

LEARNING PRODUCTION RULES FOR CONSULTATION SYSTEMS

Sholom M. Weiss, Casimir A. Kulikowski, and Bernard Nudel

Department of Computer Science
Rutgers University
New Brunswick, N.J. 08903

ABSTRACT

Production rules, with associated measures of probability or confidence, play an important role in contemporary consultation systems. Procedures have been developed which learn a set of probabilistic production rules from a training sample. We have experimented with these methods, by deriving a set of production rules from a medical data base.

Production rules, with associated measures of probability or confidence, play an important role in contemporary consultation systems [1,2]. A recent AI learning system has concentrated on extracting production rules from relatively noise-free data [3], whereas more traditional pattern recognition learning methods have to a large extent dealt with statistical approaches [4]. The major advantage of production rules (and a disadvantage of most statistical methods) is the readily understandable format of the production rule for the expert who is the source of knowledge and the user of consultation systems. A theme of the present work is to show that for some applications, probabilistic productions rules can be learned from a training sample. Moreover, the form of these production rules is similar to those used in a current large-scale consultation system [8]. Some of the procedures employed appear to be analogous to those an expert would use to generate rules from experimental data. A similar goal has motivated application of variable valued logic methods [9] to rule induction in pancreatic cancer.

The problem can be stated as: given a set of h binary features and samples for class C and its logical complement C' , find a relatively small set of exhaustive and mutually exclusive production rules (of the

type found in decision tables) and their associated probabilities. For simplicity, and to reduce dimensionality, binary features, and a single class and its complement are considered. The task of reducing the features to this form and feature extraction, has been considered separately [5].

A somewhat analogous problem has been previously investigated in the context of learning decision trees [6,7]. Our work has some similarities to version spaces [3], but we consider a set of production rules (not a single production rule) and the data may be very noisy, i.e. the same pattern often appears in both C and its complement. In our testing, the sample data was processed so that it appeared in the following form: For six features, there are 64 possible patterns, of which 39 occurred.

	PATTERN						PROB	FREQ
1)	1	2	3	4	5	6	1.0000	4
2)	1	2	3	4	5	-6	.5000	4
3)	1	2	3	4	-5	6	.8222	45
4)	1	2	3	4	-5	-6	.7104	38
5)	1	2	3	-4	5	6	.3333	3
6)	1	2	3	-4	5	-6	.0000	10
7)	1	2	3	-4	-5	6	.6874	32
8)	1	2	3	-4	-5	-6	.4374	16
9)	1	2	-3	4	5	6	.0000	0
.								
.								
.								
64)	-1	-2	-3	-4	-5	-6	.0000	0

Table 1.

Each pattern has an associated PROB, the Probability of class C given the Pattern, and FREQ, the total frequency of the pattern (among the samples).

The solution consists of finding a smaller set of production rules (with don't care conditions on the variables in addition to yes/no) such that: the set of production rules is much smaller than the number of patterns, the frequency of occurrence for

each production rule is $>k$, and the classification accuracy is as close as possible to that in the original data base. This is desirable because : (a) the expert can only understand and accept a limited number of rules (b) the frequency of occurrence of any rule must be large enough to produce relatively accurate estimates of probability.

What follows is an example of a set of 7 production rules which was derived for the above 64 possible patterns. The data consists of historical information taken from 341 patients with bronchial asthma. The derived classification rate for the 341 patterns, which was used to aid in predicting which patients would benefit from therapy, was competitive with other statistical methods, such as the Bayesian and the nearest neighbor methods.

	PATTERN				PROB	FREQ
1)	1	2	3	* *	0.769	91.
2)	1	*	3	-4 * 6	0.606	66.
3)	1	*	3	-4 * -6	0.315	54.
4)	*	*	-3	* * *	0.206	34.
5)	1	-2	3	4 * 6	0.645	31.
6)	1	-2	3	4 * -6	0.375	32.
7)	-1	*	3	* * *	0.121	33.

