Information-Theoretic Rule Induction

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

1 Background and motivation

The problem of induction or "learning from examples" can roughly be divided into two distinct categories, namely the symbolic manipulation approach and the statistically-oriented approach. Within each category exist a variety of theories and techniques for inductive inference. The better known symbolic techniques include Mitchell's version spaces algorithm [1] and the AQ11 algorithm of Michalski [2]. Statistical techniques for induction have primarily evolved from classical pattern recognition theory and embody such principles as non-parametric statistical estimation (e.g., the CART algorithm by Breiman et al. [3]) and information-theoretic ideas (e.g., the work of Quinlan [4] and Goodman and Smyth [5]).

Induction can be viewed as a search for hypotheses (restricted to some *hypothesis space*) to account for a set of instances or examples which are often assumed to be restricted to some *instance space*. The general learning problem consists of being given positive and negative instances of some *concept* and trying to find a hypothesis in the hypothesis space which "best" describes this concept. Let v be any positive instance in the instance space for some concept. The fundamental difference between the two inductive approaches lies in the fact that symbolic algorithms try to find a *deterministic* mapping, or a Boolean function F, from the instance space to the hypothesis space, to describe the concept, i.e., we seek an F such that $F(v) = 1$ for all v. The statistical approach, however, tries to find a *probabilistic* mapping, or a probability distribution, between the two spaces, i.e. $prob(F(v) = 1) \geq 1 - \delta$, where δ is some inherent function of the given hypothesis space, $0 \leq \delta < 1$.

Statistical techniques cannot easily deal with *incremental* learning (e.g., in decision tree design the entire tree algorithm must be re-run) while symbolic algorithms often incorporate incremental learning as a basic mechanism. On the other hand, symbolic techniques cannot handle noise in the instance data very well (due to the implicit assumption of a deterministic mapping) while the statistical approach inherently takes account of such noise. It is worth noting at this point, that in the light of the above of the remarks we can interpret the recent learning framework introduced by Valiant [6] in the following manner: Valiant's work extends the symbolic approach to the extent that the learning of the function F is modelled probabilistically. However this is fundamentally different to the statistical approach which learns a function which is itself probabilistic and so, unfortunately, the results obtained using the Valiant framework are not directly applicable to statistical algorithms.

Production rule systems are a good example of symbolic algorithms which are based on cognitive science and yet which might benefit from statistical techniques. While such systems have seen widespread application in recent times, there remain many fundamental limitations, such as the knowledge-acquisition bottleneck in obtaining the rules and the many problems which occur when trying to control the behaviour of large rule sets. We believe that the lack of a quantitative well-defined "rule-preference measure" is the root cause of many of these problems. Such preference measures are required both to rank hypotheses during induction (cf. Michalski [7]) and to resolve conflicts during rule-based inference control. Rule-preference measures based on symbolic techniques alone are non-robust. Hence, we recently proposed the *J-measure* [8] as an information-theoretic alternative to existing approaches. The J-measure quantifies the information content of a rule or a hypothesis. In this paper we will focus on the properties of the J-measure as it relates to induction from a cognitive science viewpoint, i.e., we will investigate how the mathematics

* This work was supported in part by Pacific Bell and Caltech's Program in Advanced Technologies, sponsored by Aerojet General, General Motors and TRW

supports the theoretical inductive mechanisms of generalisation and specialisation (information-theoretic aspects of the measure are treated in [8]). Following the theoretical discussion, we define the ITRULE algorithm which uses the newly-proposed measure to learn an optimal set of rules from a set of instances and we conclude the paper with an analysis of experimental results.

2 The information content of a rule

We propose to use the following simple model of a rule, i.e.,

$$\text{If } \mathbf{Y} = y \text{ then } \mathbf{X} = x \text{ with probability p}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are two attributes (dimensions in the instance space) with "x" and "y" being values in their respective discrete alphabets. For our purposes we may treat $\mathbf{X}$ and $\mathbf{Y}$ as discrete random variables. We restrict the right-hand expression to being a single value assignment expression while the left-hand side may be a conjunction of such expressions. Intuitively we can view the two random variables as being connected by a discrete memoryless channel. The channel transition probabilities are simply the conditional probabilities between the two variables. A rule is equivalent to the occurence of a particular channel input event.

