(9) Zivkovič, T.; Trinajstič, N.; Randič, M. "On Conjugated Molecules with Identical Topological Spectra" *Mol. Phys.* **1975**, *30*, 517–533.

(10) Randič, M.; Trinajstič, N.; Zivkovič, T. "Molecular Graphs Having Identical Spectra" *J. Chem. Soc., Faraday Trans. 2* **1976**, *72*, 244–256.

(11) Heilbronner, E.; Jones, T. B. "Spectral Differences Between ' Isospectral ' Molecules" *J. Am. Chem. Soc.*, **1976**, *100*, 6506–6507. A preprint of this article was kindly supplied by Professor Heilbronner.

(12) (a) Herndon, W. C.; Tao, F.-K., unpublished work. (b) Herndon, W. C.; Ellzey, M. L., Jr., papers submitted for publication, 1972–1974, and rejected.

(13) Gutman, I.; Milun, M.; Trinajstič, N. "Topological Definition of Delocalisation Energy" *MATCH*, **1975**, *1*, 171–175.

(14) Aihara, J. "A New Definition of Dewar-Type Resonance Energies" *J. Am. Chem. Soc.* **1976**, *98*, 2750–2758.

(15) Aihara, J. "Resonance Energies of Benzenoid Hydrocarbons" *J. Am. Chem. Soc.* **1977**, *99*, 2048–2053.

(16) Gutman, I.; Milun, M.; Trinajstič, N. "Graph Theory and Molecular Orbitals. 19. Nonparametric Resonance Energies of Arbitrary Conjugated Systems" *J. Am. Chem. Soc.* **1977**, *99*, 1692–1704.

(17) Trinajstič, N. "New Developments in Hückel Theory" *Int. J. Quantum Chem.: Quantum Chem. Symp. II* **1977**, 469–477.

(18) Hosoya, H. "Graphical Enumeration of the Coefficients of the Secular Polynomials of the Hückel Molecular Orbitals" *Theor. Chim. Acta* **1972**, *25*, 215–222.

(19) Graovac, A.; Gutman, I.; Trinajstič, N.; Zikovič, T. "Graph Theory and Molecular Orbitals. Application of Sachs Theorem", *Theor. Chim. Acta* **1972**, *26*, 67–78.

(20) Aihara, J. "General Rules for Constructing Hückel Molecular Orbital Characteristic Polynomials" *J. Am. Chem. Soc.* **1976**, *98*, 6840–6844.

(21) Hess, B. A., Jr.; Schaad, L. J.; Agranat, I. "The Aromaticity of Annulenoannulenes", *J. Am. Chem. Soc.* **1978**, *100*, 5268–5271.

(22) Herndon, W. C.; Parkânyi, C. "Perturbation-Graph Theory. I. Resonance Energies of Heteroannulene π Systems" *Tetrahedron* **1978**, *34*, 3419–3425.

(23) Herndon, W. C. "Enumeration of Resonance Structures" *Tetrahedron* **1973**, *29*, 3–12.

(24) Herndon, W. C.; Ellzey, M. L., Jr. "Closed-Shell Biradical Structures" *Tetrahedron Lett.* **1974**, 1399–1402.

(25) Herndon, W. C. "Resonance Theory and the Enumeration of Kekule Structures" *J. Chem. Educ.* **1974**, *51*, 10–15.

(26) Herndon, W. C.; Ellzey, M. L., Jr.; Raghuveer, K. S. "Topological Orbitals, Graph Theory, and Ionization Potentials of Saturated Hydrocarbons" *J. Am. Chem. Soc.* **1978**, *100*, 2645–2650.

(27) Herndon, W. C. "Resonance Energies of Benzene and Heterobenzenes from Photoelectron Spectra" *Tetrahedron Lett.*, in press.

(28) Herndon, W. C. "Perturbation-Graph Theory. III. Resonance Energies From Photoelectron Spectra" *Pure Appl. Chem.*, in press.

(29) Streitwieser, A., Jr. "Molecular Orbital Theory for Organic Chemists"; Wiley: New York, 1961.

(30) Formerly, some subgraphs of the type considered were called "Sachs graphs" referring to work of H. Sachs, cited in ref 18–20.

(31) Mallion, R. B.; Schwenk, A. J.; Trinajstič, N. "A Graphical Study of Heteroconjugated Molecules" *Croat. Chem. Acta* **1974**, *46*, 171–182.

(32) Mallion, R. B.; Trinajstič, N.; Schwenk, A. J. "Graph Theory in Chemistry – Generalization of Sach's Formula" *Z. Naturforsch.* **1974**, *29a*, 1481–1484.