OPTIMAL CLASSIFICATION RATE: 72.727 %

CURRENT CLASSIFICATION RATE: 71.261 %

Table 2. (* is a don't care)

In the order of increasing computation time, three procedures are described for deriving a set of production rules.

a) The features are ordered from highest to lowest in significance (this can be done in conjunction with the expert). A closeness threshold is given, for which rules with probabilities within the threshold can be combined (when logically compatible). The program proceeds to combine rules, beginning with the least significant feature (i.e. those patterns which differ only in the rightmost column). The procedure's criteria for combining patterns are similar to two important reasoning criteria used by experts: sensitivity and specificity. In this case we require a minimum frequency and a degree of closeness of probability for any pair of rules to be combined. However, the essential difference with the usual procedures for combining rules and replacing constants with variables, is that these rules have probability measures which must be taken into account. These measures change our interpretation of which rules may be combined. This is a relatively simple procedure that gives quick and often good

results. For example, assume that the thresholds were set so that 2 rules could be combined either when they have probability measures within .1 of each other or when the frequency of one of them is less than .10. In Table 1, patterns 1 and 2 could be combined because each has a small frequency. Patterns 3 and 4 could not be combined because both frequencies are above the threshold and the probability estimates differ by more than .1.

b) Without ordering the features as in (a), all possible orders ($n!$, where n is the number of features) are considered. This has in many instances yielded near-optimal results. Table 2 was generated using this procedure.

c) Using a search and branch and bound technique, an optimal solution can be found for a set of n production rules. AI search techniques have recently been given more attention in pattern recognition theory [7].

These procedures can result in a set of production rules which are understandable to the expert, include dependencies, and are competitive in performance with other statistical procedures. The requirement that the number of features included in a rule be no larger than eight is consistent with the maximum size that an expert is likely to provide. The number of the samples must increase if rules containing more findings are required. Rules of this form are compatible and can be incorporated in models developed with a new consultation system [8]. The examples presented correspond to a form which was suggested by a physician who felt that it would be readily understandable by other physicians. We are currently working on another application of much greater complexity in developing a consultation system in rheumatology. While this learning approach is likely to work in only small applications, it should prove interesting to see whether limited sets of rules could be learned and used in relatively de-coupled subsections of the consultation system.

ACKNOWLEDGEMENTS

This research was supported in part by grant RR-643 of the NIH Biotechnology Resources Program.

REFERENCES

1. Weiss, S., Kulikowski, C., Amarel, S., Safir, A., A Model-based Method for Computer-aided Medical Decision-making, Artificial Intelligence, Vol. 11, pp. 145-172 (1978).
2. Shortliffe, E., Computer-based Medical Consultation: MYCIN, Elsevier, N.Y. (1976).

3. Mitchell, T., Version Spaces: A Candidate elimination Approach to Rule Learning, Proceedings 5th Joint Conference on Artificial Intelligence, pp. 305-310 (1977).
4. Shimura, M., Learning Procedures in Pattern Classifiers-Introduction and Survey, Proceedings 4th International Joint Conference On Pattern Recognition, pp. 125-138 (1978).
5. Weiss, S., Kern K., Kulikowski, C., Pincus, W., An Interactive System for the Design of Classifiers in Diagnostic Applications, Proceedings 1978 International Conference on Cybernetics and Society, pp.58-62.
6. Gleser, M., Collen, M., Towards Automated Medical Decisions, Computers and Biomedical Research, Vol. 5, pp. 180-189 (1972). Vol. 11, pp. 145-172 (1978).
7. Kulkarni, A. Kanal, L., Admissible Search Strategies for Parametric and Nonparametric Hierarchical classifiers, Proceedings 4th International Joint Conference on Pattern Recognition, pp. 238-245 (1978).
8. Weiss, S., Kulikowski, C., EXPERT: A System for Developing Consultation Models, Proceedings IJCAI-1979.
9. Michalski, R., AQVAL/1--Computer Implementation of a Variable-valued Logic System VL1 and Examples of its Application to Pattern Recognition, Proceedings 1st International Joint Conference on Pattern Recognition, pp. 3-17 (1973).