As we have shown in detail elsewhere [8], the j-measure is a particular formula which we defined for calculating the information we receive about the variable $\mathbf{X}$ given the event $\mathbf{Y} = y$, or $\mathrm{I}(\mathbf{X};\mathbf{Y}=y)$. We define the j-measure as

$$j\,(\mathbf{X};\,\mathbf{Y}=y) \;=\; \sum_{x} p(x|y).\log\left(\frac{p(x|y)}{p(x)}\right) \qquad (1)$$

and henceforth we refer to $\mathrm{j}(\mathbf{X};\mathbf{Y}=y)$ as the *instantaneous information* while the *average information content* is defined as

$$J\,(\mathbf{X};\mathbf{Y}=y) \;=\; p(y)\;.\;j\,(\mathbf{X};\mathbf{Y}=y) \qquad (2)$$

Note that this measure is an average in the sense there is an implicit assumption that the instantaneous information from the other "Y-terms" is zero, which is consistent with the cognitive science approach to production rules where essentially we can only draw inferences about certain events and not others. The concept of *average* information is important for induction and so $J(\mathbf{X};\mathbf{Y} = y)$ (henceforth to be referred to as the J-measure) is the measure we use to rank hypotheses for induction (as opposed to the instantaneous measure). In the next section we will demonstrate the appropriateness of this choice. In an intuitive sense, the average measure relates to the average value of the rule information content, while the instantaneous measure can be used to rank rules after the event $\mathbf{Y}= y$ has occured.

3 Ranking hypotheses using the J-measure

We examine the nature of the J-measure as a basic preference measure among competing hypotheses. There appears to be a general consensus that the two primary criteria for evaluating a hypothesis are *simplicity* of the hypothesis and *goodness-of-fit* (Angluin and Smith [9], Gaines [10] and Michalski [7]). The problem is to combine these two criteria into a single measure so that the hypotheses can be ordered. Let us interpret the event $\mathbf{X} = x$ as the concept $F(v) = 1$ to be learned and the event (possibly conjunctive) $\mathbf{Y} = y$ as the hypothesis describing this concept.

The J-measure is the product of two terms. The first, $p(\mathbf{Y} = y)$, is the probability that the hypothesis will occur and, as such, can be interpreted as a measure of simplicity. (Angluin and Smith [9] also mention this idea, the more probable a hypothesis the simpler it should be). The second term is $j(\mathbf{X};\mathbf{Y} = y)$. This quantity as defined in equation (1) is equal to the cross-entropy of $\mathbf{X}$ with the variable '$\mathbf{X}$ conditioned on the event $\mathbf{Y} = y$'. Cross-entropy is well known as a goodness of fit measure between two distributions (cf. Shore and Johnson [11]). It can be interpreted as a distance measure where 'distance' corresponds to the amount of information required to specify a random variable. It is frequently used to find the conditional distribution which most closely agrees with the original distribution. The cross-entropy is zero if and only if the two distributions are exactly equal. For our purposes we must be very careful to interpret what we mean by goodness-of-fit. In the probabilistic manner for which the problem is defined we have an *a priori* value for $p(\mathbf{X} = x) = p(F(v) = 1)$. This represents our best probability estimate as to whether an arbitrary instance v is contained in the concept or not, without any hypothesis. By introducing a hypothesis we now have an *a posteriori* value, $p(F(v) = 1|\mathbf{Y} = y)$. Without loss of generality we can assume that

$$p(F(v) = 1|\mathbf{Y} = y) \;\geq\; p(F(v) = 1) \qquad (3)$$

since otherwise we can define $F(v) = 0$ as the hypothesis of interest. The measure attains its maximum value (for a given F and $\mathbf{Y}$) if and only if

$$p(F(v) = 1|\mathbf{Y} = y) \;=\; 1, \qquad (4)$$

whereas the measure is minimised if and only if

$$p(F(v) = 1|\mathbf{Y} = y) \;=\; p(F(v) = 1), \qquad (5)$$

i.e., the hypothesis is no better than a random guess based on *a priori* probabilities. Intermediate *a posteriori* values, $1 < p(F(v) = 1|\mathbf{Y} = y) < p(F(v) = 1)$, provide a monotonic measure of the information-theoretic distance between the uncertainty of the random guess without the hypothesis, and complete specification. Given two hypotheses, $\mathbf{Y} = y$ and $\mathbf{Z} = z$, and assuming without loss of generality that $p(\mathbf{X} = x|\mathbf{Y} = y) > 0.5$, then it can easily be shown that

$$j(\mathbf{X}; \mathbf{Y} = y) \;>\; j(\mathbf{X}; \mathbf{Z} = z) \qquad (6)$$

if and only if

$$p(\mathbf{X} = x|\mathbf{Y} = y) \;>\; p(\mathbf{X} = x|\mathbf{Z} = z). \qquad (7)$$