(33) Graovac, A.; Polansky, O. E.; Trinajstič, N.; Tyutyulkov, N. "Graph Theory in Chemistry. II. Graph Theoretical Description of Heteroconjugated Molecules" *Z. Naturforsch.* **1975**, *30a*, 1696–1699.

(34) Rigby, M. J.; Mallion, R. B.; Day, A. C. "Comment on a Graph-Theoretical Description of Heteroconjugated Molecules" *Chem. Phys. Lett.* **1977**, *51*, 178–182.

(35) Hall, L. H. "Group Theory and Symmetry in Chemistry"; McGraw-Hill: New York, 1969; Chapters 5 and 6.

(36) Heilbronner, E. "Ein graphisches Verfahren zur Faktorisierung der Sakulardeterminante aromatischer Ringsysteme in Rahamen der LCAO-MO-Theorie" *Helv. Chim. Acta* **1954**, *37*, 913–921.

(37) King, R. B. "Symmetry Factoring of the Characteristic Equations of Graphs Corresponding to Polyhedra" *Theor. Chim. Acta* **1977**, *44*, 223–243.

(38) McClelland, B. J. "Graphical Method for Factorizing Secular Determinants of Hückel Molecular Orbital Theory" *J. Chem. Soc., Faraday Trans 2* **1974**, 1453–1456.

(39) Texas Instrument calculators TI58 and TI59 are preprogrammed to solve algebraic equations using this method. A Fortran computer program to solve these equations is available from the authors.

(40) Heilbronner, E. "Das Kompositions-Prinzip: Eine anschauliche Methode zur elektronen-theoretischen Behandlung nicht oder niedrig symmetrischer Molekeln im Rahmen der MO-Theorie" *Helv. Chim. Acta* **1953**, *36*, 170–188.

(41) Dewar, M. J. S.; Gleicher, G. J. "Ground States of Conjugated Molecules. II. Allowance for Molecular Geometry" *J. Am. Chem. Soc.* **1965**, *87*, 685–692.

(42) Dewar, M. J. S.; de Llano, C. "Ground States of Conjugated Molecules. XI. Improved Treatment of Hydrocarbons" *J. Am. Chem. Soc.* **1969**, *91*, 789–795.

(43) Herndon, W. C. "Resonance Energies of Aromatic Hydrocarbons. A Quantitative Test of Resonance Theory" *J. Am. Chem. Soc.* **1973**, *95*, 2404–2406.

(44) Herndon, W. C.; M. L. Ellzey, Jr. "Resonance Theory. V. Resonance Energies of Benzenoid and Nonbenzenoid π Systems" *J. Am. Chem. Soc.* **1974**, *96*, 6631–6642.

(45) Hess, B. A., Jr.; Schaad, L. J. "Hückel Molecular Orbital π Resonance Energies. A New Approach" *J. Am. Chem. Soc.* **1971**, *93*, 305–310.

(46) Hess, B. A., Jr.; Schaad, L. J. "Hückel Molecular Orbital π Resonance Energies. The Benzenoid Hydrocarbons" *J. Am. Chem. Soc.* **1971**, *93*, 2413–2416.

(47) Hess, B. A., Jr.; Schaad, L. J., Hückel Molecular Orbital π Resonance Energies. The Nonalternant Hydrocarbons" *J. Org. Chem.* **1971**, *36*, 3418–3423.

(48) Wheland, G. W. "Resonance in Organic Chemistry"; Wiley: New York, 1955; pp 75–121.

The Probability of Dichotomization by a Binary Linear Classifier as a Function of Training Set Population Distribution

ERIK K. WHALEN-PEDERSEN and PETER C. JURS*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received June 25, 1979

The dimensionality (number of descriptors per pattern) of nonparametric binary pattern classifiers has been the topic of several papers appearing in the chemical literature.[1-5] These papers have discussed the relationship between the ratio of the number of patterns being classified, N, and the number of descriptors per pattern, d, and the probability of dichotomization, P, for a binary linear classifier. The theoretical relationship

$$P = D(N,d)/2^N \qquad (1)$$

$$D(N,d) = 2\sum_{k=0}^{d} \frac{(N-1)!}{(N-1-k)!k!} \qquad (2)$$

is derived in many pattern recognition texts (e.g., refs 6 and 7) and has been discussed in the literature and will not be dealt with in detail here. The only constraint used in the derivation is that the patterns be well distributed. The derivation makes no mention of the way the patterns forming the two classes are divided between the two classes. This relationship is commonly used to determine requirements upon the ratio of N to d for development of linear classifiers such that the probability of complete separation between the two classes due to mathematical artifacts is kept very small.

Several authors have proposed rules for practical application of the theoretical probability. Bender et al.[5] recommended using three times the number of patterns as descriptors per pattern for each class, and Gray[3] proposed that a threshold criterion could be imposed on the probability function. Both of these guidelines appeal directly to the theoretical probability of dichotomization.

