Evaluation and Comparison of Pattern Classifiers for Chemical Applications

Leonard J. Soltzberg,[1] Charles L. Wilkins,* Steven L. Kaberline, T. Fai Lam, and Thomas R. Brunner

Contribution from the Department of Chemistry, University of Nebraska—Lincoln, Lincoln, Nebraska 68588. *Received April 9, 1976*

Abstract: The application of several different evaluation criteria to the performance of pattern classifiers used for chemical pattern recognition indicates the need for caution in evaluating and comparing classification methods. Important considerations include the distinction between recognition and prediction performance, the similarity of test data sets, the distribution of compounds within the various categories in these test sets, and the distinction between binary and multicategory classification. It is shown that the class conditional probabilities and a proposed figure of merit (M) are well suited for selection of classifiers with balanced high performance on class and nonclass members.

Pattern recognition has been advocated as a generalized approach to the solution of data analysis problems in experimental chemistry.[2,3] In general, the aim of pattern recognition systems in chemistry is to detect or predict properties of compounds, elements, or mixtures of chemical interest based on observation of some previously defined and different set of properties or measurements.

Chemical pattern recognition studies have tended to focus on the classification of compounds into functional group classes or activity classes based on observation of spectral features, drawn primarily from mass spectra,[4-10] infrared,[11-16] or NMR spectra.[17-22] These studies have employed classifiers developed in different ways, including linear discriminant functions,[4] minimum distance measures,[18] and adaptive learning networks.[9,10] It is natural, confronted with a growing body of methods and results, that one should want to compare these methods in order to select the best method for a particular classification task and, further, that one should wish to have some measure of the anticipated level of successful performance of the chosen classifier in laboratory application.

Uhr has pointed out the paucity of controlled efforts to compare various pattern classifiers, particularly in regard to the lack of testing with similar sets of test data.[23] Since any selected set of test data might not reflect the characteristics of the "universe" of patterns to which a classifier will be applied in actual use, a serious comparative study ought to employ as large as possible a test data set and should ensure that the test sets used are identical or, at least, very similar, for the various classifiers being studied. For a classifier being designed for a restricted, well-defined set of patterns such as standard-font alphabetic characters, this problem is less difficult than for a "universe" as diverse as, say, mass spectra of organic compounds.

Recent theoretical work by Rotter and Varmuza[24] suggests an efficacious approach to the problem of choosing suitable measures of chemical classifier performance. In the present work, we present the results of applying various evaluation criteria to the performances of four sets of pattern classifiers developed and tested with a uniform set of 1252 mass spectra drawn from a file of 18 806 mass spectra.[25] This study represents the first attempt to compare mass spectral pattern classifiers on the basis of so large and so uniform a data set. We also formulate the evaluation criteria of Rotter and Varmuza in terms of experimental quantities which are easily tabulated during the testing of a pattern classifier.

Objective Evaluation of Pattern Classifiers

Rotter and Varmuza have discussed from a theoretical viewpoint a variety of possible criteria for the evaluation of binary pattern classifiers.[24] For ease of reference, we summarize below the probabilities used in their treatment. We retain the notation of those authors, in which "1" and "2" are used to denote the two possible class memberships of a pattern in a binary classification problem, and "j" and "n" are used to indicate the possible class assignments by the classifier. That is, "j" (ja) means the classifier assigns the pattern to class "1" and "n" (nein) denotes assignment of the pattern by the classifier to class "2".

Here follow the definitions of the various probabilities associated with the binary classification problem.

$p(1)$ and $p(2)$ are the a priori probabilities of membership of patterns in class 1 or class 2 in the population of interest. These probabilities thus give the composition of the test population.

$p(j)$ and $p(n)$ are the probabilities that the classifier in question will classify a pattern as belonging to class 1 [$p(j)$] or to class 2 [$p(n)$].

$p(j|1)$ and $p(n|2)$ are the class conditional probabilities or "predictive abilities", in the terminology of Rotter and Varmuza, of the classifier correctly classifying patterns from class 1 and patterns from class 2. $p(j|1)$ is the probability that patterns belonging to class 1 will be classified correctly and $p(n|2)$ is the probability that patterns belonging to class 2 will be classified correctly.

$p(1|j)$ and $p(2|n)$ are the a posteriori probabilities of membership of patterns in class 1 and class 2 following application of the classifier. $p(1|j)$ is the probability that a pattern actually belongs to class 1 given that the classifier says it does, and, similarly, $p(2|n)$ is the probability that a pattern actually belongs to class 2 given that the classifier says it does. Probabilities for *incorrect* assignments are also defined: these are $p(1|n)$ and $p(2|j)$.

