Group, Massachusetts Institute of Technology, Boston, USA, hamish@medg.lcs.mit.edu, (617) 258-8996

ABSTRACT: The nature of Rough sets theory gives many opportunities for applying heuristics. The prime application area of these is probably reduct calculation. This article will present a practical methodology for constructing rough classifiers from a large set of reducts by using reduct performance filtering. The goal is to create understandable classifiers without sacrificing performance. Results from an application of the methodology to chest pain datasets are presented.

INTRODUCTION

There are several approaches to synthesizing knowledge from low level data. Many of them are based on pattern recognition. The objective of pattern recognition is to find potentially useful patterns in sets of data. The usefulness of these patterns is often judged according to their performance in classification and/or description tasks. These methodologies come from the often non-exclusive fields of machine learning and statistics. Examples are generalized linear regression, neural networks, decision trees and rough sets. This article deals with a rough sets approach.

The rough sets methodology (Pawlak, 1982, 1991; Skowron, 1995) generates propositional descriptors of data originally kept in tabular form. These descriptors reflect functional dependencies within the data and can be used as classification rules. The attractions of the rough sets methodology are a relative simple theoretical foundation, explanatory/descriptive power, data reduction potential and an ability to reflect non-linear dependencies in data. The main drawback is the computational effort required.

The nature of the methodology and the computational intensity present the user with many choices of heuristics to apply. Some of these are essential to the applicability of the methodology such as the reduct computation heuristics, and some have not been formally investigated, and can currently be considered "rules of thumb". A strategy for choosing the best descriptors to include in classifiers is investigated. This strategy involves ranking sets of observations to build descriptors from. This is done using a "tuning" or "holdout" portion of the data. Such tuning sets are also used in e.g. backpropagation neural networks (Rumelhart et al., 1986) to avoid overfitting the model to the data. The explanatory power of rough sets classifiers comes from the propositional rules matching an object to a classification. A large number of matching rules might decrease this power. The construction of small classifiers that do not sacrifice performance is thus a goal. An experiment

using the performance filtering strategy on datasets describing patients presenting in the ER with chest pain is presented.

PRELIMINARIES

Let U be a set of objects. We then define an *attribute* a on U to be a mapping $a : U \to V_a$ from U to an attribute value set V_a. Considering an ordered set A of attributes on U we can map an object x to the corresponding *information vector* $\overline{x}_A$ in the attribute value space $V = \times_{a \in A} V_a$. Given a set $B \subseteq A$ of attributes, we can construct an equivalence relation $\sim_B$ on U as $x \sim_B y \iff \overline{x}_B = \overline{y}_B$. Let us denote the corresponding partition of U as $U_B = \{[x]_B | x \in U\}$. A *decision* attribute d is an attribute that reflects a classification of the objects in the universe. Non decision attributes are often called *conditional* attributes. If $\sim_A \subseteq \sim_{\{d\}}$, then the information system $(U, A \cup \{d\})$ is said to be *consistent*. A non-consistent system $(U, A \cup \{d\})$ is made consistent by defining a *generalized decision* attribute d^* as $d^*(x) = \{d(y) | [y]_d \cap [x]_A \neq \emptyset\}$. The system $(U, A \cup \{d^*\})$ is then consistent. Let $(U, A \cup \{d\})$ be consistent in the remainder of this section. A minimal cardinality set $B \subseteq A$ such that $\sim_B \subseteq \sim_{\{d\}}$ is called a *d-relative reduct* (or, abusing language slightly, just reduct.) Given a reduct B and an object x we can generate a classification rule $\wedge_{a \in B} a(x) \to d(x)$. Considering the universe U one obtains one such rule for each element of the partition U_B. The cardinality of this equivalence class can then be used as an estimate of the *strength* of the rule. Rough classifiers are then built using the set of rules generated from the universe U by a set of reducts.

There is an alternative to this so called "full reduct" approach. Systematically, for each equivalence class $[x]_A$ in $\sim_A$, delete all other equivalence classes that map to the same decision class as the one in question. The object set thus obtained is a subset U_x of U. Reducts are calculated for the system $(U_x, A \cup \{d\})$ and rules are generated using $\{x\}$. These rules are "specialized" to deal with this equivalence class, hence the name "object relative" reduct approach. In general one uses all the rules generated from all the different equivalence classes to form a classifier.

