Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 9193–9196, December 1990
Applied Mathematics

Multisurface method of pattern separation for medical diagnosis applied to breast cytology

(linear programming/pattern recognition/expert systems/cancer diagnosis)

WILLIAM H. WOLBERG* AND OLVI L. MANGASARIAN†

*Department of Surgery, Clinical Sciences Center, University of Wisconsin, Madison, WI 53792; and †Computer Sciences Department, University of Wisconsin, Madison, WI 53706

Communicated by George B. Dantzig, August 21, 1990

ABSTRACT **Multisurface pattern separation is a mathematical method for distinguishing between elements of two pattern sets. Each element of the pattern sets is comprised of various scalar observations. In this paper, we use the diagnosis of breast cytology to demonstrate the applicability of this method to medical diagnosis and decision making. Each of 11 cytological characteristics of breast fine-needle aspirates reported to differ between benign and malignant samples was graded 1 to 10 at the time of sample collection. Nine characteristics were found to differ significantly between benign and malignant samples. Mathematically, these values for each sample were represented by a point in a nine-dimensional space of real variables. Benign points were separated from malignant ones by planes determined by linear programming. Correct separation was accomplished in 369 of 370 samples (201 benign and 169 malignant). In the one misclassified malignant case, the fine-needle aspirate cytology was so definitely benign and the cytology of the excised cancer so definitely malignant that we believe the tumor was missed on aspiration. Our mathematical method is applicable to other medical diagnostic and decision-making problems.**

In this report we present a general mathematical method for aiding medical diagnosis and decision-making and demonstrate its application in diagnosing breast mass cytology. Multisurface pattern separation (1) is a mathematical method for distinguishing between pattern sets. The pattern sets are comprised of elements that consist of various scalar observations. In the present application, the pattern sets were benign or malignant and each element of the pattern sets consisted of nine cytological characteristics of benign or of malignant breast fine-needle aspirates (FNAs). These nine characteristics have been established to differ between benign and malignant samples, but no single characteristic alone or presently described pattern distinguishes between benign and malignant samples. Discrimination between benign and malignant samples was accomplished by multisurface pattern separation.

Eleven cytological characteristics of breast FNAs were valued on a scale of 1 to 10, with 1 being the closest to benign and 10 the most anaplastic. Statistical analysis (2) showed that the following nine characteristics differed significantly between benign and malignant samples: uniformity of cell shape, uniformity of cell size, clump thickness, bare nuclei, cell size, normal nucleoli, clump cohesiveness, nuclear chromatin, and mitoses. The multibranched decision tree, which was developed to distinguish benign from malignant epithelial cell clumps (3), gave 16 false-positive and 3 false-negative classifications with a subset of the present data, consisting of the first 352 FNAs (190 benign and 162 malignant) (4). However, this decision tree considered only four of the cytological characteristics and equally satisfactory separation resulted from simply summing the values for all nine of the characteristics (2). In contrast, the mathematical method (1) used in this paper simultaneously weights the values for all nine cytological characteristics and designs a piecewise-linear separating surface composed of four pairs of parallel planes that completely separates 201 benign samples from 168 malignant samples.

MATERIALS AND METHODS

FNA. A standard method was used to obtain and prepare breast FNAs (3). The sample considered in this paper is a subset of 482 breast FNAs that contained epithelial cell clumps. Excluded, but previously reported (4), were 20 unsatisfactory preparations, 65 satisfactory preparations that did not contain epithelial cells, and 27 preparations that contained only nonclumped completely free epithelial cells.

Cytological Characteristics. The following cytological characteristics were assessed: the extent to which epithelial cell aggregates were mono- or multilayered (clump thickness), cohesion of the peripheral cells of the epithelial cell aggregates (marginal adhesion), the diameter of the population of the largest epithelial cells relative to erythocytes, the proportion of single epithelial nuclei that were devoid of surrounding cytoplasm (bare nuclei), blandness of nuclear chromatin, normal nucleoli, infrequent mitoses, uniformity of epithelial cell size, and uniformity of cell shape.

All malignant aspirates were histologically confirmed whereas FNAs diagnosed as benign masses were biopsied only at the patient's request. The remainder of benign cytologies were confirmed by clinical reexamination 3 and 12 months after the aspiration. Masses that produced unsatisfactory or suspicious FNAs were surgically biopsied.