In this sense $j(\mathbf{X}; \mathbf{Y} = y)$ clearly corresponds to a goodness of fit measure. Hence we can conclude that the J-measure possesses appropriate properties for ordering hypotheses as it trades-off a simplicity component, $p(\mathbf{Y} = y)$, with a goodness-of-fit component, $j(\mathbf{X}; \mathbf{Y} = y)$.

4 Generalisation and specialisation using the J-measure

Next we investigate how our J-measure can be used to generate new rules. Holland et al. [12] define the basic operations for the induction of new rules as *condition-simplifying* generalisation, *instance-based* generalisation and specialisation.

Condition-simplifying generalisation:
The basic principle at work here is that rules which have irrelevant conditions on the left-hand side, can drop these conditions to become better rules. In a more formal sense, the operation increases the simplicity of the hypothesis. In a good generalisation scheme this increase in simplicity is traded-off against a change in goodness-of-fit in order that the overall hypothesis preference measure is increased. Consider that we have a rule with the joint event $\mathbf{Y} = y, \mathbf{Z} = z$ as the left-hand side, where $y \neq z$. If we drop the condition $\mathbf{Z} = z$ then we have

$$p(\mathbf{Y} = y) \;>\; p(\mathbf{Y} = y, \mathbf{Z} = z) \qquad (8)$$

so that the simplicity component of the more general rule is always greater. In general, the generalisation step will only increase the J-measure if

$$j(\mathbf{X}; \mathbf{Y} = y) \;>\; \alpha.j(\mathbf{X}; \mathbf{Y} = y, \mathbf{Z} = z) \qquad (9)$$

where $\alpha = \frac{p(\mathbf{Y}=y, \mathbf{Z}=z)}{p(\mathbf{Y}=y)}$, i.e., if the fractional increase in simplicity is greater than the fractional decrease in cross-entropy.

Let us consider a very simple example of generalisation. Consider a rule in the reptile domain which says that

If is.snake is true **and** habitat is desert
then no.legs is true **with** probability 1

where $prob(is.snake) = prob(no.legs) = 0.3$, $prob(habitat = desert, is.snake) = 0.2$, and we generalise this to

If is.snake is true **then**
no.legs is true **with** probability 1.

We find that $J_g = 0.52 > J_s = 0.46$ bits, where J_g and J_s are the information contents of the general and more specialised version, respectively.

Specialisation:
This technique is used to refine hypotheses and is essentially the opposite of generalisation in that a decrease in simplicity is traded-off for an increase in goodness-of-fit in return for a better hypothesis, i.e., a higher preference measure. We will see later how the ITRULE algorithm employs specialisation to discover an optimal set of hypotheses. Let J_g and J_s be the information contents as defined earlier. Consider an example in the animal domain. The general rule might be

If has.wings is true
then can.fly is true **with** probability 0.9

and the more specialised version might be

If has.wings is true **and** is.penguin is false
then can.fly is true **with** probability 1.0

It is not intuitively obvious which rule is better on the average. If we specify $prob(can.fly = true) = 0.27$ and $prob(has.wings = true) = 0.3$ then we find that $J_s = 0.47$ bits and $J_g = 0.38$ bits, i.e., the more specialised rule is better. Without a quantitative measure, such as the J-measure, it would be very difficult to rank rules in this fashion.

