Rule Combination
in
Inductive Learning

Luis Torgo

LIACC
R.Campo Alegre, 823 - 2º.
4100 PORTO
PORTUGAL
Telf. : (+351) 2 600 16 72 - Ext. 115
Fax : (+351) 2 600 3654
e-mail : ltorgo@ciup1.ncc.up.pt

Abstract. This paper describes the work on methods for combining rules obtained by machine learning systems. Three methods for obtaining the classification of examples with those rules are compared. The advantages and disadvantages of each method are discussed and the results obtained on three real world domains are commented. The methods compared are: selection of the *best* rule; PROSPECTOR-like probabilistic approximation for rule combination; and MYCIN-like approximation. Results show significant differences between methods indicating that the problem-solving strategy is important for accuracy oflearning systems.

1 Introduction

Most work in inductive learning tends to discuss the learning method details, but little attention is paid to the problem of how the learned rules are used. This paper shows that different problem solving strategies can lead to very different accuracy results. This clearly indicates the importance of these strategies when comparing performance of learning systems.

Our experiments used an attribute-based learning system to generate theories which were then tested with different problem-solving strategies. This problem is however extensible to other types of learning systems. In general, whenever different sources of knowledge are used (including in multi-strategy learning systems) we need a method for conflict resolution.

Experiments were made on three real world domains. Their goal was to observe if different classification strategies could lead to different results. The following strategies were used : two well known expert systems approaches, MYCIN [11] certainty factors and PROSPECTOR's [5] odds), together with the *best rule* strategy.

The next section describes briefly the inductive system used in the experiments. Section 3 presents the different strategies and section 4 the experiments carried out.

2 The Inductive Engine

In the context of this work the inductive system is used only as generator of rules. The system used for learning those rules was YAILS [12,13]. YAILS system belongs to the attribute-based family of learning programs. It is an incremental rule learning system capable of dealing with numerical attributes and noisy domains. It uses a kind of hill-climbing search strategy with different types of generalisation and specialisation operators. This bi-directional search is guided by an evaluation function which is described in section 3.3.

3 Strategies for Classifying Examples

The goal of inductive learning systems is to generate rules for later use. Application of rules may present some problems, however. We are concerned with the problem of several rules covering an example to be classified. We need a way for deciding which rule is to be followed. This is often referred as the conflict resolution strategy. Alternatively, we may decide to combine different opinions. Several methodologies exist to solve these problems. In the following sections three different strategies are presented. Each strategy attempts to deal with the problem of uncertainty caused, for instance, by unknown attribute values or incomplete description of examples.

3.1 Using Certainty Factors (MYCIN)

MYCIN [11] is one of the best known expert systems. MYCIN uses certainty factors (CF) as a way of modelling reasoning under uncertainty. A certainty factor is a number between -1 and 1 that represents the change in our belief on some hypothesis. A positive number means an increase in the belief and a negative number the contrary. A value of 0 means that there is no change in our belief on the hypothesis. In this work we are particularly interested in the parallel combination of rules, i.e. given $E_1 \Rightarrow H$ and $E_2 \Rightarrow H$ together with their respective confidence factors we are interested on the confidence factor of H given that E_1 and E_2 are true. The formulas used for rule combination in MYCIN are the following :

If CF(H,E1) and CF(H,E2) have opposite signs :

$$CF(H,E_1E_2) = \frac{CF(H,E1)+CF(H,E2)}{1-\min[|CF(H,E1)|,|CF(H,E2)|]} \tag{1}$$

If CF(H,E1) and CF(H,E2) are both greater or equal to zero :

$$CF(H,E_1E_2) = CF(H,E_1)+CF(H,E_2)-CF(H,E_1)\times CF(H,E_2) \tag{2}$$

If CF(H,E1) and CF(H,E2) are both less than zero :

$$CF(H,E_1E_2) = CF(H,E_1)+CF(H,E_2)+CF(H,E_1)\times CF(H,E_2) \tag{3}$$

For the probabilistic definition of CF's we use the following [7] :

$$CF(H,E) = \begin{cases} \dfrac{\lambda(H,E)-1}{\lambda(H,E)} & \text{if } \lambda(H,E) \geq 1 \\ \lambda(H,E)-1 & \text{if } 0 \leq \lambda(H,E) \leq 1 \end{cases} \tag{4}$$

where

$$\lambda(H,E) = \frac{P(E|H)}{P(E\,|\,\overline{H}\,)}$$

This formalisation is derived from a set of axioms [7] which imply that the rules must be conditionally independent given the hypothesis and its negation [9]. This assumption does not hold in general. Nevertheless, this approach has been widely used and achieved good practical results.