Multisurface Method of Pattern Separation. The computational problem addressed here is that of discriminating between the elements of two disjoint sets of points, sets B and M. Set B consists of 201 points each of which represents nine measurements made on FNAs of benign breast masses. Set M consists of 168 points representing the same nine measurements made on malignant masses. This is a classical problem of pattern classification or separation (1, 5, 6) that we have solved by a nonparametric method of pattern classification (1, 6) that employs linear programming (7) as the basic computational tool. A brief outline of the underlying mathematical ideas of the method follows.

Point sets B and M are linearly separable if and only if there exists a plane or, more precisely, a hyperplane in the nine-dimensional space of real variables such that points of sets B and M are on opposite sides of this plane. Such a plane can be easily constructed by solving a single linear program (6). Since, in general, sets B and M are not linearly separable, as is the case here, we resort to the technique of multisurface

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: FNA, fine-needle aspirate.

pattern separation (1) whereby a piecewise-linear separating surface is constructed sequentially by means of a finite number of pairs of parallel planes as follows. Any two planes in an n-dimensional real space determine three disjoint regions. By linear programming, we determine two such planes, as close together as possible, so that only the region between contains points from both sets B and M. Having done this, we discard the outside points and repeat the process with the inside points and so on, until what is left is linearly separable. Classification of the two sets is thus achieved by checking whether each point lies outside the region between the first pair of parallel planes. If it does, then it is classified at stage 1 as being in set B or in set M depending on which side it is on. If it is between the first pair of parallel planes (or on them), the point is then checked by the same process using each of the remaining pairs of parallel planes in a prescribed order. The process is continued until every point is classified at some stage by some pair of parallel planes. When sets B and M are disjoint, such a classification procedure is always possible (1); however, its usefulness in general decreases if a large number of parallel planes are required. For the present problem, complete classification of 369 samples was achieved with four pairs of parallel planes. The planes were obtained by using a procedure similar to that of Mangasarian (1) and employed the MINOS linear programming package developed at Stanford University (8). If we denote by $(x_1, x_2, \ldots, x_9)$ the nine real numbers representing the nine measurements associated with each sample, then the precise classification procedure consists of the following:

If $c_{11}x_1 + c_{12}x_2 + \ldots + c_{19}x_9 > b_1$, then it is malignant at stage 1.

If $c_{11}x_1 + c_{12}x_2 + \ldots + c_{19}x_9 < a_1$, then it is benign at stage 1.

If $c_{21}x_1 + c_{22}x_2 + \ldots + c_{29}x_9 > b_2$, then it is malignant at stage 2.

If $c_{21}x_1 + c_{22}x_2 + \ldots + c_{29}x_9 < a_2$, then it is benign at stage 2.

If $c_{31}x_1 + c_{32}x_2 + \ldots + c_{39}x_9 > b_3$, then it is malignant at stage 3.

If $c_{31}x_1 + c_{32}x_2 + \ldots + c_{39}x_9 < a_3$, then it is benign at stage 3.

If $c_{41}x_1 + c_{42}x_2 + \ldots + c_{49}x_9 \geq (a_4 + b_4)/2$, then it is malignant at stage 4.

If $c_{41}x_1 + c_{42}x_2 + \ldots + c_{49}x_9 < (a_4 + b_4)/2$, then it is benign at stage 4.

The 44 numbers c_{ij}, a_i, b_i, for $i = 1, 2, 3, 4$ and $j = 1, 2, \ldots, 9$, were obtained using linear programming in an iterative fashion (1) on the 369 samples. The nine-dimensional vector $(c_{i1}, c_{i2}, \ldots, c_{i9})$ is the normalized perpendicular vector to the ith pair of parallel planes, and the absolute values of a_i and b_i give the distance of the two ith parallel planes to the origin in the nine-dimensional space. A Fortran program was written to implement the above classification procedure for both a Sun 4/110 workstation and for an IBM PC. Given a nine-dimensional vector, the program instantly classifies the point into a benign or malignant category and gives the stage, 1, 2, 3, or 4, at which the classification was made.

The validity of the method was tested by randomly splitting the 369 samples into training samples (i.e., samples that generate the separating planes) and into testing samples (i.e., samples that test the correctness of the classification by the separating planes).

RESULTS

We first note that the FNA cytology in the one excluded malignant case was unmistakably benign. The cells were approximately 7 μm in diameter and were uniform in size and shape, the nuclear chromatin was fine and homogeneous, and the nucleoli were normal (Fig. 1). The cytology of the excised cancer unmistakably was malignant. The tumor was very cellular, the nuclei measured 15 μm in diameter, the cells were irregular in size and shape, the nuclear chromatin was course and clumped, and the nucleoli were very prominent (Fig. 2). Since there were also 18 benign cases with identical characteristics to those of this malignant case, we decided to exclude this sample from the classification procedure. This case was also incorrectly classified by our other classification schemes.