An information vector is classified by matching it to the left hand side of the rules. If several rules match, voting strategies are often applied, e.g. considering each firing rule as a voter with as many votes as it has strength. The strength or voting results allow the classifier to state it's "belief" in an object being of a certain class.

THE EXPERIMENT

The primary dataset of the experiment collected in Sheffield, England, consisted of 500 cases with one decision attribute, one identifier and 43 conditional attributes. Three out of the 43 were non-binary. The decision was 1 or 0, representing the presence and absence of myocardial infarction (MI) respectively. The prevalence of class 1 was 30%.

The data set was randomly split into a training set and a test set with 342 and 159 cases respectively. The means for the MI decision attribute were 0.31 and 0.30 respectively.

The reduct calculations were done by sampling of 10 subtables of sizes 50%, 60%, 70%, 80%, and 90% of the original table size, and calculating reducts for each table.

Attribute value clustering The training set was randomly split into two sets E_0 and T_0 with 113 and 228 objects respectively. Reducts were calculated from T_0 and rules were generated from all of them forming a rule set R_0. Then the non-binary attributes of T_0 were discretized using a simple algorithm. The interval cutoffs were found scanning the sorted attribute values, and adding a cutoff point if the set of objects having this value did not have a classification in common with the sets of objects belonging to the neighboring value forming a discretized set T_d. A rule set R_d was calculated from T_d. As the 0.5 threshold accuracy of R_d on E_0 was higher than R_0 on E_0, discretization was felt to be in order. Discretization cutoffs were calculated using the entire training-set, and the training-set was discretized.

Reduct calculation The discretized training-set was again randomly split into two sets E_1 and T_1 with 113 and 228 objects respectively. 2942 reducts were calculated from T_1. All reducts of length 10 and larger were

removed, leaving 595. For each of these a rule set was created. These rule sets were evaluated on E_1, the "tuning" set, and the "hold-out accuracies" (HA) ((true positives + true negatives)/113) for a 0.5 threshold were calculated.

Classifier construction The union of reducts having a HA above a "hold-out accuracy threshold" (HAT) of 0.83 comprised the full conditional attribute set. Rule Sets for all reducts with HA above 0.83 were calculated using the entire training-set. Incrementing the HAT with 0.01 each time, a classifier using the rule sets of the reducts having HA above this HAT was created and labeled with the HAT.

Classifier evaluation The classifiers built were evaluated on the test set. For 100 different thresholds evenly distributed between 0 and 1, sensitivity (true positives fraction) and specificity (true negatives fraction) was calculated. These values were used to generate receiver operating (ROC) curves that are plots of 1 - specificity vs. sensitivity for each classifier. Also the area below each of the ROC curves (C-index) and the corresponding standard errors (Hanley and McNeil, 1982) were calculated. A summary of the results can be seen in table 1, and ROC curves can be seen in figure 1. The classifiers were also evaluated on a set of 1253 cases collected in Edinburgh, Scotland. The MI mean for this dataset was 0.22. The results were not statistically different (Hanley and McNeil, 1983) from the Sheffield data results (smallest p value was 0.41014.) An argument for the use of ROC curve related indices can be found amongst others in (Swets and Pickett, 1982).

HAT	reducts	rules	attributes	Acc	sens	spec	Cindex / Std.err
0.83	153	16599	43	0.86	0.6	0.96	0.927 / 0.027
0.84	97	10474	42	0.86	0.6	0.96	0.925 / 0.023
0.85	37	3858	39	0.87	0.65	0.96	0.925 / 0.023
0.86	14	1413	33	0.87	0.67	0.95	0.920 / 0.024
0.87	8	837	28	0.86	0.65	0.95	0.922 / 0.023
0.88	4	407	19	0.87	0.71	0.94	0.888 / 0.031
0.89	1	104	8	0.8	0.52	0.92	0.807 / 0.041