$p(1,j)$, $p(2,n)$, $p(1,n)$, and $p(2,j)$ are the overall probabilities for the four possible circumstances which may exist following the application of a binary classifier. $p(1,j)$ is the joint probability that a pattern is both a member of class 1 and is classified as such by the classifier. Similar definitions apply for the other three quantities.

The probabilities mentioned above are not independent of one another, but are interconnected by a number of useful relationships. Obviously,

$$p(1) + p(2) = p(j) + p(n) = 1 \quad (1)$$

Further,

$$p(1,j) + p(1,n) = p(1) \quad (2)$$

$$p(2,j) + p(2,n) = p(2) \quad (3)$$

$$p(1,j) + p(2,j) = p(j) \quad (4)$$

$$p(1,n) + p(2,n) = p(n) \quad (5)$$

Also,

$$p(1,j) = p(1|j)p(j) = p(j|1)p(1) \quad (6)$$

$$p(2,n) = p(2|n)p(n) = p(n|2)p(2) \quad (7)$$

$$p(1,n) = p(1|n)p(n) = p(n|1)p(1) \quad (8)$$

$$p(2,j) = p(2|j)p(j) = p(j|2)p(2) \quad (9)$$

In addition to these probabilities, Rotter and Varmuza employ a quantity called the "information gain" of a classifier under consideration. The information gain $I(A,B)$ is the difference between the a priori uncertainty $H(A)$, or entropy,[26] regarding class membership and the residual uncertainty $H(A|B)$ following application of the classifier. That is,

$$I(A,B) = H(A) - H(A|B) \quad (10)$$

where

$$H(A) = -p(1) \log_2 p(1) - p(2) \log_2 p(2) \quad (11)$$

and

$$H(A|B) = p(j)H(A|j) + p(n)H(A|n) \quad (12)$$

$$H(A|j) = -p(1|j) \log_2 p(1|j) - p(2|j) \log_2 p(2|j) \quad (13)$$

$$H(A|n) = -p(1|n) \log_2 p(1|n) - p(2|n) \log_2 p(2|n) \quad (14)$$

The use of base 2 logarithms determines the units of these uncertainties to be "bits".

A more compact expression for $I(A,B)$ can be shown to be

$$I(A,B) = \sum_{i=1,2} \sum_{k=j,n} p(i,k) \log_2 \frac{p(i,k)}{p(i)p(k)} \quad (15)$$

To apply the foregoing statistical measures to the evaluation of an actual classifier, these quantities must be computed from tabulations of correct and incorrect responses. We next show how all the quantities above can be computed from four experimental measures which are easily tabulated during the testing of a classifier.

Consider a test population consisting of N_{total} patterns, of which N patterns are in category 1 and $N_{total} - N$ are in category 2. Let the number of patterns which the classifier assigns to category 1 be N^{pred} and the number which it *correctly* assigns to category 1 be N^{corr}. The four quantities N, N^{pred}, N^{corr}, and N_{total} are sufficient to compute all the relevant probabilities defined previously.

$$p(1) = N/N_{total} \quad (16)$$

$$p(2) = 1 - p(1) = (N_{total} - N)/N_{total} \quad (17)$$

$$p(j) = N^{pred}/N_{total} \quad (18)$$

$$p(n) = 1 - p(j) = (N_{total} - N^{pred})/N_{total} \quad (19)$$

$$p(j|1) = N^{corr}/N \quad (20)$$

$$p(1,j) = p(j|1)p(1) = N^{corr}/N_{total} \quad (21)$$

$$p(1|j) = p(1,j)/p(j) = N^{corr}/N^{pred} \quad (22)$$

$$p(1,n) = p(1) - p(1,j) = (N - N^{corr})/N_{total} \quad (23)$$

$$p(2,j) = p(j) - p(1,j) = (N^{pred} - N^{corr})/N_{total} \quad (24)$$

$$p(2,n) = p(2) - p(2,j) = \frac{(N_{total} - N - N^{pred} + N^{corr})}{N_{total}} \quad (25)$$

$$p(2|n) = \frac{p(2,n)}{p(n)} = \frac{(N_{total} - N - N^{pred} + N^{corr})}{(N_{total} - N^{pred})} \quad (26)$$

$$p(n|2) = \frac{p(2,n)}{p(2)} = \frac{(N_{total} - N - N^{pred} + N^{corr})}{(N_{total} - N)} \quad (27)$$

