RITIO - Rule Induction Two In One

David Urpani†, Xindong Wu‡, Jim Sykes*

† Division of Building, Construction & Engineering, CSIRO
Graham Road, Highett, VIC 3190, Australia

‡ Department of Software Development
Monash University
900 Dandenong Road
Melbourne, VIC 3145, Australia

* School of Information Systems
Swinburne University of Technology
Hawthorn 3122
Melbourne, Australia

Abstract

The identification of relevant attributes is an important and difficult task in data mining applications where induction is used as the primary tool for knowledge extraction. This paper introduces a new rule induction algorithm, RITIO, which eliminates attributes in order of decreasing irrelevancy. The rules produced by RITIO are shown to be largely based on only the most relevant attributes. Experimental results, with and without feature selection preprocessing, confirm that RITIO achieves high levels of predictive accuracy.

Introduction

The pervasive use of sensor and information technology in all aspects of our life has resulted in the generation of vast amounts of digital data. Converting raw sensor data into useful information for human decision makers is one of the driving forces behind research into applications of data mining.

Inductive learning is the primary method for discovering new knowledge from a database by providing concept descriptions which are able to generalize. The two major forms of representation used in inductive learning are the decision tree (DT) and the rule structure. DTs are known to cause fragmentation of the database whenever a high-arity attribute is tested at a node (Pagllo & Haussler 1990) and also have a tendency to repeat subtrees when expressing disjunctive concepts. As a consequence of these two problems DTs tend to grow very large in most realistic problem domains, and they are usually decompiled into a set of rules. Rule-like structures are a more convenient form to express knowledge. Rules are similar to the way human experts express their expertise and human users are comfortable with this way of expressing newly extracted knowledge. Algorithms directly inducing a set of rules are therefore at a distinct advantage as they immediately create the rule set.

Hypothesis generation in induction involves searching through a vast (possibly infinite) space of concept descriptions. Practical systems constrain the search space through the use of bias (Utgoff 1986). One such bias, which has not been given much attention is to minimize the number of features in the concept description. We propose a rule induction algorithm, RITIO (Rule Induction Two In One), which prefers hypothesis spaces containing fewer attributes, generalizing by removing irrelevant attributes. We show that this form of attribute based induction can very efficiently provide syntactically simple concept descriptions with high generalizing powers even in noisy environments. We provide an empirical evaluation of RITIO with C4.5 (Quinlan 1993), C4.5rules and HCV (Wu 1995).

In the next section we describe the RITIO algorithm, and we follow this with the results of an empirical investigation. Two sets of results are presented. The first set is of RITIO without any feature selection (FS) preprocessing and the second set uses FS preprocessing. An algorithm called Classer, reported elsewhere (Urpani & Sykes 1995) is used for this purpose. Classer uses genetic search to identify a subset of features that result in identical or improved classification rates on a nearest neighbour classifier.

RITIO - Rule Induction, Two in One

RITIO carries out a data driven, specific-to-general search for a consistent set of rules which describes the different classes in the data. Like ID3-like algorithms, RITIO makes use of the entropy measure albeit in a different way as a means of constraining the hypothesis search space but unlike ID3-like algorithms the hypotheses language is the rule structure. ID3-like algorithms including ID3 and C4.5 need a decompiler (such as C4.5rules) to transform decision trees into rules, whereby RITIO carries out rule induction without decision tree construction.

Initially the rule set is a copy of the training set representing a set of maximally specific generalizations referred to as the rule matrix (RM). In the initial RM,

which is the rule matrix at level $L = 0$, a rule exists for each training instance. Each level L refers to one stage in the induction process, with higher levels denoting increasing rule generalization. There are a maximum of $N - 1$ levels where N is the number of attributes in the database.