An important and relevant consideration has been omitted from all previous treatments. The fraction of the patterns comprising each of the two classes will affect the probability

0095-2338/79/1619-0264$01.00/0 © 1979 American Chemical Society

Table I. Results of Training 60 Patterns with 30 Descriptors Each for 20 Training Sets at Each Set of Class Populations: Uniform Distribution

no. in yes class	no. in no class	fraction converging	av predictive ability
30	30	0.550	50.4
20	40	0.450	50.0
17	43	0.600	48.8
16	44	0.775	48.0
15	45	0.800	47.6
14	46	0.850	48.2
46	14	0.900	47.4
13	47	0.800	47.0
12	48	0.750	47.4
10	50	0.925	45.0

of dichotomization. Clearly, if all the patterns were members of one class, then the probability of dichotomization would be unity. If the patterns were equally divided between the two classes, the probability would be minimized. It is the probability for equal class populations that is calculated from the probability equation (eq 2). For the development of linear classifers with the same overall number of patterns but different distributions between the two classes, there would be corresponding deviations in the values for the probability of dichotomization. The effective probability for any specific distribution would be expected to be intermediate between the theoretical value computed from eq 2 and the extreme case where all patterns were members of one class and $P = 1$. While this qualitative behavior of the probability of dichotomization as a function of division of the patterns between classes is understood, the problem is not amenable to analytic solution. Therefore, we have investigated the nature of the probability of dichotomization using Monte Carlo methods.

For the investigation a data set comprised of random numbers was generated. A total of 30 descriptors were developed for each of 250 patterns. Both Gaussian and uniform distributions were employed. The descriptors were autoscaled to have equal ranges and thus equal weights. From each of the uniform and Gaussian distributions 125 patterns were arbitrarily assigned to a "Yes" class and the remaining 125 patterns assigned to a "No" class. Training sets were randomly selected from each of the uniform and Gaussian distributions with specified numbers of Yes and No class members. The patterns not selected for the training set were put in the corresponding prediction set for that training set and were used to determine the predictive ability of the linear classifier.

The linear classifier used was a linear learning machine using an error correction feedback routine. The routine was constrained to have a maximum number of feedbacks with which to develop a weight vector. If the routine found linear separability it was said to have converged. If the routine failed to converge, the final weight vector was found by an averaging procedure over an additional number of feedbacks. During training, patterns were presented in a nonrandom sequence and the weight vectors were initialized between training sets. Predictive abilities were assessed by classifying members of the prediction set—all of which were not used during training. The results of the training and predictions are given in Tables I–IV.

Table I shows the results of training 60 patterns with 30 descriptors each from the data set formed with uniformly distributed random numbers. The value of λ [defined as $\lambda = N/(d + 1)$] for this set of trainings is 1.935 and the corresponding probability of dichotomization (P) is 0.602. With the total number of patterns held constant at 60 and the population of the two classes varied, deviations from the theoretical are observed as shown in Table I.

The fraction of training sets converging will approximately represent the probability of dichotomization. For equal class populations, the predicted probability is observed. As the population of the two classes become increasingly unequal, the fraction of convergences increases and approaches unity (which would be the case if all patterns were members of one class). The data are random numbers and the separations obtained are mathematical artifacts as indicated by the predictive abilities which are approximately random.

Table II. Results of Training 90 Patterns with 30 Descriptors Each for 40 Training Sets at Each Set of Class Populations: Uniform Distribution

no. in yes class	no. in no class	fraction converging	av predictive ability		
				yes class	no class
45	45	0.0	48.7	48.8	48.6
30	60	0.0	47.3	38.7	59.8
20	70	0.0	44.6	34.0	64.9
19	71	0.050	45.0	33.9	66.9
18	72	0.100	44.8	33.6	67.5
17	73	0.075	42.9	33.1	64.9
16	74	0.150	41.2	29.0	64.7
15	75	0.225	42.1	29.2	68.5
14	76	0.250	41.5	29.6	68.4
13	77	0.325	39.9	25.4	72.3
12	78	0.475	39.3	24.8	74.3

Table III. Results of Training 90 Patterns with 30 Descriptors Each for 40 Training Sets at Each Set of Class Populations: Gaussian Distribution

no. in yes class	no. in no class	fraction converging	av predictive ability		
			yes class		no class
45	45	0.0	50.4	49.6	51.3
30	60	0.0	49.4	43.2	58.5
22	68	0.025	44.4	30.8	68.9
21	69	0.125	44.2	30.6	69.4
20	70	0.200	44.4	30.4	70.9
19	71	0.125	43.8	29.5	71.4
18	72	0.175	43.3	28.3	71.9
17	73	0.300	42.5	27.7	72.5
16	74	0.300	42.3	27.6	73.8
15	75	0.350	41.1	28.6	75.6
14	76	0.475	40.8	25.0	76.6

A more realistic situation is illustrated in Table II. Here, training sets of 90 patterns with 30 descriptors each from the uniform distribution were trained. This corresponds to a $\lambda = 2.903$ and $P = 0.001$. As the population becomes increasingly biased toward one class, the fraction of convergences increased.