Thus, the tabulation of N, N^{pred}, N^{corr}, and N_{total} permits computation of various statistics describing classifier performance as well as the information gain $I(A,B)$ when expressed in terms of $p(1)$, $p(2)$, $p(j)$, $p(n)$, $p(1,j)$, $p(1,n)$, $p(2,j)$, and $p(2,n)$.[27]

Currently, perhaps the most commonly used measure of classifier performance in chemical applications is the overall "percent correct classifications". In terms of the four tabulations recommended above, this measure is

$$\% \text{ correct classification} = 100 \frac{N^{corr} + N_{total} - N - N^{pred} + N^{corr}}{N_{total}} \quad (28)$$

The process of evaluating a pattern classifier has two goals. First, one wishes to obtain a measure which permits objective comparison of various classifiers so that the best classifier can be selected for any particular task. Second, one wishes a measure which allows the potential user to anticipate the level of performance that a particular classifier will yield in actual use (i.e., allows *interpretation* of the predictions made). One may ask to what extent a particular measure satisfies these requirements and whether the conditions under which the measure is determined affect its validity.

Several variables in the testing process can act to interfere with an objective evaluation. If classifiers to be compared are tested on very different sets of patterns, one cannot be sure whether differences in performance are due to differences in difficulty of the classification problems or to inherent differences in the classifiers. The size and makeup of the data set used to test classifiers can influence the usefulness of an evaluation; a small test data set is unlikely to be as representative of real applications as a large set and even with a large test set, the composition of the set can significantly affect the apparent performance of a classifier as measured by certain evaluators, as we shall show later. The development of a classifier involves training with a selected set of patterns and it is naturally expected that the classifier will perform better in "recognizing" the class identities of the training patterns than in "predicting" the identities of patterns not encountered in training. Thus, the incorporation of recognition data in the evaluation of a classifier will tend to exaggerate the quality of the classifier's performance.

In the following sections, we report the application of five different evaluative measures to collections of pattern classifiers developed in four different ways. The advantages and weaknesses of each measure will be discussed. In what follows, we shall confine our attention to prediction, as opposed to recognition.

The Evaluators

Of the probabilities discussed by Rotter and Varmuza, two types are potentially most useful as performance evaluators for binary classifiers.

The class conditional probabilities $p(j|1)$ and $p(n|2)$ measure the probability that the classifier will classify correctly patterns which are drawn from class 1 on the one hand and from class 2 on the other.

The a posteriori probabilities $p(1|j)$ and $p(2|n)$ are, on first consideration, attractive candidates for performance evaluation. They give the probability that an assignment by the classifier is correct; $p(1|j)$ for assignment to class 1 and $p(2|n)$ for assignment to class 2. As we shall show, these measures,

although intuitively attractive, must be interpreted with extreme care, as they are highly dependent on the makeup of the test data set.

The overall percent correct prediction, a commonly reported statistic, summarizes both of the class conditional probabilities $p(j|1)$ and $p(n|2)$, but is weighted toward the performance of the more populous class. Thus, this overall measure can obscure certain aspects of a classifier's performance.

In order to avoid dependence on test set composition, Rotter and Varmuza propose that the information gain $I(A,B)$ be used as an objective measure of classifier performance. This quantity measures the amount by which the classifier reduces the uncertainty regarding class membership and is measured in bits. Further consideration, however, shows that the composition of the test set does indeed impose limits on $I(A,B)$ which may not be the same for all classifiers being tested. Recall that, as stated in eq 10, $I(A,B)$ is equal to the difference in pre- and post-classification entropies ($H(A)$ and $H(A|B)$). The value for $I(A,B)$ is thus zero bits in the event that the classifier adds no information; that is, the uncertainty $H(A|B)$ after use of the classifier is equal to $H(A)$, the uncertainty prior to its use. However, in examining the upper limit on $I(A,B)$, we see that the minimum *residual* uncertainty $H(A|B)$ is zero bits, in which case

$$I(A,B)_{max} = H(A) \tag{29}$$

Thus, the maximum possible information gain for a classifier is limited by the initial uncertainty, which depends on the composition of the test set employed (see eq 11). Thus, a classifier being tested on a data set composed of equal numbers of class 1 and class 2 patterns [$p(1) = p(2) = 0.5$] would have a maximum possible information gain of 1 bit, whereas a classifier tested on a data set with $p(1) = 0.1$ and $p(2) = 0.9$ would have a maximum information gain of only 0.5 bit. Comparison of two classifiers on such a basis would be misleading.