RITIO examines each attribute in the training set at each level L, and selects the least relevant. The heuristic used to identify relevancy is the information theoretic function designed by Shannon (Shannon & Weaver 1949) and popularized in Quinlan's ID3 (Quinlan 1986). In contrast to ID3 the heuristic used in RITIO selects the attribute providing the lowest information gain. The entropy of an attribute A, $E(A)$ is the information required to classify an instance based on the information in that attribute. It is defined as follows:

$$E(A) = \sum_{j=1}^{V} (RelFreq_j \times Inf_j)$$

where V is the total number of distinct values in attribute A, and

$$RelFreq_j = P_j/T$$

with P_j being the number of occurrences of value j in attribute A and T being the number of training instances,

$$Inf_j = -\sum_{k=0}^{C} [P_{j_k}/P_j \times log_2(P_{j_k}/P_j)],$$

with P_{j_k} being the number of occurrences of value j in attribute A belonging to class k and C the number of classes in the training set.

$E(A)$ is calculated for each attribute in the database. In ID3 the attribute with the minimum entropy is selected at a decision tree node to split the tree. RITIO chooses the attribute with the maximum entropy as the candidate for elimination from the RM. This guarantees that the least relevant attribute (according to information theory) is eliminated. The induction process will make a total of $N - 1$ entropy calculations.

On identifying the first least relevant attribute, RITIO checks against each training instance to see whether removal of that attribute results in an inconsistency. An inconsistency is here defined as the occurrence of the same example with different classifications after the removal of the least relevant attribute. An attribute whose removal from a rule causes no inconsistency anywhere in the training set is termed a 'removed' attribute.

After removal of the least relevant attribute from all instances where such removal does not cause an inconsistency, a new, more general RM results. The RM has now been partitioned into two distinct groups: one which still retains the full, initial dimensionality, N, the 'retain' group, and the other with a reduced dimensionality, $N - 1$, the 'remove' group. The RM is now at level 1.

In succeeding rounds of entropy calculations all previous 'least relevant' attributes are not considered. In this case the least relevant attribute from the remaining, $N - 1$ attributes across the two existing partitions is chosen as the next candidate for elimination.

Once this attribute has been chosen the next round in the RM generalization process commences. While checking for consistency the following rules henceforth apply when identifying members of the training set to be used in the checking process:

1. If checking a rule belonging to the 'retain' group all training instances are used. In some cases this means checking also against previously eliminated attributes.

2. If checking a rule belonging to the 'remove' group only training instances belonging to that group are used. As before, the current least relevant attribute is dropped form those rules which do not cause an inconsistency.

The RM at level $L = 2$ has now been partitioned into four groups (ie retain & retain, remove & retain, retain & remove and remove & remove groups). The process repeats itself iteratively $N - 1$ times with new entropy calculations, consistency checks and further partitioning of the RM. At any point in the induction process the RM will contain a maximum of 2^{L-1} partitions up to a final maximum of 2^{N-1} different partitions.

The final RM contains a set of maximally generalized rules guaranteed for consistency. The generalizing process results in a reduction in the number of rules from the original training set size. This manifests itself by repeating rules which are eliminated in the rule extraction process. Another effect of the generalizing process is a reduction in the average dimensionality of the RM as attributes are progressively eliminated. The rules are finally presented as an unordered list in conjunctive normal form.

Real world databases are often noisy, contradictory, incomplete and redundant. To be of any practical use RITIO needs to be able to generalize in the presence of noisy data. RITIO handles noise by a series of processes distributed throughout the induction process. These techniques are discussed in detail elsewhere (Urpani 1996).

Experimental Evaluation

In this section we present the results of an empirical investigation into the performance of RITIO (with and without feature selection preprocessing), using Classer) and compare the results obtained by RITIO with those using C4.5, C4.5rules and HCV.

Throughout the experiments the same default conditions were used for all the databases. Obviously fine tuning different parameters in RITIO would have achieved higher accuracy rates. This however would have been at the expense of a loss in generality and applicability of the conclusions. The default conditions used in RITIO were as follows:

1. The induced rule set was pruned by eliminating those rules which had the same class coverage of less than 5.0%.

2. The maximum number of mismatches allowed during deduction is the number of attributes minus one.

Similarly default conditions were adopted for the three other programs C4.5, C4.5rules and HCV as recommended by the respective authors.