Instance-based generalisation:
Instance-based generalisation proceeds not from rules but from examples. Most induction algorithms (statistical methods in particular) use this technique as the fundamental rule-generation mechanism. In our discussion so far we have used probabilities without

indicating where these probabilities were obtained. Theoretically there is an implicit assumption that these probabilities are "true" probabilities which translates into a practical assumption that they are estimated from frequency counts on very large samples of data. When large sample data *is* available (as for example in vision problems) then we would like to use the maximum-likelihood estimator for p, namely

$$\hat{p} = \frac{r}{N} \tag{10}$$

as an estimate for the true p (where r is the number of successes in N random trials). However if there are very few samples we would prefer to use a more conservative technique such as a maximum entropy estimate.

We propose the use of a point estimator which interpolates between an initial estimate and the maximum likelihood estimate as the sample size increases. We believe techniques such as this are an important step in bridging the gap between symbolic and statistical approaches. The estimator we use is

$$\hat{p} = \frac{\alpha + r + 1}{\alpha + \beta + N + 2} \tag{11}$$

where α and β are parameters of an intial density and can be obtained from initial subjective estimations [13]. The use of this estimator effectively introduces a sample-size dependent noise term with the desired effect that the information content of the rule (J-measure) increases with sample-size.

We have clearly shown that that the J-measure satisfies the basic requirements of an induction measure as defined by Holland et al. [12], i.e., it supports the basic inductive mechanisms of condition-simplifying generalisation, instance-based generalisation, and specialisation.

5 The ITRULE algorithm

We have introduced elsewhere [14] the notion of *generalised rule induction*. This is is the particular learning problem we address in this paper, i.e., the problem of discovering an optimal set of rules from a set of instances. Most previous work in the area of learning has concentrated on single concept learning or classification, e.g., decision trees. However in data-driven applications such as rule-based expert systems a more flexible approach is desirable. For example, decision trees are necessarily restrictive in representation and do not easily handle missing or uncertain information. Generalised rule induction is more general than single concept learning or discovering classification rules. In essence we wish to learn multiple hypotheses for multiple concepts — so we need a method for ranking hypotheses not only for the *same* but also for *different* concepts. The J-measure allows us to do exactly this. Automatic rule induction for expert systems is but one application of this idea, where we need not only classification rules but also rules linking intermediate concepts.

The ITRULE (Information-Theoretic **Rule** Induction) algorithm takes as input a set of feature vectors (where the N features are restricted to being discrete-valued) and it produces as output a set of K probabilistic rules. The rules are the K most informative rules available from the data as ranked by the J-measure (the parameter K is defined by the user). As previously mentioned, the rules are restricted to conjunctive expressions on the left, and a single expression on the right.

The algorithm cycles through each feature in turn as a right hand side. It keeps a ranked list of the K most informative rules determined up to that point. The information content of the Kth rule is used as a running minimum to determine whether or not new rules should be inserted in the list. For each feature the algorithm must find all possible left-hand side expressions which yield rules with greater information content than the running minimum. The search is constrained considerably using information-theoretic bounds on specialising the J-measure [15].

6 Experimental results

We have implemented the ITRULE algorithm together with comprehensive data manipulation tools, into a software package that runs on MacII and SUN workstations. In this section we show sample outputs of the algorithm, using published statistical data on mutual funds [16].

Figure 1 shows a set of typical raw data on mutual funds. Each line is an instance of a fund, and each column represents an attribute of the fund. Attributes can be numerical or categorical. From this raw data a second set of data is produced to serve as the input to ITRULE (Figure 2). The expert has a significant say in this process which serves to both categorise numerical data, and select attributes of interest. Numerical data is categorised using two techniques. First, the expert can identify "obvious" categorisations. For example, the 5 year return can be compared with the Standard and Poor's 500 index, ("S&P", above/below). Second, the software uses a maximum entropy algorithm to automatically identify statistically significant categorisations. The expert can accept this advice or modify the value to make the categorisation more meaningful.

