Machine Learning, 6, 67-80 (1991)

Information-Based Evaluation Criterion for Classifier's Performance

IGOR KONONENKO
Faculty of Electrical and Computer Engineering, Trzaska 25, Ljubljana, Yugoslavia

IVAN BRATKO
Faculty of Electrical and Computer Engineering, Trzaska 25, Ljubljana, Yugoslavia and Jozef Stefan Institute, Jamova 39, Ljubljana, Yugoslavia

Editor: David Haussler

Abstract. In the past few years many systems for learning decision rules from examples were developed. As different systems allow different types of answers when classifying new instances, it is difficult to appropriately evaluate the systems' classification power in comparison with other classification systems or in comparison with human experts. Classification accuracy is usually used as a measure of classification performance. This measure is, however, known to have several defects. A fair evaluation criterion should exclude the influence of the class probabilities which may enable a completely uninformed classifier to trivially achieve high classification accuracy. In this paper a method for evaluating the information score of a classifier's answers is proposed. It excludes the influence of prior probabilities, deals with various types of imperfect or probabilistic answers and can be used also for comparing the performance in different domains.

Keywords. Classifier, evaluation criteria, machine learning, information theory

1. Introduction

The following is a rather general statement of the problem of learning from examples:

Given: A definition of a problem domain and a set of training examples with known classes.

Find: A rule that explains the learning set and can be used to classify new objects from the same domain with unknown classes.

This definition applies to many known systems for learning decision rules from examples, such as ID3 (Quinlan, 1979), ACLS (Paterson & Niblett, 1982), CART (Breiman et al., 1984) C4 (Quinlan, 1986), CN2 (Clark & Niblett, 1987, 1989), AQ and its successors (Michalski et al., 1986), Assistant (Bratko & Kononenko, 1987; Cestnik et al., 1987) and many others.

Several authors compared the performance of different inductive learning systems with the performance of human experts, other learning systems and classifiers based on statistical pattern recognition (Michalski & Chilausky, 1980; Horn et al., 1985; Mozetic et al., 1986; Bratko & Kononenko, 1987; Clark & Niblett, 1989). The aspect of comparison is usually

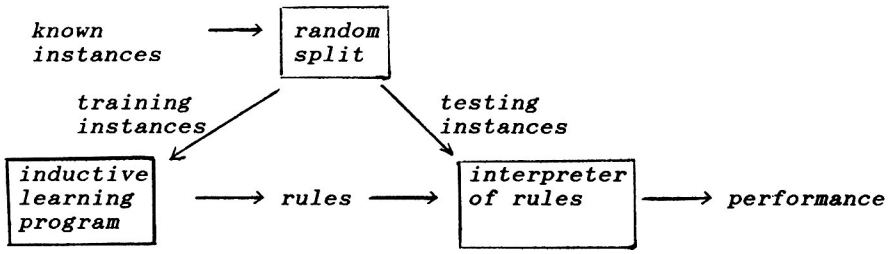

Figure 1. One cycle of the procedure for evaluating the performance of a machine learning systems.

the classification accuracy of a classifier on new, not previously seen objects. Usually, to assess the accuracy of generated rules the procedure depicted in Figure 1 is repeated N times and the results are averaged.

As different classifiers return different kinds of answer and as there are several interpretations of the term "performance," the comparison is difficult. There is clear need for unbiased method of comparison that would take into account the informativity of a classifier's answers.

In Section 2, problems with evaluation of classifiers' performance are discussed in detail. In Section 3, an information based evaluation criterion is formally defined and some of its properties are described. Experimental results in several medical domains are described in Section 4.

2. Problems with estimation of performance

We start the discussion of problems in performance evaluation with an episode from our own work. In experiments in two medical domains, 'breast cancer' and 'primary tumor' (Cestnik et al., 1987), the Assistant learning program achieved respectively 77% and 44% classification accuracy (medical experts achieve similar, slightly lower accuracy, see Section 4). At first glance the result in diagnosing the location of primary tumor is poor while the accuracy of prognosing the recurrence of breast cancer is acceptable. In fact these results were frequently interpreted in that way by serious discussants.

A closer look shows a completely different picture. In 'primary tumor' there are 22 possible classes while in 'breast cancer' only 2. In addition, in 'breast cancer' the prior probability of one of the two classes is 80% while the prior probability of the majority class in 'primary tumor' is only 25%. Therefore 80% classification accuracy in 'breast cancer' can be trivially achieved, while 44% in 'primary tumor' is fairly hard to achieve. Obviously, the percentage of correct classifications in these cases is not a very appropriate measure of performance.