All databases used were divided randomly into training and testing partitions by a 70/30 split of the instances. This was carried out 10 times to obtain 10-fold cross-validation trials.

The Data

The data used in our experiments (see Table 1) can be divided into three groups. The first group is made up of data with 100% nominal attributes. The second group contains data of mixed nominal and continuous attributes. Most of these two groups of data were obtained from (Murphy & Aha 1995), and are available from the HCV software data suite (Wu 1995). The third group of data originates from an aluminium smelter. Further information on all data sets can be found in (Urpani 1996).

Rule Accuracy

Tables 2 shows the accuracy results obtained by the four programs, HCV (Version 2.0), C4.5, RITIO and C4.5rules. The best result for each problem is highlighted with **boldface** font in the table. Results for C4.5 are the pruned ones. The RITIO results are the average of ten fold cross-validated results on unseen test cases. Also included for RITIO is the 95% confidence interval estimate of the mean of the accuracy results. This estimate shows the variance associated with the results and is a good indication of the stability of the algorithm over different databases. For 8 databases out of 10 in the first group in Table 2, RITIO (with FS) obtained the best results. RITIO (without FS) performed just as well relative to the other data sets, however its accuracy was nearly always lower than that from Classer+RITIO.

Out of the 4 databases with continuous data in the second group, RITIO obtained the best results on 2 of them. RITIO produced a particularly good result on the water treatment plant database WTP which is a notoriously difficult real world database, significantly exceeding the next best. FS preprocessing did not seem to particularly improve RITIO's performance on these data sets.

With the industrial databases shown in the last group, RITIO again obtained the best accuracy results for 5 out of the 7 databases. Furthermore in 4 out of 7 cases FS preprocessing obtained a better result than when using RITIO without preprocessing. In several cases such as the 'Temperature' and 'Prediction 3' databases RITIO obtained a very significant improvement over the next best result.

Conclusion

We have presented a new induction algorithm, RITIO, which uses the information theoretic function in a novel way to induce directly a set of rules. It is similar to HCV (Version 2.0) in its approach of using matrices but has stronger noise handling capabilities by eliminating attributes, starting with the least relevant attribute. This is in direct contrast to the DT inducer in C4.5 which uses the most relevant attribute first to branch on. Results also indicate that RITIO's induction accuracy can in many cases be improved through the use of a FS preprocessing procedure.

The algorithm has been shown in the experiments carried out on a wide variety of databases to produce concept descriptions of consistently high accuracy which perform better in most cases than C4.5, C4.5rules or HCV. Future work involves looking at different evaluation functions to employ when selecting attributes for elimination. The consistency check procedure will also be modified to take a 'softer' fuzzy approach (Wu & Måhlén 1995). This ability to tolerate different levels of inconsistency should add to the already good noise tolerance of the RITIO algorithm.

References

Murphy, P.M. & Aha, D.W., UCI Repository of Machine Learning Databases, Machine-Readable Data Repository, Irvine, CA, University of California, Department of Information and Computer Science, 1995.

Paglio, G. & Haussler, D., Boolean feature discovery in empirical learning, *Machine Learning*, **5**(1990): 71-99.