The above considerations lead us to propose a figure of merit, M,

$$M = I(A,B)/H(A) \tag{30}$$

where M is the information gain relative to the maximum possible information gain imposed by the composition of the test set. Our results indicate that M, as a measure of classifier performance, does not suffer from the defect of being test set dependent.

The Classifiers

In the present work, we have applied each of the five types of evaluator (class conditional probabilities, a posteriori probabilities, percent correct prediction, information gain, and figure of merit) to a set of 44 classifiers. These classifiers were developed to assign organic compounds to functional group classes based on their low resolution mass spectra. Functional group categories (11) are represented and are listed in Table I. Classifiers for the 11 categories were developed in four different ways; the methods are described in detail elsewhere.[28,29] In each case, the classifier set described here is the end product of an effort aimed at producing the best classifiers of that particular type.

Three of the four 11-classifier sets were developed as strictly binary classifiers. Set 1 was developed by the linear learning machine method, based on error correction feedback;[30] the resulting weight vector classifiers utilize from 20 to 45 features in the patterns to which they are applied. With these few features, the patterns in the training sets were not linearly separable. Set 2 was developed by the sequential simplex optimization procedure[8,28] using the weight vectors in set 1 as the starting point. These classifiers utilize the same numbers of features as those in set 1. Set 3 was developed by the linear learning machine method using more features to improve linear separability of the training set patterns. These classifiers employ 60 features each, which results in linear separability of six of the 11 training sets.

Table I. Composition of Data Set from Which Training and Test Sets Were Drawn

Category	Number of spectra
1 C_6H_5R (R = straight chain)	249
2 RC(=O)R′ (R′ may be H)	96
3 ROR′	103
4 ROH	185
5 C_6H_5OH	84
6 RC(=)OH	51
7 RSR′ (R′ may be H)	135
8 RC(=O)OR′	125
9 RNR′R″ (R′ and R″ may be H)	131
10 $RC(=O)NH_2$	56
11 RC≡N	37

The classifiers comprising set 4 are based on adaptive digital learning networks and thus are not simple binary classifiers. In the digital learning network (DLN) method, the classifiers are templates which are used as a group to effect multicategory classification instead of being employed individually as with the binary classifiers.[9,10] In this work, 128 training compounds were used to develop the set of 11 DLN classifiers. Although one can still tabulate quantities N, N^{pred}, and N^{corr} for each of the 11 categories and thus evaluate the multicategory classifier as if it were an adjustable binary classifier, certain distinctions must be borne in mind. The multicategory classifier assigns each pattern to just one class out of the 11, making a single unambiguous (though not necessarily correct) classification. The same pattern, submitted to an array of 11 binary classifiers, might be classified, say, as an ether by one binary classifier and as an amine by a different binary classifier. Of course, for polyfunctional compounds, assignment to more than one category might be appropriate, but the prospect of misclassification or ambiguous classification is increased by the possibility of assignment to more than one category. Also, the collection of binary classifiers is capable of assigning a pattern to none of the categories in question, but the DLN multicategory classifier is designed to always make a positive class assignment. Thus, assignments by a group of binary classifiers are less restricted than those of the DLN classifier, but the former are also subject to uncertainties not characteristic of the latter. A choice between these two types of classifier might depend, in part, on the desired application, and a comparison of their performance must be made on some common basis. In the present study, by treating the DLN classifier as if it were a variable topic binary classifier, we compare its performance on each particular topic or category with that of the corresponding binary classifier. Alternatively, one could build a multicategory classifier from a parallel array of binary classifiers[31] and compare the resulting multicategory assignments with those of the DLN system.

The Data Set

The results reported here and are based on a rather larger data set than has been employed in most chemical pattern recognition studies. From a file of 18 806 mass spectra,[25] a set of 1252 spectra was selected, distributed among the 11 functional group categories as shown in Table I. The category definitions do not overlap, except in the case of categories 1 and 5 (phenyl compounds and phenols, respectively). With this exception, no compound in the data set contains functional groups from more than one category, but some of the com-

Table II. Performance Measures for Set 1, Reduced Feature Linear Learning Machine Weight Vectors

