ITRULE : AN Information Theoretic Rule-Induction Algorithm

Rodney M.F.Goodman and Padhraic Smyth
Department of Electrical Engineering
California Institute of Technology, 116-81
Pasadena, CA 91125

In this paper we propose a general information-theoretic model for rule-based systems and describe an algorithm which can induce an optimal set of concepts or rules from a set of examples. This paper is an extension of a companion paper [1] which analysed the information theory model in the more restricted problem domain of deriving classification decision trees from a set of examples. We show in this paper that previous work in this area has concentrated on learning single concepts from examples and is classification-oriented. Motivated by the need for an induction method which can derive a more flexible set of rules for expert systems we define a non-parametric information-theoretic methodology. This information-theoretic representation is shown to be compact, mathematically consistent and very powerful. The main advantage of our approach over previous algorithms is in the flexibility of representation, allowing the algorithm to learn many different concepts.

The generalised rule induction problem

The ability to extract general information from a set of examples is a fundamental characteristic of knowledge acquisition. The term *example* is taken to mean a feature vector of attribute values where the values are assumed to be either categorical or discrete numerical. Usually in this notation each feature vector has another component called a classification which contains the class label for that example. However for our purposes we treat the classification as just another attribute, albeit an important one. The *set* of specific examples can be thought of as a large table of these feature vectors, where 'large' means statistically sufficient for whatever computations may be required. The *general information* we wish to extract from these examples is taken to be in the form of production rules, i.e.

if *condition A* then *condition B* with probability p

where we allow the antecedent A to be the conjuction of several simple conditions and the consequent B is restricted to being a single simple condition. A simple condition is defined as 'attribute X takes on value y'. The rationale for choosing this representation is based on both theoretical and practical considerations. Theoretically it has been postulated that such rule representations are a good model for knowledge representation [2] while from a practical viewpoint such rules can be used as the knowledge-base for an expert system.

In this paper we consider the problem of obtaining the best set of rules given a set of specific examples. What is new is the fact that we do not restrict the consequent clause in our rules to be the classification attribute, i.e. we are interested in attribute-attribute rules as well as attribute-class rules. Previous work in rule-induction has concentrated on classification-type rules where the consequent clause only involves the class variable. However in data-driven applications a more flexible and general approach is desirable. One wishes to have rules relating attributes. In this sense, by treating the class as 'just another attribute', the *generalised* rule-induction problem is formulated. The more standard classification approach is seen to be a special case of the general problem.

Consider an example from the animal domain (mammals actually). Say we have a large list of animals (our example set), each animal being described in terms of various attributes such as size (large or small), colour, number of legs, dangerous (yes or no), species, etc. A classification-type of induction algorithm might induce rules such as

If *the animal has no legs* then *it is a snake* with probability p

A more general induction algorithm would induce rules such as

If *the animal is large* then *it has 4 legs* with probability p

in addition to the classification rules. The advantages of the more general approach are obvious in terms of providing a more accurate representation of the important concepts about the domain from which the examples are taken.

For the purposes of this paper we limit our attention to selective induction (where the induction algorithm retains the same representation as given by the examples) as opposed to constructive induction.

Earlier work in the area of rule induction

A good taxonomy of automatic induction algorithms is given in Cohen and Feigenbaum [3]. These algorithms can loosely be categorised into two main areas ; those which use symbolic manipulation techniques and those which use statistically-oriented techniques. A good example of the former would be the AQ11 algorithm of Michalski and Larson [4] which achieved success in the domain of plant disease diagnosis [5]. Typically these algorithms examine the examples *sequentially* and refine what is known as the rule space until a set of general rules covering the examples are arrived at. However noisy examples (as in the case where not all but nearly all large animals in the domain have 4 legs) are not easily handled by the symbolic approach. In addition the algorithms are computationally unattractive. Consequently their use has been limited to research-oriented endeavours rather than practical applications such as knowledge acquisition for expert systems.