The ITRULE software then processes this table

Fund Type	5 Year Return%	diver-sity	Beta (Risk)	Bull Market Perfom.	Bear Market Perfom.	Stocks %	Largest Holding	Distributed Dividends (per share)	Distributed Cap. Gains (per share)	Net Asset Value $	Distri-butions as % of NAV	Portfolio Turnover Rate %	Total Assets ($M)
Balanced	135.6	C	0.8	B	D	87	information	0.5	6.07	37.27	17.63	34	414.6
Growth	32.5	C	1.05	E	B	81	manufacture	0.11	0	12.49	0.88	200	16.2
Growth&Income	88.3	A	0.96	C	D	82	financial	0.16	0.41	11.92	4.78	127	26.8
Agressive	-24.4	A	1.23	E	E	95	oil	0	0.6	6.45	9.30	161	64.2
Growth&Income	172.2	E	0.59	A	B	73	officeequip	0.52	0.84	13.64	9.97	31	112.5
Growth&Income	80.4	A	0.58	D	B	98	computers	0.45	2.2	13.7	19.34	64	76.4
Growth	52.7	B	0.86	E	D	95	airlines	0.24	0	8.72	2.75	54	11.7
Balanced	143.8	C	0.71	B	B	51	materials	0.66	0.7	13.03	10.44	239	190
Agressive	91.7	B	1.24	C	E	100	technology	0.04	1.8	8.11	22.69	218	96.2
Growth	105.4	A	0.99	B	D	98	energy	0.55	3.21	13.62	27.61	20	253.8
Growth	93	C	0.91	C	B	94	drugs	0	0	27.44	0.00	8	3.9
Growth	148.9	A	0.9	A	E	87	financial	0.5	5.8	32.4	19.44	37	452.9

Figure 1: Part of initial data set

Fund Type	5 Year Return%	diver-sity	Beta (Risk)	Bull Market Perfom.	Bear Market Perfom.	Common Stocks >90%	Distributions	Portfolio Turnover Rate	Total Assets
Balanced	below	average	under1	good	poor	no	high	low	large
Growth	below	average	over1	poor	good	no	low	high	small
Growth&Income	below	high	under1	average	poor	no	low	high	small
Agressive	below	high	over1	poor	poor	yes	low	high	small
Growth&Income	above	low	under1	good	good	no	low	low	large
Growth&Income	below	high	under1	poor	good	yes	high	low	small
Growth	below	high	under1	poor	poor	yes	low	low	small
Balanced	above	average	under1	good	good	no	low	high	large
Agressive	below	high	over1	average	poor	yes	high	high	small
Growth	below	high	under1	good	poor	yes	high	low	large
Growth	below	average	under1	average	good	yes	low	low	small
Growth	above	high	under1	good	poor	no	high	low	large

Figure 2: Quantised data

	IF	AND	THEN	p(x/y)	p(y)	p(x)	J(X;y)	J(X;y)
1	5yrReturn aboveS&P		Bull_perf good	0.966	0.311	0.511	0.75411	0.23461
2	Bull_perf good	Bear_perf NOT poor	5yrReturn aboveS&P	0.242	0.356	0.689	0.60743	0.21597
3	Bull_perf NOT good		5yrReturn belowS&P	0.978	0.489	0.689	0.40940	0.20015
4	5yrReturn belowS&P	Assets small	Bull_perf NOT good	0.159	0.478	0.511	0.39009	0.18638
5	Bull_perf good	Bear_perf good	5yrReturn aboveS&P	0.167	0.189	0.689	0.84334	0.15930
6	type NOT A	Bull_perf good	Assets large	0.786	0.456	0.456	0.32960	0.15015
7	diversity NOT low	Assets large	Bull_perf good	0.875	0.344	0.511	0.43274	0.14906
8	Bear_perf NOT poor	Assets large	5yrReturn aboveS&P	0.269	0.278	0.689	0.53537	0.14872
9	Beta over1	Bear_perf poor	stocks>90% yes	0.111	0.189	0.611	0.78686	0.14863
10	type NOT G	Bull_perf good	5yrReturn aboveS&P	0.286	0.300	0.689	0.49371	0.14811
11	Beta over1	Bear_perf poor	type A	0.722	0.189	0.244	0.72779	0.13747
12	type NOT A	Assets large	Bull_perf NOT poor	0.025	0.433	0.267	0.31528	0.13662
13	type NOT A	Assets large	Bull_perf good	0.825	0.433	0.511	0.31050	0.13455
14	Bull_perf poor		Assets small	0.080	0.267	0.456	0.49554	0.13214
15	type NOT A	Bull_perf good	5yrReturn aboveS&P	0.381	0.456	0.689	0.28889	0.13161
16	stocks>90% no	Assets large	Bull_perf good	0.848	0.356	0.511	0.36439	0.12956
17	type NOT GI	Bull_perf NOT good	Assets small	0.139	0.389	0.456	0.33154	0.12893
18	Bear_perf NOT poor	Assets large	Bull_perf good	0.885	0.278	0.511	0.45974	0.12771
19	Bull_perf good	stocks>90% no	5yrReturn aboveS&P	0.361	0.389	0.689	0.32676	0.12707
20	Assets large		Bull_perf good	0.810	0.456	0.511	0.27804	0.12666