Category[a]	Number of features	$p(1)$[b]	% correct pred/100	$p(j\|1)$	$p(n\|2)$	$p(1\|j)$	$p(2\|n)$	$I(A,B)$, bits	M
1	20	0.14	0.91	0.65	0.96	0.71	0.94	0.20	0.34
2	35	0.04	0.88	0.75	0.89	0.21	0.99	0.06	0.25
3	45	0.04	0.88	0.40	0.90	0.15	0.97	0.02	0.07
4	30	0.08	0.87	0.60	0.90	0.34	0.96	0.07	0.18
5	20	0.04	0.87	0.95	0.87	0.22	0.998	0.09	0.38
6	25	0.02	0.86	0.71	0.87	0.10	0.99	0.02	0.18
7	20	0.05	0.86	0.93	0.86	0.26	0.995	0.11	0.36
8	25	0.04	0.83	0.67	0.84	0.15	0.98	0.04	0.14
9	20	0.06	0.78	0.61	0.79	0.15	0.97	0.03	0.09
10	25	0.02	0.82	0.81	0.82	0.10	0.99	0.03	0.19
11	25	0.02	0.94	0.82	0.94	0.20	0.997	0.04	0.36

Table III. Performance Measures for Set 2, Simplex Weight Vectors, Using Same Features as Set 1

Category	$p(1)$[a]	% correct pred/100	$p(j\|1)$	$p(n\|2)$	$p(1\|j)$	$p(2\|n)$	$I(A,B)$, bits	M
1	0.14	0.95	0.92	0.95	0.77	0.99	0.37	0.63
2	0.04	0.89	0.78	0.90	0.23	0.99	0.06	0.27
3	0.04	0.83	0.77	0.84	0.17	0.99	0.05	0.20
4	0.08	0.80	0.84	0.79	0.26	0.98	0.09	0.23
5	0.04	0.95	0.92	0.95	0.42	0.997	0.12	0.53
6	0.02	0.91	0.71	0.91	0.15	0.99	0.03	0.23
7	0.05	0.95	0.93	0.95	0.49	0.996	0.16	0.55
8	0.04	0.87	0.76	0.87	0.21	0.99	0.06	0.23
9	0.06	0.95	0.92	0.95	0.53	0.99	0.18	0.55
10	0.02	0.90	0.65	0.91	0.16	0.99	0.03	0.20
11	0.02	0.96	0.82	0.96	0.27	0.997	0.05	0.42

[a] For each category, class 1 refers to members of that category and class 2 refers to nonmembers.

pounds contain more than one functional group of a particular type.

Preprocessing of the spectra for the three linear discriminant classifier sets consisted of discarding peaks with intensity less than 1% of the base peak and scaling the remaining peaks by taking the square root of their intensities. m/e values up to 166 were utilized. The nature of the DLN classifier required peak/no peak binary coding of the spectra at 256 m/e values from m/e = 1 to 256.

Because of the differences in the methods, the training sets used in developing the various classifiers were not identical. The linear discriminant classifiers were trained using 200 of the 1252 spectra, leaving 1052 patterns for testing. Necessarily, different training sets were used to train the classifiers for the various categories, so that the test sets for the various functional group classifiers, though similar, were not strictly identical. However, identical test sets were employed for each of the three groups of linear discriminant classifiers. The DLN classifier required 128 training patterns, leaving 1124 test spectra. Here, since the whole classifier is trained as a unit, the test set was uniform for all categories.

Results and Discussion

Tables II–V summarize the evaluation results for the four sets of classifiers analyzed in this work. Scrutiny of these values reveals several important characteristics of the various evaluators.

The limitations of percent correct prediction as an evaluator become apparent upon consideration, for example, of classifier 1 as developed by the linear learning machine method with differing numbers of features (Tables II and IV). The percent correct predictions for these classifiers are 91 and 92%, respectively. However, the actual performances are rather divergent, as measured by the class conditional probabilities. The higher dimensioned weight vector (Table IV) performs with equal ability (92%) in classifying both class members and nonmembers, while the weight vector with fewer components (Table II) correctly classifies only 65% of the class members. The percent correct prediction is actually the weighted average of $p(j|1)$ and $p(n|2)$ and thus will always be nearer the class conditional probability for the more populous class. For a test involving several categories, class 2 (nonmembers) will generally be far more populous than class 1 (members) for any particular category of patterns. Another example of this difficulty is seen with classifier 3 as initially developed by the linear learning machine method (Table II) and as subsequently optimized by the simplex procedure (Table III). Although the performance of the classifier on class members [$p(j|1)$] is dramatically improved by the optimization, the overall percent correct prediction figure is lower for the optimized weight vector as a result of a drop in performance on the more numerous nonclass members.