Methods which can termed as statistical, exploit *average* properties of the example set. Perhaps the best known of such techniques is Quinlan's ID3 algorithm [6] based on information theory. The ID3 algorithm derives an efficient classification tree for the example

set. This algorithm has been used in several practical applications [7,8]. Recent work has shown that the basic strategy of the algorithm is theoretically well-founded [9]. An interesting experiment which we carried out was to use the information measure to derive hierarchical edge detectors [1]. Because we formulated the problem in a classification framework, the algorithm in essence was able to learn basic edge detection capabilities. In addition the experiment yielded new domain-specific information as to which attribute was the single most important edge discriminator.

The information-theoretic approach

In this paper we intend to view the generalised induction problem from a statistical viewpoint, using information theory techniques in particular. To begin with we are given a sample set of data, in the domain or universe of interest (e.g. mammals), i.e. a list of examples from this domain described in terms of attributes. We assume that the size of the sample set is sufficiently large to allow us to estimate statistical parameters to some required degree of accuracy - we need not be overly concerned with the sampling theory aspects of the problem.

The attributes ('class' included) can be viewed as random variables and, as assumed earlier, are discrete-valued. The generalised rule induction problem can then be reformulated from this random variable viewpoint - in some sense we want to 'capture' all the important interdependencies among the set of random variables. So the best set of rules should (somehow) be the most important dependency relations between the random variables - but how can we quantify this ? If p is near, or equal to, 1, does that mean that the rule is good ? Not if one considers the example

If *the animal is a kangaroo* then *it does not drink beer* with probability = 1

Although $p = 1$ this cannot be viewed as a good rule, at least in most parts of the world. It is reliable but it is simply not useful. On the other hand a rule such as

If *the animal is red* then *it is a bird* with probability = 0.6

is quite unreliable but in fact it could be quite useful. Hence whatever measure we choose to represent the 'goodness' of a rule must at least be a function of the usefulness and the reliability of the rule, although we have not specified how to measure either parameter.

Consider the average mutual information measure between 2 random variables A and B, I(A;B), as defined by Shannon [10]. This measure has certain interesting properties. The definition can be given as

$$I(A;B) = H(A) - H(A|B)$$

i.e. the average information that A provides about B is our original average uncertainty about A ($H(A)$) less our remaining average uncertainty about A given B ($H(A|B)$). Another way to put it is that it is our average *decrease* in uncertainty about A once we know

B. $H(A)$ is the entropy function, a measure of how uncertain we are about the value of A.* The information measure was originally defined by Shannon in 1948 and has found almost exclusive (but extremely profitable) use in the area of communication theory. Nonetheless this measure has more general implications than just measuring the information sent across a channel.

We propose that the information measure is a natural and intuitive measure for the 'goodness' of a rule. Further we argue that it is powerful in terms of induction properties while being consistent and mathematically rigorous from a statistical viewpoint. As we stated previously, rules can be viewed as defining the relative dependency of two random variables. The information measure $I(A;B)$ is a measure of this dependency. Independent variables have zero mutual information ($I(A;B) = 0$), while completely dependent variables have the maximum amount of information possible, i.e $I(A;B) = H(A) = H(B)$. But as we saw in the kangaroo example, dependent variables do not necessarily yield good rules. Consider the information contained in this rule. The information is the decrease in our uncertainty about the variable 'beer-drinker' (yes/no), in the animal domain. But $p_{yes} = 0$ and $p_{no} = 1$ so that our initial uncertainty is zero, i.e. we are certain that animals do not drink beer. Hence knowing whether the animal is a kangaroo or not provides us with zero information. The rule is useless according to our measure - just the result we desired. If we look at the basic definition of $I(A;B)$ above we can identify in a general sense that the $H(A)$ term can be a measure of the usefulness of a rule (as it pertains to A), while the equivocation or noise term ($H(A|B)$) is a measure of the unreliability of the rule.