The predictive abilities shown in Table II illustrate another important effect of unequal class populations. As the population of a class was increased, its predictive ability increased as well. A large population in a class is certainly more representative of that class than a smaller population would be. This explains how the linear classifier could find a better than random separation for the more populous class and a less than random separation for the less populous class from random data where no genuine difference exists.

In Table III are the results of training 90 patterns with 30 descriptors each for the data set consisting of Gaussian distributed random numbers. The observed behavior is the same as that for the uniformly distributed data set. The inflection points for deviation from the theoretical are not identical but are very similar. The differences may be attributed to the differences in the natures of the uniform and Gaussian distributions.

Table IV gives the results of training linear classifiers with 30 descriptors per pattern and 30 patterns in the less populous class and increasing numbers of patterns in the more populous class. This illustrates that even with four times as many patterns in the more populous class as in the less populous class, when the population of the less populous class does not fall

Table IV. Results of Training with 30 Descriptors and Keeping the Number of Patterns in the Less Populous Class Equal to the Number of Descriptors: Gaussian Distribution[a]

no. in yes class	no. in no class	fraction converging	av predictive ability		
				yes class	no class
30	30	0.550	48.8	48.6	48.8
30	40	0.225	49.6	44.8	52.2
30	50	0.100	49.1	41.2	49.1
30	60	0.0	49.4	43.2	58.5
30	70	0.0	45.5	34.4	64.8
30	80	0.0	44.1	33.7	66.3
30	90	0.0	40.9	31.3	67.1
30	100	0.0	38.8	30.2	71.2
30	110	0.0	34.4	29.6	69.6
30	120	0.0	30.2	28.0	71.0

[a] There were 40 training sets for each set of class populations.

below the number of descriptors, the probability of dichotomization does not detectably deviate from the theoretical. It also illustrates that the predictive ability of the more populous class increases with population and the predictive ability of the less populous class undergoes a corresponding decrease even though its populations remains constant.

In the preceding experiment, it has been shown that in addition to the sample size and dimensionality, class populations affect the probability of dichotomization. In addition to this effect, the unequal distribution of patterns between classes increases the predictive ability of the more populous class and decreases the predictive ability of the less populous class. With careful consideration, these effects are logical.

When the decision function is developed, the class which has the most members representing it will be more accurately depicted. With a finite number of class members, the greater the fraction of the total patterns used, the more accurately the class mean and distribution are characterized. The converse is true of small fractions. Thus the linear classifier's predictive abilities are dependent upon the class population.

With fewer patterns in one of the classes, the dichotomization becomes easier. The linear classifier is less constrained when there are less patterns representing one class and the probability of finding a separation increases. The probability that the separation is trivial, if one is found, increases.

Reconsidering the familiar probability of dichotomization vs. λ plots, a more complete interpretation is possible. It would appear that these plots represent the lower limit of the probability for a given number of patterns and descriptors. This would naturally occur where the class populations were equal and is calculated from the theoretical probability equation which does not consider the distribution of patterns between the two classes. A plot of the probability of dichotomization vs. the fraction of the total number of patterns in a class would give a symmetric plot with minimum at a fraction of 0.5 and dual asymptotic maxima at fractions of 0 and 1.

With this interpretation of the probability of dichotomization plots and the awareness of the effects of class populations on the behavior of linear classifiers, it is possible to establish training parameters to maximize the likelihood of finding a separation with physical significance if one exists.

The probability of finding a linear discriminant that completely separates the two classes due to chance can be kept extremely low by ensuring that N/d is greater than ~3 and that the number of patterns in the less populated class exceeds d. Optimum probability is achieved when the class populations are equal.

REFERENCES AND NOTES

(1) G. L. Ritter and H. B. Woodruff, *Anal. Chem.*, **49**, 2118 (1977).

(2) C. P. Wiesel and J. L. Fasching, *Anal. Chem.*, **49**, 2114 (1977).

(3) N. A. B. Gray, *Anal. Chem.*, **48**, 2265 (1976).

(4) A. J. Stuper and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **16**, 238 (1976).

(5) L. F. Bender, H. D. Shepard, and B. R. Kowalski, *Anal. Chem.*, **45**, 617 (1973).

(6) J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Inc., Reading, Mass., 1974.

(7) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, 1965.