As stated above, when all of the 369 samples were used as a training set, complete classification was achieved by means of four pairs of parallel planes. When 50% of the samples were used as a training set, only two pairs of parallel planes were needed to completely classify the training set. When these two pairs of parallel planes were tested on the remaining 50% of sample points, 6.5% of the samples were misclas-

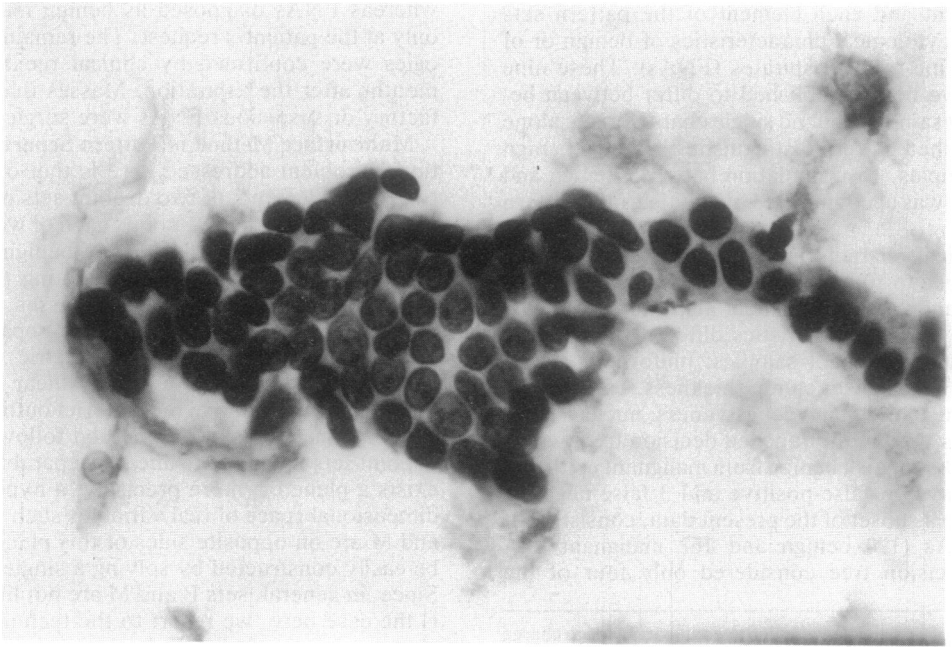

FIG. 1. FNA cytology of the one misclassified malignant specimen. (×900.)

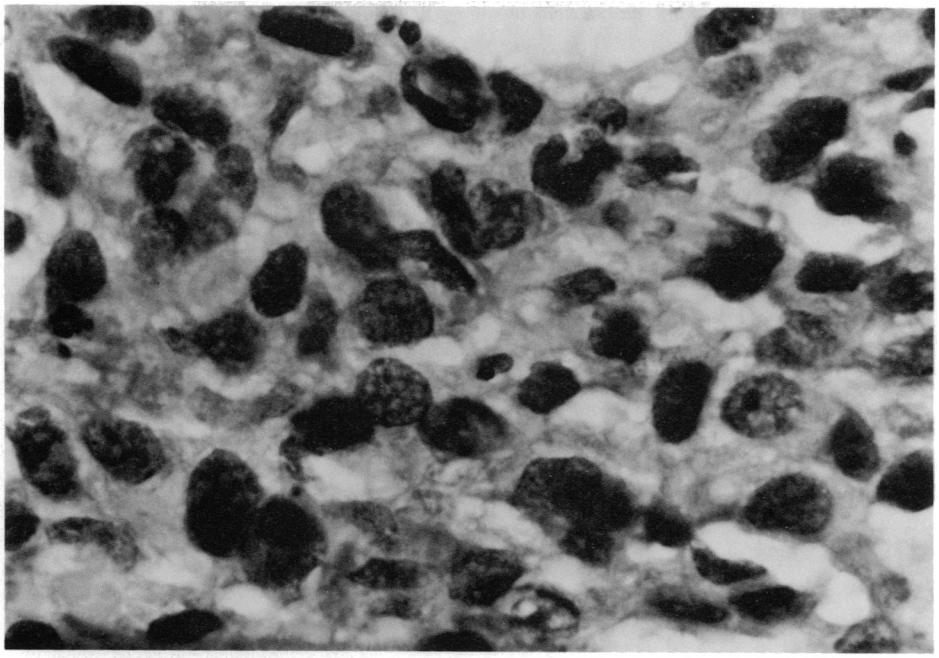

FIG. 2. Histology of the misclassified malignant specimen. (×900.)