In the following subsections the problems with evaluating the classifier's performance are described more systematically. We discuss these difficulties under the following categories:

1. Different classifiers produce different forms of answers that cannot be directly compared.

2. Taking into account the prior probabilities of classes in performance evaluation.
3. Comparison of performance in different domains.

The following notation will be used for probabilities:

$P(C)$ — the prior probability of class C
$P'(C)$ — the probability of class C returned by the classifier

2.1. *Different kinds of answers*

The following are possible answers of a classifier when classifying a particular instance:

1. The classifier returns a unique class C, then $P'(C) = 1$. We call such a classifier *categorical*.
2. The classifier returns a set of N possible classes, for example the AQ algorithm (Michalski et al., 1986) or the SEARCH leaf in the ID3 program (Quinlan, 1979). One way of treating such an answer is as:

   $P'(C) = 1/N$, for all classes C in the returned set.

3. The classifier cannot classify a particular instance at all, for example the NULL leaf in ID3 (Quinlan, 1979). Such an answer can be interpreted as

   $P'(C) = P(C)$, for all classes C.

4. The classifier returns a distribution $P'(C)$ over all classes C, as in Assistant (Cestnik et al., 1987).

The problem is how to handle such different kinds of possible answers when comparing the performance of different classifiers. The answer of the first kind above (exact answer) can be correct or incorrect. The other kinds of answers may also be correct (although less precise), or they can be partially correct. Such imperfect answers can still be useful as partial information is better than nothing. In some domains exact answers are even impossible since there is not enough information available for exact classification.

The fourth kind of answer is obviously the most general and covers all the others. An evaluation criterion that can handle this kind of answer would be general enough for comparison of different classifiers.

2.2. *The influence of prior probabilities*

In a problem domain where the prior probability of one class is very high the 'naive' classifier that classifies all instances into the most likely class would achieve a high classification accuracy. We feel that such a high accuracy result, achieved in a simple-minded way in

an easy domain, should not score as high as in a more difficult domain. The evaluation criterion should therefore take into account the prior probabilities of classes.

Obviously the correct classification into a more probable class is an easier task than the correct classification into a less probable class. On the other hand, the misclassification of a more probable class should count as a more serious mistake than misclassification of a less probable class since the former is not expected while the latter would not be surprising.

Let us have two classes C_1 and C_2 and let $P(C_1) > P(C_2)$. If we denote the credit for correct classification into class C with $V_c(C)$, and the penalty for misclassification with $V_m(C)$, then the following should hold:

$$V_c(C_1) < V_c(C_2)$$

and

$$V_m(C_1) > V_m(C_2)$$

The usual evaluation criteria such as the relative frequency of correct classifications or the number of correct classifications minus the number of incorrect classifications do not take into account the difference between prior probabilities of classes.

To overcome this problem, multiple evaluation parameters can be used which all together give more detailed picture of the performance. For example, in the case that only two diagnoses are possible (positive and negative) five standard measures of diagnostic performance are usually used which combine four possible outcomes: true positive *TP* (the number of correct positive predictions), true negative *TN* (the number of correct negative predictions), false positive *FP* (the number of incorrect positive predictions), and false negative *FN* (the number of incorrect negative predictions). These five measures are (Williams, 1982; Weiss et al., 1987):

$$\textit{Sensitivity} = TP/(TP + FN),$$
$$\textit{Specificity} = TN/(FP + TN),$$
$$\textit{Positive predictive value} = TP/(TP + FP),$$
$$\textit{Negative predictive value} = TN/(FN + TN) \text{ and}$$
$$\textit{Efficiency} = (TP + TN)/(TP + TN + FP + FN) = \textit{Classification accuracy}$$

However, it is cumbersome to use five parameters at once. Usually efficiency is used to compare the performance of different classifiers (we will refer to it as *accuracy*). Besides, as they stand these measures cannot be directly applied to probabilistic answers and, except accuracy, to decision problems with more than two classes. It is not clear how they can be generalized in these two directions.