Use of the a posteriori probabilities $p(1|j)$ and $p(2|n)$ as evaluators is also likely to cause difficulty. According to these measures, classifier 1 as developed by the DLN method (Table V) has a likelihood of 78% of being correct if it predicts that a compound contains phenyl and a 99.5% likelihood of being correct if it predicts that the compound does not contain this functional group; these values seem reasonable enough. However, for classifier 11 in that same set, a posteriori probabilities of 100 and 99% are observed for class members and nonmembers, respectively. These values are placed in perspective by examination of the values for $p(j|1)$ and $p(n|2)$ for classifier 11, which show that the classifier correctly assigns only 58% of the compounds actually containing -C≡N. We see here a "conservative" classifier, which tends to classify patterns as nonmembers. All those patterns which it does classify as containing -C≡N do indeed contain that moiety (hence, $p(1|j)$ = 1.00), but this performance is at the expense of misclassifying 42% of the nitrile spectra presented to the classifier. Depending on the application, this kind of trade-off may or may not be acceptable.

Table IV. Performance Measures for Set 3, 60 Feature Linear Learning Machine Weight Vectors

Category	$p(1)^a$	% correct pred/100	$p(j\|1)$	$p(n\|2)$	$p(1\|j)$	$p(2\|n)$	$I(A,B)$, bits	M
1	0.14	0.92	0.92	0.92	0.65	0.99	0.32	0.53
2	0.04	0.95	0.87	0.96	0.45	0.99	0.12	0.50
3	0.04	0.84	0.77	0.85	0.18	0.99	0.05	0.21
4	0.08	0.88	0.82	0.89	0.40	0.98	0.14	0.34
5	0.04	0.92	0.90	0.92	0.31	0.996	0.10	0.42
6	0.02	0.91	0.71	0.92	0.15	0.99	0.03	0.23
7	0.05	0.93	0.91	0.93	0.43	0.99	0.14	0.49
8	0.04	0.91	0.51	0.93	0.24	0.98	0.04	0.15
9	0.06	0.94	0.95	0.94	0.51	0.997	0.18	0.57
10	0.02	0.94	0.73	0.95	0.26	0.99	0.05	0.32
11	0.02	0.97	0.88	0.97	0.31	0.998	0.06	0.49

[a] For each category, class 1 refers to members of that category and class 2 refers to nonmembers.

Table V. Performance Measures for Digital Learning Network Classifiers

Category	$p(1)^a$	% correct pred/100	$p(j\|1)$	$p(n\|2)$	$p(1\|j)$	$p(2\|n)$	$I(A,B)$, bits	M
1	0.20	0.94	0.98	0.93	0.78	0.995	0.50	0.69
2	0.08	0.95	0.55	0.98	0.68	0.96	0.12	0.31
3	0.08	0.89	0.40	0.93	0.35	0.95	0.05	0.11
4	0.15	0.86	0.66	0.90	0.53	0.94	0.15	0.24
5	0.07	0.96	0.45	0.99	0.83	0.96	0.10	0.29
6	0.04	0.93	0.20	0.96	0.17	0.97	0.01	0.04
7	0.11	0.98	0.95	0.98	0.87	0.99	0.38	0.77
8	0.10	0.90	0.35	0.97	0.54	0.93	0.07	0.14
9	0.10	0.93	0.54	0.97	0.68	0.95	0.14	0.28
10	0.04	0.95	0.48	0.97	0.47	0.98	0.06	0.23
11	0.03	0.99	0.58	1.00	1.00	0.99	0.09	0.50

[a] For each category, class 1 refers to members of that category and class 2 refers to nonmembers.

The values of the a posteriori probabilities are highly dependent on the composition of the data set used to test the classifiers. The more populous is class 1 (members) for any category, the more likely is an assignment to class 1 to be correct. Thus, comparison of the category 4 classifiers as developed by the linear learning machine (Table IV) and by the DLN method (Table V) shows that the linear discriminant classifier has higher performance [$p(j|1)$] for class members than does the DLN classifier. However, the a posteriori probability is higher for the DLN classifier. This seeming contradiction arises from the higher a priori probability of class 1 membership in the DLN test set [$p(1) = 0.15$ compared with $p(1) = 0.08$ for the linear learning machine test set]. For the same reason, the values of $p(2|n)$ are uniformly high for all 44 classifiers, with no values falling below 0.93. Thus, the a posteriori probabilities are useful for comparative evaluations *only* if the test set a priori probabilities are the same for all classifiers being compared. Further, the a posteriori probabilities are indicative of predictive reliability in an absolute sense only if the composition of the test set is known to be representative of the universe of patterns to which the classifier will be applied in actual use.