Table 1: Rough Classifier summary

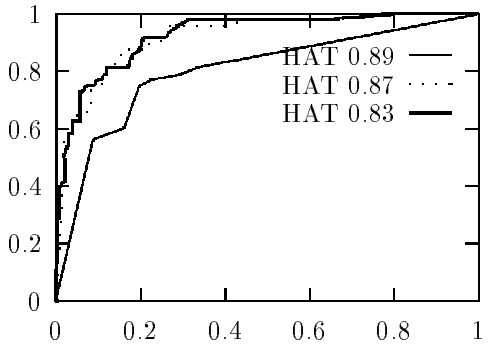

Figure 1: ROC curves from Sheffield data

DISCUSSION

There are several classification methodologies that do quite well on a dataset like the Sheffield dataset used in this paper. They include regression, connectionist, and tree generating approaches. Not all of these offer classifiers with semantic content. Approaches like artificial neural networks have this disadvantage. Rough sets classifiers offer simple, readable propositional rules, but real world classifiers tend to include a large number of these, thus potentially obscuring some of the semantics content. Using the so-called "full" reduct approach (as opposed to the "object related" approach for which a different strategy would have to be employed due to the coupling of reducts to equivalence classes), and assuming no values missing, we can give an upper bound on the number of rules that fire on each incoming information vector. This upper bound is the number of reducts

used in the classifier. On the other hand, we might experience that no rules fire. In that case we suggest a classification according to the apriori probabilities estimated from the training-set. Looking at the HAT 0.83 classifier (table 1) with 16599 rules, the explanation of a classification (barring missing values) would at most involve 153 propositional rules. Preferably the HAT 0.87 classifier should be used, giving at most 8 such rules in an explanation, and doing just as well on our test set. The strategy behind this result might be seen as the main result of the paper.

Another aspect generally of great interest, especially in medicine, is the behavior of classifiers across site boundaries. Several studies have been made using the two datasets featured in this paper, amongst others (Kennedy et al., 1996; Tsien and Fraser, 1997). Both report a loss in performance when crossing site boundaries, although both use the larger Edinburgh set to construct the classifiers, and evaluate on the smaller Sheffield set. The results obtained in the experiment presented seems to suggest that the rough sets classifiers constructed generalize well.

The lowest HAT examined was the one that contained the full conditional attribute set. Models that used higher HATs would contain fewer attributes and could therefore be compared not only in terms of classification accuracy but also in terms of number of attributes. Heuristics were used in the selection of the 0.5 threshold and the size of HAT increments, but our approach indicates that an objective determination of a "minimal" set of rules for a rough set classifier may be attainable.

ACKNOWLEDGMENTS

Staal Vinterbo was funded by project grant 107409/320 from the Norwegian Research Council, and Veritas Research AS. Dr. Ohno-Machado was funded by grant LM/OD06538-01 from the National Library of Medicine and National Institutes of Health. Dr. Fraser was funded by the National Heart Lung and Blood Institute grant R01-HL33041 and Pfizer Inc. We would like to thank Dr. Lee Kennedy for the use of the data, and Aleksander Øhrn for quickly implementing essential features of this strategy on request in the ROSETTA (Øhrn et al., 1997) system. This was the system used to carry out the experiment in this paper.

REFERENCES

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, April 1982.

J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843, September 1983.

R.L. Kennedy, A.M. Burton, H.S. Fraser, L.N. McStay, and R.F. Harrison. Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: Derivation and evaluation of logistic regression models. *European Heart Journal*, 17:1181–1191, Aug 1996.

A. Øhrn, J. Komorowski, A. Skowron, and P. Synak. A software system for rough data analysis. *Bulletin of the International Rough Set Society*, 1(2), 1997.

Zdzislaw Pawlak. Rough sets. *International Journal of Information and Computer Science*, 11, 1982.

Zdzislaw Pawlak. *Rough Sets, Theoretical Aspects of Reasoning about Data*, volume 9 of *Series D: System Theory, Knowledge Engineering and Problem Solving*. Kluwer Academic Publishers, 1991.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learing internal representations by error propagation. *Parallel Distributed Processing*, 1, 1986.

Andrzej Skowron. Synthesis of adaptive decision systems from experimental data. In A. Aamodt and J. Komorowski, editors, *SCAI - 95 Fifth Scandinavian Conference on Artificial Intelligence*, volume 28 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 1995.

J. A. Swets and R. M. Pickett. *Evaluation of Diagnostic Systems. Methods from Signal Detection Theory*. Academic, New York, 1982.

Christine L. Tsien and Hamish S. F. Fraser. Optimizing diagnosis of myocardial infarction in the emergency room: A new flowchart decision aid. American College of Cardiology, Sept 1997. Presented at the American College of Cardiology Dearborn Summit: Reducing Costs and Improving Performance in Cardiovascular Care-Practical Lessons.