Spackman (1989) suggested the use of ROC (Relative Operating Characteristic) curves for evaluating the performance of empirical learning systems. ROC curves are in fact a graphical representation of the relation between sensitivity and specificity and are used for evaluating classification performance in signal detection (for example Winter, 1982; Ripley et al., 1975; Hansmann et al., 1976). However, the ROC analysis assumes classification

based on a test (continuous attribute, possibly constructed from several attributes) whose result is normally distributed. The classification result depends on whether the test result exceeds a decision threshold (Winter, 1982). For our purpose such assumptions are inappropriate an often unrealistic. In addition, again, it is not clear how the ROC analysis can be naturally extended to more than two classes and probabilistic answers.

2.3. Comparison of performance in different domains

To compare the performance in different domains one should be able to estimate the difficulty of decision problems. A problem with more classes is generally more difficult than a problem with fewer classes. If we split one class into two classes then the obtained classification problem is harder than the original one. Also the problem with equal prior probabilities for all classes is harder than the problem with great differences between prior probabilities.

The comparison of performance in different domains is questionable also because in different domains different amounts of information may be available for classification. If a domain is incomplete then exact classification is even impossible. (A domain is said to be incomplete if the available attributes do not suffice to completely distinguish between classes. As a consequence, there can be clashes between examples — different classes corresponding to the same attribute vetor.) The amount of available information also depends on the number of available training instances. In this paper we will not address the problem of the amount of information available for classification.

In some domains there may be also other criteria that influence the classification such as the time and cost requirements. Such parameters should also be included when evaluating the classification performance. In this paper we will not be concerned with such parameters.

3. Informativity based evaluation criterion

3.1. The amount of information

The *uncertainty* or the *entropy* of an event C with prior probability $P(C)$ is defined (Shannon & Weaver, 1949) as

$$-log\ P(C)\ [bit] \tag{1}$$

Throughout the paper the log function is assumed to be of base 2. Expression (1) also defines the amount of information necessary to correctly classify an instance into class C whose prior probability is $P(C)$. Analogously the amount of information necessary to correctly decide that an instance does not belong to class C is

$$-log\ (1 - P(C))\ [bit] \tag{2}$$

If we have N classes C_i, $i = 1..N$, then the *expected* amount of information for classification of one instance is equal to the *entropy* of a distribution:

$$E = -\sum_{i}^{N} P(C_i) * log\ P(C_i)\ [bit] \tag{3}$$

where

$$\sum_{i}^{N} P(C_i) = 1$$

3.2. *Information score of an answer*

Let the correct class of an instance be C. Recall that $P(C)$ is the prior probability of class C and $P'(C)$ is the posterior probability returned by the classifier. We consider two cases:

(a) $P'(C) > P(C)$

Here the probability of class C has changed in the right direction, therefore we will call such an answer *useful*. It should be awarded a positive score.

(b) $P'(C) < P(C)$

Here the probability of class C has changed in the wrong direction, therefore we will call such an answer *misleading*. It should be assigned a negative score.

In addition we have the case $P'(C) = P(C)$. Such an answer contains no information and should score 0.

We reserve the terms "correct answer" and "incorrect answer" just for two special cases when $P'(C) = 1$ and $P'(C) = 0$ respectively. This means that only categorical answers are correct or incorrect. Of course, a correct answer is also useful, and an incorrect one is also misleading.

Reasons why we cautiously use the terms "useful" and "misleading" deserve explanation. In the case $P'(C) > P(C)$ the classifier is essentially saying: "After I have examined the available evidence about the instance, my belief that the instance is of class C has increased." The classifier has thus changed its belief in the right direction, although it is not necessarily saying that the class is now C. It is tempting to simply call such an answer "correct." This might, however, cause terminological confusion associated with the interpretation of the classifier's answer. The difficulty is illustrated by the following example. Suppose we have two classes, C_1 and C_2, with prior probabilities $P(C_1) = 0.9$ and $P(C_2) = 0.1$. Now, let the correct class be C_2 and the classifier's answer is $P'(C_2) = 0.4$ (and correspondingly $P'(C_1) = 0.6$). Many people would here implicitly assume the maximum likelihood decision rule and interpret this answer as if the classifier is saying: class = C_1. Accordingly,

they would conclude that the answer is incorrect. Notice that this judgment derives from the assumed maximum likelihood rule. In this paper, however, the maximum likelihood rule is never assumed and the interpretation of the classifier's answer is left to the user. All that the classifier is really saying is: "Now I think that class C_2 has become more likely than it looked before."