A perfect classifier would have $p(j|1) = p(n|2) = p(1|j) = p(2|n) = 1.00$. Of course, automatic classifiers of this quality are rarely, if ever, available and, for most problems, there is reason to expect that such performance is not achievable. In a less-than-perfect classifier, one would usually (though not always) seek balanced performance on class members and nonmembers at the highest achievable level. One would like to have, therefore, an objective measure of classifier performance with which to make comparisons among classifiers. The information gain $I(A,B)$ discussed by Rotter and Varmuza was intended by those authors to meet this need.[24] If one were selecting, for example, a classifier for the presence of phenyl from among those in Tables III, IV, and V, a choice based on the class conditional probabilities would not be clear cut. The information gain $I(A,B)$, however, indicates that the DLN classifier supplies more information than do the two linear discriminant classifiers. The figure of merit M gives the same ordering of these three classifiers.

In the case of classifier 2, a limitation becomes apparent regarding the use of the information gain as an evaluator. Here, the linear learning machine derived weight vector (Table IV) shows the most balanced high performance of the three class 2 classifiers (Tables III, IV, and V), but the information gain of that classifier is no larger than that of the DLN classifier. Because of differences in the a priori probabilities $p(1)$ and $p(2)$ in the test sets, the information gain for the DLN classifier is not strictly comparable with that of the other two classifiers. The same situation is observed with classifier 5 in Tables IV and V. This difficulty is met by the figure of merit, M, which is the ratio of the information gain to the maximum possible information gain imposed by the test conditions, as discussed above. Examination of the M values of the various classifiers shows that high M values are indicative of balanced high performance on both class members and nonmembers.[32]

Summary

Several measures of performance for automatic binary pattern classifiers have been applied to classifiers which were developed by various methods and tested using a large collection of mass spectra. The different evaluators do not always agree in their comparisons of classifiers. Of the various measures, the class conditional probabilities $p(j|1)$ and $p(n|2)$ and the figure of merit M, derived from the information gain, are well suited to selecting classifiers with balanced high performance on class members and nonmembers. The a posteriori probabilities can be useful in a restricted way for reporting the reliability of a classifier's prediction, but these values are very sensitive to the composition of the test set.

We have refrained from value judgments such as "good" or "better than" in discussing both the classifiers and the evaluators. The actual characteristics which make a classifier suitable for a particular application may depend on the application itself, as, for example, when the penalty associated with misclassification of a class member is not the same as for misclassification of a nonmember.[24] Although the percent correct prediction should probably be abandoned as a measure of the performance of binary classifiers, the other measures discussed can be useful in developing a total picture of relative and absolute classifier performance.

Acknowledgment. Support of this research under Grant MPS-74-01249 from the National Science Foundation is gratefully acknowledged. We also gratefully acknowledge the University of Nebraska Research Council, which provided partial support for purchase of the computer-readable mass spectra data set.

References and Notes

(1) L. J. Soltzberg is a visiting Associate Professor during the 1975–1976 academic year from Simmons College, Boston, Mass.
(2) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **94**, 5632 (1972).
(3) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **95**, 686 (1973).
(4) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 690 (1969).
(5) B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 695 (1969).
(6) L. Pietrantonio and P. C. Jurs, *Pattern Recognition*, **4**, 391 1972).
(7) B. R. Kowalski and C. F. Bender, *Anal. Chem.*, **45**, 2334 (1973).
(8) G. L. Ritter, S. R. Lowry, C. L. Wilkins, and T. L. Isenhour, *Anal. Chem.*, **47**, 1951 (1975).
(9) T. J. Stonham, I. Aleksander, M. Camp, W. T. Pike, and M. A. Shaw, *Anal. Chem.*, **47**, 1817 (1975).
(10) T. J. Stonham and M. A. Shaw, *Pattern Recognition*, **7**, 235 (1975).
(11) B. R. Kowalski, P. C. Jurs, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **41**, 1949 (1969).
(12) S. R. Lowry, H. B. Woodruff, G. L. Ritter, and T. L. Isenhour, *Anal. Chem.*, **47**, 1126 (1975).
(13) H. B. Woodruff, S. R. Lowry, and T. L. Isenhour, *Appl. Spectrosc.*, **29**, 226 (1975).
(14) R. W. Liddell and P. C. Jurs, *Appl. Spectrosc.*, **27**, 371 (1973).
(15) D. R. Preuss and P. C. Jurs, *Anal. Chem.*, **46**, 520 (1974).
(16) R. W. Liddell and P. C. Jurs, *Anal. Chem.*, **46**, 2126 (1974).
(17) B. R. Kowalski and C. A. Reilly, *J. Phys. Chem.*, **75**, 1402 (1971).
(18) B. R. Kowalski and C. F. Bender, *Anal. Chem.*, **44**, 1405 (1972).
(19) C. L. Wilkins, R. C. Williams, T. R. Brunner, and P. C. McCombie, *J. Am. Chem. Soc.*, **96**, 4182 (1974).
(20) T. R. Brunner, R. C. Williams, C. L. Wilkins, and P. J. McCombie, *Anal. Chem.*, **46**, 1798 (1974).
(21) T. R. Brunner, C. L. Wilkins, R. C. Williams, and P. J. McCombie, *Anal. Chem.*, **47**, 662 (1975).
(22) C. L. Wilkins and T. L. Isenhour, *Anal. Chem.*, **47**, 1849 (1975).
(23) L. Uhr, "Pattern Recognition, Learning, and Thought", Prentice-Hall, Englewood Cliffs, N.J., 1973, p 26.
(24) H. Rotter and K. Varmuza, *Org. Mass Spectrom.*, **10**, 874 (1975).
(25) E. Stenhagen, S. Abrahamssen, and F. W. McLafferty, "The Registry of Mass Spectral Data", Wiley-Interscience, New York, N.Y., 1974.
(26) C. H. Chen, "Statistical Pattern Recognition", Hayden Book Co, Inc., Rochelle Park, N.J., 1973, p 54.
(27) The information gain I(A,B) could, of course, be expressed directly in terms of N, N^{pred}, N^{corr}, and N_{total}, but the equation is cumbersome. In computing I(A,B) from $p(i)$, $p(k)$, and $p(i,k)$, the situation can arise where $p(i,k) = 0$. The corresponding term $p(i,k) \log_2 p(i,k)/p(i)p(k)$ in the expression for I(A,B) is then indeterminate. However, by L Hospital's Rule:

$$\lim_{x \to 0} x \log cx = \lim_{x \to 0} \frac{c' \ln cx}{1/x} = \lim_{x \to 0} \frac{\frac{d}{dx} c' \ln cx}{\frac{d}{dx} \frac{1}{x}} = \lim_{x \to 0} c'' x = 0$$

Thus, such terms contribute nothing to the summation.
(28) T. F. Lam, C. L. Wilkins, T. R. Brunner, L. J. Soltzberg, and S. L. Kaberline, *Anal. Chem.*, in press.
(29) L. J. Soltzberg, C. L. Wilkins, S. L. Kaberline, T. F. Lam, and T. R. Brunner, *J. Am. Chem. Soc.*, following paper in this issue.
(30) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, N.Y., 1965.
(31) W. L. Felty and P. C. Jurs, *Anal. Chem.*, **45**, 885 (1973).
(32) DLN classifier 11 (Table V) is exceptional because of its perfect performance on nonmembers of the class and thus has a high figure of merit in spite of its low performance $p(j|1)$ on class members.

Evaluation and Comparison of Pattern Classifiers for Chemical Applications: Adaptive Digital Learning Networks and Linear Discriminants

Leonard J. Soltzberg,[1] Charles L. Wilkins,* Steven L. Kaberline, T. F. Lam, and T. L. Brunner

Contribution from the Department of Chemistry, University of Nebraska–Lincoln, Lincoln, Nebraska 68588. *Received April 9, 1976*

Abstract: Recent research on the use of adaptive networks of digital learning elements for chemical pattern recognition has stressed the high performance of such classifiers and their applicability to linearly inseparable data. In the present work, we apply a new performance measure, the figure of merit, and a large set of test data in a rigorous evaluation of the performance of digital learning networks. The results herein reported show that, when confronted with a large data set selected without particular consideration of the peculiarities of the network, the digital learning network continues to give good performance, although this performance is substantially below the levels previously reported. A comparison of the performance of the digital learning network classifiers with that of a set of linear discriminant functions indicates similar levels of performance for the two types of classifier.

The formal problem of pattern recognition can be approached from the viewpoint of various paradigms.[2] Those models based on the establishment of templates of various sorts resemble what one might expect in a biological pattern recognition apparatus. At the same time, abstract or ad hoc algorithms based on purely mathematical notions of pattern resemblance can also function quite successfully.

Pattern classifiers applied to chemical data have generally been of the abstract or ad hoc type. Specifically, most chemical applications have employed linear discriminant functions computed by error correction feedback[3] or by sequential simplex methods.[4] Some use has been made of the k nearest neighbor algorithm[5] as well as certain other abstract statistical methods.[6]

Recently, Stonham et al.[7,8] have described the machine recognition of chemical classes from a limited group of mass