One of the fundamental criteria for any worthwhile induction mechanism is that it must have the property of *generalisation* (see Michalski [13]). Consider for example the rules

If *the animal is brown and it can fly* then *it has wings*

and

If *the animal can fly* then *it has wings*

Clearly the second rule is preferable to the first since it is more general. Can one measure this preference quantitatively ? The information measure has the property of additivity, i.e.

$$I(A;B,C) = I(A;B) + I(A;C|B)$$

where $I(A;B,C)$ is the average information provided about A by both B and C together, and $I(A;C|B)$ is the average information provided about A by C, given that we know B already. This latter quantity we can call the relative information of C to A given B. If this is zero, then the more specific rule contains no information *relative* to the more general rule. Hence the information provided by the variable 'is-brown' about 'has-wings' is zero, given that we already know that 'can-fly' is true.

---

* $H(A) = \sum_{i=1}^{n} p_i . \log \frac{1}{p_i}$ where A can take on n values and p_i is the probability of the ith component.

antlers	wings	dangerous	legs	size	dog	reindeer	snake	bird	probability
yes	no	no	yes	large	no	yes	no	no	0.10
yes	no	yes	yes	large	no	yes	no	no	0.01
yes	no	no	yes	small	no	yes	no	no	0.03
no	no	yes	yes	large	yes	no	no	no	0.16
no	no	yes	yes	small	yes	no	no	no	0.03
no	no	no	yes	large	yes	no	no	no	0.09
no	no	no	yes	small	yes	no	no	no	0.08
no	no	yes	no	large	no	no	yes	no	0.18
no	no	yes	no	small	no	no	yes	no	0.07
no	no	no	no	large	no	no	yes	no	0.01
no	no	no	no	small	no	no	yes	no	0.02
no	yes	no	yes	large	no	no	no	yes	0.01
no	yes	yes	yes	large	no	no	no	yes	0.06
no	yes	no	yes	small	no	no	no	yes	0.15

Figure 1

In summary, the information-theory model provides a consistent measure of the 'goodness' of a rule and we demonstrated that it is clearly an appropriate model for the induction problem. Unlike many computational methods based on statistics the model is non-parametric, requiring no restrictive statistical-model assumptions.

The rule induction algorithm

Having demonstrated the utility of the information theory approach in a general sense we present initial results from a new algorithm called the ITRULE algorithm (Information-Theoretic Rule induction algorithm). ITRULE induces the best set of rules from from an example set where best is defined in an information-theoretic sense.

The earlier work of Quinlan [6], ourselves [1,9] and others [11] was based on algorithms which when given a set of training examples ultimately produce a decision tree for classifying the examples. The disadvantage of this approach lies in the restrictive format of the tree structure. While it is true to say (as pointed out by Bundy et al. [12]) that the tree represents a form of rule structure, this structure may be too limited for applications such

ITRULE.1 output : the most informative set of 20 rules

rule # 1 If the animal does not have legs then it is a snake with probability 1.00
rule # 2 If the animal is a snake then it does not have legs with probability 1.00
rule # 3 If the animal has wings then it is a bird with probability 1.00
rule # 4 If the animal is a bird then it has wings with probability 1.00
rule # 5 If the animal has antlers then it is a reindeer with probability 1.00
rule # 6 If the animal is a reindeer then it has antlers with probability 1.00
rule # 7 If the animal is dangerous and it is not a dog then it does not have legs with probability 0.93
rule # 8 If the animal is dangerous and it is not a dog then it is a snake with probability 0.93
rule # 9 If the animal has legs then it is not a snake with probability 1.00
rule #10 If the animal is not a snake then it has legs with probability 1.00
rule #11 If the animal does not have wings then it is not a bird with probability 1.00
rule #12 If the animal is not a bird then it does not have wings with probability 1.00
rule #13 If the animal does not have legs and it is large then it is dangerous with probability 0.96
rule #14 If the animal is large and it is a snake then it is dangerous with probability 0.96
rule #15 If the animal has legs and it is not a dog then it is not dangerous with probability 0.94
rule #16 If the animal is not a dog and it is not a snake then it is not dangerous with probability 0.94
rule #17 If the animal is dangerous and it has legs then it is a dog with probability 0.90
rule #18 If the animal is dangerous and it is not a snake then it is a dog with probability 0.90
rule #19 If the animal does not have wings and it is not a dog then it does not have legs with probability 0.67
rule #20 If the animal is not a dog and it is not a bird then it does not have legs with probability 0.67

Figure 2

as deriving rule bases for expert systems where a more flexible data-driven representation is generally required.