Another example will show that an answer that is incorrect under the maximum likelihood interpretation can otherwise be regarded as highly informative and useful. Suppose in 1950 a classifier was shown a list of one million names and asked: "Who of these people will be the president of the U.S.A. in 1990? It looks they are all equally likely at the moment." So we have one million classes all with prior probabilities 10^{-6}. Suppose the classifier in 1950 answered:

$$P'(Bush) = 0.45$$
$$P'(Dukakis) = 0.55$$
$$P'(all\ others) = 0$$

Strictly speaking, this answer can be viewed as incorrect. However, normally it would be considered as highly insightful, informative and useful.

Let the correct class of an instance be C. According to the two cases of a classifier's answer being useful or misleading, we define the *information score I of the answer* separately for each case as follows:

(1) if $P'(C) \geq P(C)$ then

$$I = V_c(C) = -log\ P(C) + log\ P'(C)\ [bit] \tag{4}$$

The intuitive basis for this definition is this: information score indicates the amount of obtained information. This is the entire amount of information necessary to correctly classify an instance into class C (see (1)), minus the remainder of information necessary to correctly classify that instance. Note that if $P'(C) = P(C)$ then $I = 0$; in this case the classifier did not change the prior probability of the correct class, therefore its answer did not provide any information.

(2) if $P'(C) < P(C)$ then the answer is misleading and we define the penalty

$$V_m(C) = -log\ (1 - P(C)) + log\ (1 - P'(C))\ [bit] \tag{5}$$

Here the amount of information returned by the system is the entire amount of information necessary to decide that an instance does not belong to class C (see (2)), minus the remainder of information necessary to make that decision. As the classifier's answer here is misleading, the information score of the answer is negative and defined as the negation of the penalty (5):

$$I = -\ V_m(C)\ [bit] \tag{6}$$

It would be tempting to simplify these definitions and to define the information score of a classifier's answer simply by (4) for *any* $P'(C)$. Then in the case of $P'(C) < P(C)$ (misleading answer), the information score would be negative which is appropriate. There is a difficulty, however. Intuitively, two answers, a useful and a misleading one, may compensate each other so that their total score should be 0. The simplified definition, however, does not amount to this effect. Let us consider a counter example.

Let $P(C) = 1/2$, and the classifier has classified two objects of class C. The first one was misclassified ($P'(C) = 0$), and the second was classified correctly ($P'(C) = 1$). Intuitively, in this case the total score should be 0. The simplified definition, however, gives

$$I_1 + I_2 = (- log(1/2) + log\ 0) + (- log(1/2) + log\ 1) = - \infty$$

Thus, one classifier's mistake is more costly than any finite number of correct classifications! On the other hand our definition, (4) and (5), gives:

$$I_1 + I_2 = - (- log(1/2) + log\ 1) + (- log(1/2) + log\ 1) = 0$$

which is intuitively satisfying.

Note that the definition assumes that prior probabilities of classes are sufficiently accurately known. In practice they are usually approximated with relative frequencies from training instances. Note also that the probabilities of other classes returned by the classifier are ignored, that is only the probability of the correct class is considered.

The average information score I_a of an answer over a test set can be calculated as the sum of information scores of all T testing instances divided with T:

$$I_a = 1/T \sum_{j}^{T} I(j) \tag{7}$$

We define the *relative information score* I_r as:

$$I_r = I_a/E * 100\% \tag{8}$$

where E is the entropy (3). Note that for fair evaluation, the definition of I_r requires that the distribution in a testing set is equal to the distribution in the domain.

An alternative attempt would be to define I_r as the average of information scores normalized with $-log\ P(C)$ for case (a) and $-log(1 - P(C))$ for case (b). However, this would not be appropriate as it excludes the effect of the prior probability in answer scoring. As a consequence, the contribution to I_a of correct prediction of a likely class would count the same as that of an unlikely class.

3.3. Properties of the information based evaluation criterion

1. The proposed performance evaluation measure applies to all kinds of answers of different classification systems (see 2.1) because it deals with probability distributions.