3.2 Using Degree of Sufficiency and Necessity (PROSPECTOR)

PROSPECTOR [5] can be considered another successful expert system. In PROSPECTOR the uncertainty associated with a rule is described by two values (LS and LN) which express the degree of sufficiency and necessity with which the conditional part of a rule (E) implies the conclusion (H):

$$LS = \frac{P(E \mid H)}{P(E \mid \overline{H})} \quad \text{and} \quad LN = \frac{P(\overline{E} \mid H)}{P(\overline{E} \mid \overline{H})} \tag{5}$$

If we define the prior and posterior odds on H given E respectively as

$$O(H) = \frac{P(H)}{P(\overline{H})} \quad \text{and} \quad O(H/E) = \frac{P(H \mid E)}{P(\overline{H} \mid E)} \tag{6}$$

we obtain the following definition

$$O(H \mid E) = LS \times O(H) \quad \text{and} \quad O(H \mid \overline{E}) = LN \times O(H) \tag{7}$$

The formula used in PROSPECTOR for rule combination is as follows :

$$O(H \mid E_1^*, \dots, E_n^*) = \prod_{i=1}^{n} L_i^* \times O(H) \tag{8}$$

where

$$L_i^* = \begin{cases} LS_i & \text{if } E_i^* = E_i \\ LN_i & \text{if } E_i^* = \overline{E_i} \\ 1 & \text{if } E_i^* \text{ is unknown} \end{cases}$$

This approach also assumes conditional independence on all E_i 's.

3.3 Using *Best Rule* Strategy

This strategy represents a very simple but efficient way of producing the classification of an example given a set of potentially conflicting rules. It assumes that each rule is characterised by a value which expresses its "quality". When rules are generated by an inductive system this is easily obtained during the learning phase. Here we use a measure of quality provided by YAILS which is a function of two properties: its consistency and completeness. Rule quality is calculated as follows:

$$\text{Quality}(R) = [0.5 + W_{cons}(R)] \times \text{Cons}(R) + [0.5 - W_{cons}(R)] \times \text{Compl}(R) \tag{9}$$

where

$$\mathrm{Cons}(R) = \frac{\#\{\text{correctly covered exs.}\}}{\#\{\text{covered exs.}\}} \qquad \mathrm{Compl}(R) = \frac{\#\{\text{correctly covered exs.}\}}{\#\{\text{exs. of same class as R}\}}$$

$$W_{\mathrm{cons}}(R) = \frac{\mathrm{Cons}(R)}{4}$$

The notion of quality used here is a weighted sum of the consistency and completeness of the rule. The weights are proportional to the value of consistency giving thus some degree of flexibility (see [12, 13] for more details). Our formula for the calculation of quality is a heuristic one. Many other possibilities exist for evaluating a composite effect of various rule properties (see for instance [1] for a function which also includes simplicity).

Let us now come back to the *best rule* strategy. All rules applicable to a given example form a candidate set. After the candidate set has been formed, the rule with the highest quality value is chosen. The conclusion of this rule is followed.

4 Experiments

The experiments performed consisted of comparisons of the classification accuracies obtained by the three approaches described earlier. The same data was used in all experiments. Three medical domain datasets (obtained from Ljubljana) -Lymphography, Breast Cancer and Primary Tumour were used in these comparisons. Each of the datasets was divided in two subsets, one for learning and other for testing (70% for learning and 30% for testing). The three classification strategies were tried using the same learned theory. Table 1 presents the average of ten repetitions of these experiments (standard deviations are between brackets). In order to examine the differences, t-tests with a 95% confidence level were performed. The values which represented a significant difference are in italics on the table.

Table 1. Results of experiments.

	MYCIN-like	*PROSPECTOR-like*	*Best Quality*
Lymphography	78% (5%)	*63% (9%)*	81% (3%)
Breast Cancer	*67% (6%)*	78% (3%)	77% (4%)
Primary Tumour	*23% (4%)*	33% (6%)	32% (7%)

The results of table 1 were quite surprising. The best rule strategy was expected to be the worst since it does not take into account combinations of opinions. MYCIN's certainty factors performed worse than the others, with the exception of Lymphography dataset. PROSPECTOR's approach performed quite badly on Lymphography dataset. Both Breast Cancer and Primary Tumour datasets are known to be rather noisy. In this context the results suggest that the degree of uncertainty of the dataset counteracts in some way the advantages of combination of rules (at least for PROSPECTOR's approach as MYCIN's approach is always bad).

These differences show that the classification strategy can significantly affect the accuracy obtained by learning systems. These experiments seem to indicate that the *best rule* strategy can be a good strategy especially if we take in to account its simplicity when compared to other methods.