Table 1: Databases Characteristics

Database	# of Instances	Attributes	Classes	Majority Class (%)	Continuous Attributes (%)	Avg # of Values per Attributes	Unknown Values (%)
Hayes-Roth	160	4	3	40.60	0.00	4.00	0.00
Monk1	556	6	2	50.00	0.00	2.80	0.00
Monk2	601	6	2	65.70	0.00	2.80	0.00
Monk3	554	6	2	52.00	0.00	2.80	0.00
Tic-tac-toe	958	9	2	65.30	0.00	3.00	0.00
Soybean	683	35	19	13.50	0.00	2.80	12.00
Vote	435	16	2	61.40	0.00	3.00	5.00
Breast Cancer	286	9	2	70.3	0.00	5.80	0.00
Lymphography	148	18	4	54.7	0.00	3.30	0.00
Primary Tumor	339	17	21	24.8	0.00	2.20	3.00
Aus-Credit	690	15	2	56.00	40.00	4.56	0.65
Lab Neg	56	16	2	65.00	50.00	2.62	35.75
Wine	178	13	3	40.00	100.00	n/a	0.00
WTP	523	38	13	52.18	100.00	n/a	2.95
Pot Noise	542	22	2	50.00	100.00	n/a	0.00
UFT	407	20	2	50.00	100.00	n/a	0.00
Temperature	321	22	3	33.33	100.00	n/a	0.00
Pot Difference	195	22	2	50.00	100.00	n/a	0.00
Prediction 1	721	56	2	50.00	100.00	n/a	0.00
Prediction 2	721	68	2	50.00	100.00	n/a	0.00
Prediction 3	616	68	2	50.00	100.00	n/a	0.00

Table 2: Accuracy (%) Results on Discrete Data

Database	HCV (Version 2.0)	C4.5	RITIO	95% Estimate	C4.5rules	Classer+RITIO
Hayes-Roth	85.70	85.60	87.92	4.14	71.4	**90.4**
Monk1	**100.00**	83.30	97.37	5.96	**100.0**	**100.00**
Monk2	85.20	69.70	**94.97**	3.39	65 3	**94.97**
Monk3	98.10	97.20	99.45	0.61	96.3	**100.00**
Tic-tac-toe	88.00	94.30	98.32	1.11	**100.0**	94.37
Soybean	80.20	82.4	96.80	0.68	80.6	**97.10**
Vote	97.80	97.00	98 97	1.10	93.6	**99.96**
Breast Cancer	72.3	72.8	91.26	7.19	73.5	**92.11**
Lymphography	74.30	69.00	84.90	6.35	72.4	**93.10**
Primary Tumor	38.2	34.00	**75.67**	5.81	33.80	73.10
Aus-Credit	82.50	91.0	**93.22**	2.38	90 0	89.00
Lab Neg	76.50	**88.2**	77.16	5.37	**88.2**	80.5
Wine	90.40	**98.1**	92.37	2.16	**98.1**	92.70
WTP	58.62	60.90	**94.77**	0.83	59.2	94.77
Pot Noise	94 48	**97.80**	96.00	1.47	97.2	94.33
UFT	70.37	69.60	**89.11**	4.46	72.1	84.40
Temperature	48.57	61.90	**82.71**	4.60	67.6	81.96
Pot Difference	89.23	96.90	93.38	3.40	**96.90**	**99.23**
Prediction 1	74.90	68.60	89.58	5.58	75.3	**91.60**
Prediction 2	67.78	57.9	83.71	2.13	57.5	**85.70**
Prediction 3	60.50	53.80	86.63	2.35	54.6	**87.20**

Quinlan, J.R., Induction of decision trees, *Machine Learning*, 1(1986).

Quinlan, J.R., *C4.5: Programs for Machine Learning*, CA: Morgan Kaufmann, 1993.

Shannon, C.E. & Weaver, W., *The Mathematical Theory of Communications*, The University of Illinois Press, Urbana, IL, 1949.

Urpani, D., Knowledge acquisition from real-world data, *PhD Thesis*, School of Information Systems, Swinburne University of Technology, Australia, 1996.

Urpani, D, & Sykes, J., Facilitating knowledge acquisition from industrial process data by automating feature selection, *Proc. of the 8th Intl. Conf. on Industrial and Engg. Applications of Arti. Intelligence and Expert Systems*, Melbourne, Australia, June 6-8, 1995, 161–170.

Utgoff P.E., Shift of Bias for Inductive Concept Learning, *Machine Learning: An AI Approach*, Volume 2, Chapter 5, Morgan Kaufmann Pub., 1986, 107–148.

Wu, X., *Knowledge Acquisition from Databases*, Ablex Publishing Corp., U.S.A., 1995.

Wu, X., and Måhlén, P., Fuzzy interpretation of induction results, *Proc. of the 1995 International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada, August 20-21, 1995, 325–330.