2. The correct classification into a more probable class is less valuable than the correct classification into a less probable class (see 2.2):

$$P(C_1) > P(C_2) \Rightarrow V_c(C_1) < V_c(C_2), \text{ see (4)},$$

since

$$-\log P(C_1) < -\log P(C_2)$$

3. The incorrect classification of an instance that belongs to a more probable class is a more serious mistake than the incorrect classification of an instance that belongs to a less probable class (see 2.2):

$$P(C_1) > P(C_2) \Rightarrow V_m(C_1) > V_m(C_2), \text{ see (5)},$$

since

$$-\log (1 - P(C_1)) > -\log (1 - P(C_2))$$

4. If the classifier always answers correctly with the probability of the correct class $P'(C) = 1$ and if the distribution in a testing set of instances is equal to the distribution in the domain then the average information score (7) of the system's answer is equal to the expected necessary information for the correct classification of one instance (3):

$$I_a = E$$

5. If the classifier always correctly answers the probability of the correct class $P'(C) = 1$ and if the distribution in a testing set of instances is equal to the distribution in the domain then the relative information score (8) is

$$I_r = 1$$

However, if the classifier always answers incorrectly with the probability of the correct class $P'(C) = 0$ then I_r in general differs from -1. The lack of symmetry shows that the information score of correct classification is not equal to the penalty of incorrect classification. This lack of symmetry should not be regarded as an anomaly. It merely reflects the asymmetry in the classification problem itself: it is not equally difficult to predict the correct class as to predict an incorrect class.

6. If the classifier returns prior probabilities of classes then the information score of such a classification is:

$$I = 0$$

7. Entropy (3) can be used to evaluate the difficulty of a decision problem (see Section 2.3). For example, consider a decision problem with entropy E and we split one class C into two subclasses C_1 and C_2 then the new entropy E' is *greater*:

$$E' > E$$

since

$$P(C) = P(C_1) + P(C_2), \quad P(C_1), P(C_2) > 0,$$

and from (3)

$$E' = E + P(C) * log\ P(C) - P(C_1) * log\ P(C_1) - P(C_2) * log\ P(C_2)$$

and

$$-P(C) * log\ P(C) < - P(C_1) * log\ P(C_1) - P(C_2) * log\ P(C_2)$$

since

$$- log\ P(C) < - log\ P(C_i), \text{ for } i = 1, 2$$

Also the entropy in a domain with equal prior probabilities of classes is greater than the entropy in a domain with unequal prior probabilities because the entropy function has its maximum when all probabilities are equal (Shannon & Weaver, 1949). However, this measure does not take into account other properties of a classification problem, such as the number of attributes and their informativity.

8. The relative information score I_r (8) can be used to compare the performance of different classifiers in the same domain while the average information score I_a (7) of an answer can be used to compare the performance in different decision problems. In a more difficult problem, a lower relative information score suffices to attain the same average information score.

   For example, let us have two domains D_M with D_N with M and N classes respectively. let $M \leq N$ and let in both domains the prior probabilities of all classes be equal (that is $1/M$ and $1/N$ respectively). Let us further have an ideal categorical classifier that 100% correctly classifies in the domain D_M. The average information score of this ideal classifier is:

$$I_a' = log\ M \tag{9}$$

Next, consider a categorical classifier that correctly classifies with probability p in the domain D_N. The average information score of this classifier is

$$I_a'' = p * log\ N + (1 - p) * log\ (1 - 1/N) \tag{10}$$

If we want to obtain the same average information score per answer in both cases

$$I_a' = I_a'' \tag{11}$$

what is the corresponding probability of correct classification of the second classifier? This follows from (9), (10) and (11):

$$p = \frac{log((N - 1)/(N * M))}{log((N - 1)/(N * N))} \tag{12}$$

For example, for $M = 2$ and $N = 4$, p is 0.59. In other words, an imperfect categorical classifier for a domain with 4 equally likely classes has to predict correctly 59% of the time to attain the same average information score as the perfect classifier in a domain with 2 classes.

4. Some experimental results

In Table 1 the classification accuracy and information score per answer of the Assistant inductive learning program (Cestnik et al., 1987) in four medical domains is compared with the performance of the 'naive' Bayes classifier (Bayes classifier with the assumption that the attributes are independent) and the performance of human experts (Bratko & Kononenko, 1987).

The results with Assistant and the naive Bayes classifier are the averages of 10 runs of the procedure of Figure 1 by randomly selecting 70% of available instances for training and 30% for testing each time. The performance of physicians was obtained by testing 4 specialists (in breast cancer and primary tumor domains also nonspecialists) and calculating the average performance. For classification, human experts had available exactly the same set of attributes.