Figure 3: ITRULE rules

to produce a set of rules. The rules are ranked in order of decreasing information according to the J-measure. The K most informative rules are output where K is specified by the user. The user can also specify the maximum order of the rules, that is, the maximum number of left hand conjunctives. Figure 3 shows a portion of the ITRULE output for the mutual fund data set (the transition probabilities as listed may correspond to either $x|y$ or $\bar{x}|y$). Several points of interest emerge. We note that "obvious" rules appear, confirming that the algorithm is on the right track. For example, "if the performance in a Bull market is good and the performance in a Bear market is good then the 5 year return is better than S&P500." We also see that a rule does not have to have a very high transition probability to be in the set. Rules such as rule 20, "if the fund assets are large then the 5 year return is better than S&P500" are interesting in that they represent *new* hypotheses which have been discovered in the data. The potential application of ITRULE for automated knowledge-acquisition is clearly demonstrated even in this quite simple example.

7 Conclusion

In this paper we have clearly demonstrated the applicability of the recently proposed J-measure for induction from both a theoretical and practical standpoint. In the early sections of the paper we developed an interpretation of the measure as a hypothesis preference criterion which trades off simplicity and goodness-of-fit. We followed this by investigating how the measure supports the basic inductive mechanisms of generalisation and specialisation. Finally we described the ITRULE algorithm and gave a practical example of its use. We can conclude that the relatively simple idea of the J-measure can support inductive procedures very well and is an interesting application of a statistical technique which takes into account the theoretical aspects of induction.

References

1. T. M. Mitchell, 'Version spaces : a candidate elimination approach to rule learning,' *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Los Altos : CA, Morgan Kaufmann, 1977.
2. R. S. Michalski and R. L. Chilausky, 'Learning by being told and learning from examples', *International Journal of Policy Analysis and Information Systems* 4, pp.125-161, 1980.
3. L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworth, 1984.
4. 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.
5. R. M. F. Goodman and P. Smyth, 'Decision tree design from a communication theory standpoint,' accepted for publication in *IEEE Transactions on Information Theory*.
6. L. G. Valiant, 'A theory of the learnable,' *Communications of the ACM*, vol.27, no.11, pp.1134-42.
7. R. S. Michalski, 'Pattern recognition as rule-guided inference', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2, pp.349-361, 1980.
8. R. M. F. Goodman and P. Smyth, 'An information-theoretic model for rule-based expert systems,' to be presented at the 1988 International Symposium on Information Theory, Kobe, Japan.
9. D. Angluin and C. Smith, 'Inductive inference: theory and methods,' *ACM Computing Surveys, 15(3)*, pp. 237-270.
10. B. R. Gaines, 'Behaviour/structure transformations under uncertainty,' *Int. J. Man-Mach. Stud. 8*, pp. 337-365.
11. J. E. Shore and R. W. Johnson, ' Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy,' *IEEE Transactions on Information Theory*, vol.IT-26, no.1, Jan 1980, pp.26-37.
12. J. H. Holland, K. J. Holyoak, R. E. Nisbett, P. R. Thagard, *Induction: Processes of Inference, Learning and Discovery*, Cambridge, MA: MIT Press, 1986.
13. I. J. Good, *The estimation of probabilities: an essay on modern Bayesian methods*, Research monograph no.30, M.I.T. Press, Cambridge: MA, 1965.
14. R. M. F. Goodman and P. Smyth, 'ITRULE: an information-theoretic rule induction algorithm', *Proceedings of the First European Workshop on Knowledge Acquisition*, Reading, England, Sept. 1987.
15. P. Smyth, *The Application of Information Theory to Problems in Decision Tree Design and Rule-based Expert Systems*, Phd. thesis, Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, June 1988.
16. American Association of Investors, *The individual investor's guide to no-load mutual funds*, International Publishing Corporation: Chicago, 1987.