sified. When 67% of the samples were used as a training set, only three pairs of parallel planes were needed to completely classify the training set. When these three pairs of parallel planes were tested on the remaining 33% of sample points, 4.1% of the samples were misclassified. These results indicate that the accuracy of separation increases with the size of the training set. When all the samples are used as a training set, the accuracy of separation depends on how well the whole sample set represents all other possible points.

DISCUSSION

The mathematical method used in this paper simultaneously weights the values for nine diagnostically important cytological characteristics (1) and designs four parallel planes to separate benign samples from malignant ones. Correct separation was accomplished in 369 of 370 samples (201 benign and 169 malignant). Because of the marked cytological differences between the FNA and the excised histology, we do not believe the single excluded sample resulted from cytological misinterpretation but rather resulted from missing the tumor on aspiration. Clinical features support this theory. The cancer was 2 cm in diameter and was located deeply in a rather large breasted premenopausal woman.

The multisurface method of pattern separation is more powerful than previously used methods for breast cytology diagnosis because it utilizes all of the available diagnostic information. This is accomplished by simultaneously weighting the values of nine diagnostic characteristics (1). In contrast, statistically based schemes previously employed for diagnosing breast FNAs considered only a few cytological characteristics or else weighted all nine equally. The multibranched decision-tree misclassified 2 malignant and 19 benign cases from the present data set excluding the malignant case that was missed by aspiration. Nevertheless, classification by the previous schemes produced results comparable to the best reports in the literature (4) (16 false-positive and 3 false-negative classifications with a 352-case subset of the present data set). With the present method, cytological diagnosis correlates so closely with histological diagnosis that diagnostic differences cannot be assumed to be due to inaccuracies in cytological diagnosis. Albeit, invasion cannot be diagnosed cytologically.

The accuracy with which benign and malignant breast masses were diagnosed is an illustrative example of the value of the proposed mathematical method in medical problem solving. A personal computer program incorporating the separating surfaces constructed by our method is available. Instant diagnosis by the program simply requires the entry of the numerical values for the nine cytological variables. Utilization by others of the method is predicated on the reproducibility of numerical values assigned to cytological variables. Such reproducibility has indeed been the case for values assigned by different trained observers.

We finally note the importance of the universality of the training set as demonstrated by our numerical experiments in which the total data set was divided unequally into training and test sets. With larger training sets more planes were required to separate benign from malignant points, but the separation became more accurate. The accuracy of predicting outcome by the present method depends on how universally representative the original data base is, the precision of the numerical representation of the cytological variables, and the inherent mutual exclusiveness of the pattern sets, in this case the benign and malignant histological diagnoses. Nevertheless, we feel that the multisurface method provides a very useful diagnostic tool.

We are indebted to Mr. Rudy Setiono for his expert and creative computational implementation of the algorithm described in this work. This research is supported in part by National Science Foundation Grants CCR-8723091 and DCR-8521228 and Air Force Office of Scientific Research Grants AFOSR-86-0172 and AFOSR-89-0410.

1. Mangasarian, O. L. (1968) *IEEE Trans. Inf. Theory* **IT-14,** 801–807.
2. Wolberg, W. H., Tanner, M. A. & Loh, W.-Y. (1988) *Anal. Quant. Cytol. Histol.* **10,** 225–228.
3. Wolberg, W. H., Tanner, M. A., Loh, W.-Y. & Vanichsetakul, N. (1987) *Acta Cytol.* **31,** 737–741.
4. Wolberg, W. H., Tanner, M. A. & Loh, W.-Y. (1989) *Arch. Surg.* **124,** 814–818.
5. Duda, R. O. & Hart, P. E. (1973) *Pattern Classification and Scene Analysis* (Wiley, New York).
6. Mangasarian, O. L. (1965) *Oper. Res.* **13,** 444–452.

7. Dantzig, G. B. (1963) *Linear Programming and Extensions* (Princeton Univ. Press, Princeton, NJ).

8. Murtagh, B. A. & Saunders, M. A. (1983) *MINOS 5.0 User's Guide* (Stanford Univ., Stanford, CA), Tech. Rep. SOL 83.20.