Table 1. The comparison of performance of Assistant 86, Bayes classifier and human experts (specialists and nonspecialists) in four medical domains: diagnosis in rheumatology, diagnosing thyroid diseases, prognosing the recurrence of the breast cancer and localizing the primary tumor.

Medical Domain	Classifier	Accuracy	Inf. Score	Rel. Inf. Score
Rheumatology	Assistant	61%	0.46 bit	27%
Rheumatology	Bayes	57%	0.28 bit	16%
Rheumatology	Specialists	56%	0.26 bit	15%
Thyroid	Assistant	73%	0.87 bit	55%
Thyroid	Bayes	68%	0.70 bit	44%
Thyroid	Specialists	64%	0.59 bit	37%
Breast cancer	Assistant	77%	0.07 bit	10%
Breast cancer	Bayes	79%	0.06 bit	9%
Breast cancer	Specialists	64%	0.05 bit	7%
Breast cancer	Nonspecialists	64%	0.03 bit	4%
Primary tumor	Assistant	44%	1.38 bit	38%
Primary tumor	Bayes	49%	1.59 bit	44%
Primary tumor	Specialists	42%	1.22 bit	34%
Primary tumor	Nonspecialists	32%	0.95 bit	26%

In Table 2 characteristics of each domain are presented. Notice that the evaluation of the information score of an answer is well motivated because in all domains the majority class is much more probable than the other classes. More details about the experiments can be found in (Bratko & Kononenko, 1987).

At first glance it seems that 77% classification accuracy in the breast cancer domain is a much better result than 44% in the primary tumor domain. But the avrage information score per answer shows that in breast cancer the classifier in fact does very little (0.07 bit per answer) while in the primary tumor domain the classifier achieves much better average information score (1.38 bit). This is in fact the best achievement in the four domains, although it looks the weakest simply from the accuracy point of view.

In fact, the calculated entropy (3) in Table 2 shows that the localization of the primary tumor is the hardest problem (entropy = 3.64 bit) among the four, while the breast cancer problem is the easiest (entropy = 0.72 bit). The classification accuracy of 80% can be achieved in the breast cancer domain with a default classifier which classifies always into the majority class.

It also seems that the naive Bayes classifier achieved slightly better result (79%) than Assistant (77%) in the breast cancer domain. In fact the information score of Assistant's answers is slightly higher than that of the naive Bayes classifier. It can be concluded accordingly that the Bayesian classifier emphasizes the prior probability while Assistant tends to optimize information score. This also directly follows from their learning schemes.

The difference between the performance of machine classifiers and physicians can be partially attributed to the fact that, when tested, physicians were asked to select only one diagnosis for each patient although in certain cases the probability distribution would be more natural and probably result in better performance (this can be also seen from Table 1). Another reason for the difference is the fact that physicians, when diagnosing, do not only try to maximize the accuracy of their answer. They also take into account some other factors, like the influence of a diagnosis to the choice of medical treatment and what discomfort that would cause to the patient.

5. Conclusions

The information based evaluation criterion for evaluating the performance of classifiers can deal with incomplete classifications and takes into account the prior probabilities of classes. Its interpretation is natural.

Table 2. Characteristics of four medical domains

Medical Domain	No. Attributes	No. Classes	No. Available Instances	Probability of Majority Class	Entropy (bits)
Rheumatology	32	6	355	66%	1.70
Thyroid	15	4	884	56%	1.59
Breast cancer	10	2	288	80%	0.72
Primary tumor	17	22	339	25%	3.64

The proposed evaluation criterion can explain unusual situations such as one in Table 1, where in the breast cancer domain nonspecialists achieved the same predictive accuracy as specialists. In fact the informaton score of specialists' answers was slightly higher. However, physicians believe that today's medicine knows very little about prognosing the recurrence of breast cancer, and this is confirmed by the poor results in Table 1.

Relative information score of an answer can be used to compare the performance of different classification systems and human experts in the same problem domain. The average information score can also be used to compare the performance of different classifiers in different problem domains.

Figure 2 shows the relation between the number of classes and the required percentage of correct guesses defined with (12) by a categorical classifier to achieve the same average information score per answer as a perfect classifier in a two-class domain (100% correct answers). Equal prior probabilities of all classes are assumed ($1/N$ for a domain with N classes).