5 Relations to other work

Recently several people have studied the effects of multiple sources of knowledge (see [2] for a survey). All approaches share the problem of conflict resolution which is one of the issues tackled by the two probabilistic approaches examined in this paper.

Gams et al., [6] made several experiments with several knowledge bases when classifying new instances. They tried two different strategies to obtain the classification : *best-one* which uses the opinion with highest confidence factor (this is a strategy similar to ours) and the *majority* strategy where confidence factors add up in order to reach a conclusion. This latter strategy represents a kind of combination of different opinions. The authors made extensive experiments on artificial domains and the results showed that the *best-one* strategy scored better whenever few knowledge bases were used. When the number of knowledge bases increased the majority strategy was tbetter. These results seem to suggest that if flexible matching were introduced (which would increase the potential number of opinions) the probabilistic combination strategies examined in this paper might perform better.

Brazdil and Torgo [3] used different learning algorithms to generate several knowledge bases which were combined into one using a kind of best quality strategy. This work suggested that good results could be obtained with this simple strategy, but no comparisons were made with other possible combination strategies.

6 Future Work

The experiments carried out did not admit rules whose conditional parts were not completely satisfied (i.e. flexible matching). It would be interesting to see how accuracy would be affected if flexible matching were used.

The experiments could be extended to other datasets and other learning systems. Experiments with some existing ILP systems are under consideration. This later extension requires not only parallel evidence combination methods (as presented in the paper), but also sequential combination methods to cover the case of rule chaining.

The main extension should be to broaden the range of methods used to combine rules. Some effort could be invested towards the use of a model which does not exhibit the limitations of conditional independence [8] that both certainty factors and degrees of sufficiency and necessity suffer from. Some experiments could be done with Dempster-Shaffer [4] theory of evidence and Pearl's belief networks [10].

7 Conclusions

The experiments carried out in this paper suggest that a simple and quite naive *best rule* strategy performs quite well in comparison with the two other more complex strategies tested. As for PROSPECTOR-like approach, the results on Lymphography were quite bad, and similar to the *best rule* on the other datasets. With respect to MYCIN's certainty factors the performance was quite disappointing altogether.

The results did not show a clear advantage of the two traditional methods which combine different opinions. A possible cause for this could be a small number of rules to combine. This could perhaps improve if flexible matching were used.

The differences in classification accuracy observed between three different combination strategies indicate that more care should be taken when discussing the performance of learning systems. A great deal of work done in the area of approximate reasoning and uncertainty management could be exploited by the ML community.

Acknowledgements

I would like to thank both Pavel Brazdil and Zdenek Kouba for their help and comments.

REFERENCES

1. Bergadano, F., Matwin,S., Michalski,R., Zhang,J. : "Measuring Quality of Concept Descriptions", in *EWSL88 - European Working Session on Learning*, Pitman, 1988.
2. Brazdil,P., Gams,M., Sian,S., Torgo,L., Van de Velde,W. : "Learning in Distributed Systems and Multi-Agent Environments", in *Machine Learning - EWSL91 ,European Working Session on Learning*, Kodratoff,Y. (Ed), Lecture Notes on Artificial Intelligence, Springer Verlag, 1991.
3. Brazdil,P., Torgo,L. : "Knowledge Acquisition via Knowledge Integration", in *Current Trends in Knowledge Acquisition*, IOS Press, 1990.
4. Shafer, G. : *A Mathematical Theory of Evidence*, Priceton University Press, Princeton, 1976.
5. Duda,R., Hart,P., Nilsson,N. : "Subjective Bayesian methods for rule-based inference systems", in Proceedings of the AFIPS National Computer Conference, vol. 47, pp. 1075-1082.
6. Gams,M., Bohanec,M., Cestnik,B. : "A Schema for Using Multiple Knowledge", Josef Stefan Institute, 1991.
7. Heckerman,D. :"Probabilistic interpretation for MYCIN's certainty factors", in *Uncertainty in Artificial Intelligence*, Kanal,L. et al.(eds.), North-Holland, 1986.
8. Kouba,Z. : "Data Analysis and Uncertainty Processing", in *Advanced Topics in Artificial Intelligence*, Marik,V. et al. (eds.), Lectures Notes in Artificial Intelligence, Springer-Verlag, 1992.
9. Mántaras,R., : *Approximate Reasoning Methods*, Ellis Howood Limited, 1990.
10. Pearl, J. : *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, 1988.
11. Shortliffe,E., Buchanan,B., "A Model of Inexact Reasoning in Medicine", in *Mathematical Biosciences*, 23, 1975.
12. Torgo,L. : YAILS an incremental learning program, LIACC-ML Group, Internal report nº 92.1, 1992.
13. Torgo,L. : "Controlled Redundancy in Incremtnal Rule Learning", in this volume.