The ITRULE algorithm derives a set of rules rather than a decision tree. A special case of ITRULE, where it focuses on the class attribute alone (in terms of consequent clauses) would be equivalent to the decision tree algorithms.

ITRULE.1 is an initial implementation of the general ITRULE algorithm. The approach taken was to simulate an example set using the prototype examples of figure 1. The domain consists of the set of animals [dogs, reindeer, snakes, birds] described in terms of the attributes *antlers, wings, dangerous, legs, size.* The number of examples generated for each prototype is proportional to its probability. These probabilities were chosen rather arbitrarily, e.g. large snakes are much more likely to be dangerous than not dangerous.

The results of the algorithm on this example set are shown in figure 2. The induced set of rules represent the twenty most informative rules, listed in order from most informative downwards. The are seen to correspond well with what rules a human might consider important, capturing all the important concepts such as 'birds have wings,' 'snakes don't have legs,' etc. The results on this simple problem are quite encouraging.

Conclusion

Having defined the problem of generalised rule induction, we proposed Shannon's information measure as a natural and coherent model from which to build a solution. We showed that the measure has desirable properties both as a measure of the goodness of a rule and as a basic induction mechanism. Initial results with the ITRULE.1 algorithm show that the measure yields rules which are intuitive and common-sense to humans. Current and future work is to build on the information theory principle and develop much more powerful induction algorithms, addressing such problems as dealing with example sets which, in a statistical sense, are not large.

References

[1] R.M.F.Goodman and P.Smyth, 'Learning from examples using information theory', accepted for presentation at the 2nd AAAI Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, October 1987.

[2] A.Newell and H.A.Simon, *Human Problem Solving*, Englewood Cliffs, N.J : Prentice Hall, 1972.

[3] P.R.Cohen and E.A.Feigenbaum, *The Handbook of Artificial Intelligence, Volume 3* William Kaufmann Inc., 1982.

[4] R.S.Michalski and J.B.Larson, 'Selection of most representative training examples and incremental generation of VL1 hypotheses', *Report No. 867*, Computer Science Department, University of Illinois, 1978.

[5] R.S.Michalski and R.L.Chilausky, 'Learning by being told and learning from examples', *International Journal of Policy Analysis and Information Systems* 4, 125-161, 1980.

[6] J.R.Quinlan, 'Learning efficient classification procedures and their application to chess endgames', *Machine learning : an artificial intelligence approach*, R.S.Michalski, J.G.Carbonell and T.M.Mitchell (editors), Palo Alto, CA. : Tioga, 1983.

[7] D.Michie, S.Muggleton, C.Riese and S.Zubrick, 'RuleMaster - a second generation knowledge engineering facility', *Proceedings of the First Conference on Artificial Intelligence Applications*, Denver CO, December 1984.

[8] J.D.Stuart, S.D.Pardue, L.S.Carr, D.A.Feldcamp, 'TITAN : an expert system to assist in troubleshooting the Texas Instruments 990 minicomputer system', *Radian Technical Report ST-RS-00974*, Radian Corporation, Austin TX, December 1984.

[9] R.M.F.Goodman and P.Smyth, 'An information theoretic approach to decision tree design', presented at the 1986 IEEE Int. Symp. Inform. Theory, Ann Arbor, Michigan, October 1986.

[10] C.E.Shannon, 'A Mathematical Theory of Communication', *Bell Syst. Tech. Journal*, vol. 27, (pt 1) pp. 379-423, (pt 2) pp. 623-656.

[11] L.Breiman, J.H.Friedman, R.A.Olshen and C.J.Stone, *Classification and regression trees*, Belmont, CA. : Wadsworth, 1984.

[12] A.Bundy, B.Silver and D.Plummer, 'An analytical comparison of some rule-learning programs', *Artificial Intelligence* ,27, 1985, pp.137-181.

[13] R.S.Michalski, 'Pattern recognition as rule-guided inference',*IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-2, pp 349-361, 1980.