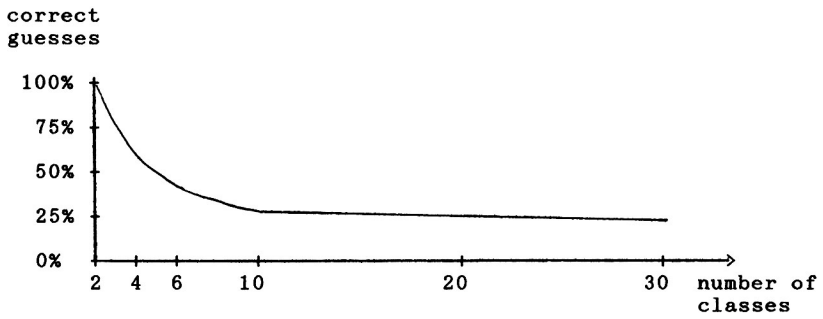

Figure 2. The required percentage of correct guesses (11) by an exact classifier to achieve the same average information score per answer as a perfect categorical classifier in a two-classes domain.

Acknowledgments

We thank Donald Michie for suggesting the entropy as a basis for evaluating the performance of classifiers. Dr. Sergej Hojker, Dr. Vlado Pirnat and Dr. Matjaz Zwitter from University Clinical Center in Ljubljana provided the medical data and tested the physicians' performance. The comments of reviewers significantly helped improving the paper.

References

Bratko, I., & Kononenko, I. (1987). Learning rules from incomplete and noisy data. In B. Phelps (Ed.), *Interactions in artificial intelligence and statistical methods*. Hampshire, England: Technical Press.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*, Belmont, California: Wadsworth, Int. Group.

Cestnik, B., Kononenko, I., & Bratko, I. (1987). ASSISTANT 86: A knowledge elicitation tool for sophisticated users. In I. Bratko, N. Lavrac (Eds.), *Progress in machine learning*. Wilmslow, England: Sigma Press.

Clark, P., & Niblett, T. (1987). Learning if then rules in noisy domains. In B. Phelps (Ed.), *Interactions in artificial intelligence and statistical methods*. Hampshire, England: Technical Press.

Hansmann, D.R., Sheppard, J.J., & Yeshaya, A. (1976). Evaluation of the Dyna-Gram Holter ECG Analysis System, *Computers in Cardiology, 1976*, 171–182.

Horn, K.A., Compton, P., Lazarus, L., & Quinlan, J.R. (1985). An expert system for the interpretation of thyroid assays in a clinical laboratory. *The Australian Computer Journal, 17*, 7–11.

Michalski, R.S., & Chilausky, R.L. (1980). Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems, 4*, 125–161.

Michalski, R.S., Mozetic, I., Hong, J., & Lavrac, N. (1986). The multipurpose incremental learning system AQ15 and its testing application to three medical domains. *Proceedings of the National Conference on Artificial Intelligence AAAI 86*. Philadelphia.

Mozetic, I., Lavrac, N., & Kononenko, I. (1986). Automatic construction of diagnostic rules. *Proceedings of the Fourth Mediterranean Conference on Medical & Biological Engineering*. Sevilla, Spain.

Paterson, A., & Niblett, T. (1982). *The ACLS user manual*. Glasgow: Intelligent Terminals Ltd.

Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the microelectronic age*. Edinburgh University Press.

Quinlan, J.R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Ripley, K.L., & Arthur, R.M. (1975). Evaluation and comparison of automatic arrhythmia detectors, *Computers in Cardiology, 1975*, 27–32.

Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communications*. Urbana, IL: The University of Illinois Press.

Spackman, K.A. (1989). Signal detection theory: Valuable tools for evaluating inductive learning, *Proceedings of the Sixth International Workshop on Machine Learning*, (pp. 160–163) Ithaca, NY: Cornell University.

Weiss, S.M., Galen, R.S., & Tadepalli, P.V. (1987). Optimizing the predictive value of diagnostic decision rules. *Proceedings of the Sixth National Conf. on Artificial Intelligence AAAI-87*, (pp. 521–526) Seattle, Washington.

Williams, B.T. (Ed.) (1982). *Computer aids to clinical decisions* (Vol. I & II). Boca Raton, FL: CRC Press.

Winter, J. (1982). Computer assessment of observer performance by receiver operating characteristic curve and information theory, *Computers and Biomedical Research, 15*, 555–562.