Pattern Classification by Adaptive Machines

Patterns are categorized by the use of fixed and adaptive networks.

Charles A. Rosen

Man has been striving for centuries to conceive and fabricate machines to replicate or improve his own performance. In the past 200 years he has made astonishing progress in first replacing and then exceeding his purely mechanical abilities. In the past 20 years, new developments in a host of scientific fields ranging from biology to computer technology have made possible exciting new approaches to the far more difficult task of supplementing or ultimately supplanting the intelligent capabilities of man.

There is no universally accepted definition of intelligence (*1*). We can describe an intelligent automaton operationally as a machine that can perform tasks that normally require almost continuous human control and intervention. Man's pattern recognition process—that is, his ability to select, classify, and abstract significant information from the sea of sensory information in which he is immersed—is a vital part of his intelligent behavior. In the past 10 years pattern recognition has become a major focus of research by scientists working in the field of artificial intelligence.

In its most general form, pattern recognition is a complex of functions operating at many levels. At the lowest level it consists of the ability, shared by the simplest biological species, to sense, select, combine, and identify environmental data necessary to preserve life and to continue the species. At a much higher level, man is able to use advanced forms of pattern recognition to deal with highly abstracted information. This function is embedded in a complex of mental processes that includes logic, induction, association, and invention, among others. At present knowledge of pattern recognition is essentially at the first level; a relatively small but growing effort is being devoted to more difficult problems involving both elementary pattern recognition and simple logical processing, both woven into a single system.

At the lowest level, general pattern recognition reduces to pattern classification, which consists of techniques to separate groups of objects, sounds, odors, events, or properties into classes, based on measurements made on the entities being classified. This article is primarily concerned with pattern classification, with a brief introduction to methods and concepts being explored to attack the more difficult problems in pattern recognition.

Effective techniques for pattern classification have been developed; these techniques represent a small but significant part of pattern recognition systems. No general methods now exist for either the selection of appropriate measurements or the logical embedding of classification techniques in an overall recognition system. These are the areas of research today.

Processing and Categorization

In dealing with the query "Is this a man or a mouse?" we can recognize at least two distinguishable processes that are themselves so closely related that it is often difficult to clearly define the domain of each. For most classification problems it is adequate to perform two operations: first, to select a set of measurements (features), such as size, shape, color, odor, number of feet, absence or presence of tails and clothing, and the like; and second, to use these features either to distinguish between this man and this mouse, or, more generally, to define the classes of mice and men. For the more difficult query "Is this a boy or a girl?" it may be necessary either to select a more complex set of distinguishing features, or to augment the list of simple features. The method, and the form of the machine which uses the method to classify the phenomena with the selected features, can be the same for answering both questions. The pattern classifier can be divided into two subsystems: the data filter and the categorizer. Data filtering consists of selecting distinguishing features and representing them in terms of sets of real numbers, each set being termed a pattern. Each pattern is categorized by being assigned to one of several classes.

While there are a number of methods for the design of categorizers, there is no general method to select the distinguishing features. At present the data filter is designed with empirical methods based on "common sense," trial and error, educated guesses, imitation of biological data filters, or statistical analysis (*2, 3*). As the problem becomes more complex, more features are usually required. It becomes economically infeasible, however, to use all the information available to describe a given entity. In fact, such all-inclusive descriptions will usually degrade the performance of the categorizer by contributing irrelevant information that can be lumped together with other disturbances as "noise." In this, as in other information processes, significant features or useful information is masked by the irrelevant data.

We do not yet know whether there are any general methods that can be used to design data filters. It is possible that the selection of features for any given domain of pattern classification is relevant to that domain alone. It thus becomes important to determine the relative efficiency of features selected by any method. Some statistical methods can be used to test the relative importance of selected features. These methods become unwieldy as the number of such features increases. Newer methods based on adaptive techniques and evolutionary methods of generation and testing of features are promising (*4, 5*).

Much of the future research in pattern classification will be devoted to understanding and elaborating the role

The author is manager of the Applied Physics Laboratory, Stanford Research Institute, Menlo Park, California.

of the data filter, devising adequate means of testing the effectiveness of features, and finally, integrating data filter and classifier into one working system.

There are two general approaches in the design of a categorizer. Both approaches usually require a statistically large sample of correctly identified and labeled patterns represented by appropriate features. The first approach assesses the relative importance of the features by statistical means, with known or assumed probabilities of occurrence of each feature, or combinations of features in the patterns comprising each category. With known or derivable statistics, a categorizer can be designed with fixed internal organization to yield results with minimum errors (*6*). In many, if not most cases, neither the probability distributions of the features nor any simple assumptions about the form or nature of the distributions are valid, and more powerful techniques are required. Exploration of these techniques has led to the second general method: the adaptive network approach. In its ideal form, a machine composed of adaptive or alterable networks would automatically and progressively change its internal organization as a result of being "trained" by exposure to a succession of correctly classified and labeled patterns. In this case there is no prior knowledge of the relative importance or distribution of the features comprising each pattern. After a training or learning process, the final organization would effect, ideally with minimum error, the classification of patterns that are not contained in the training set but are quite similar to them. The ideal machine does not exist, not even biologically. Some relatively simple structures have been made (*7, 8*), and many more have been simulated on a digital computer, which do, in fact, progressively alter their internal structure with experience. These structures perform acceptably with certain classes of problems.

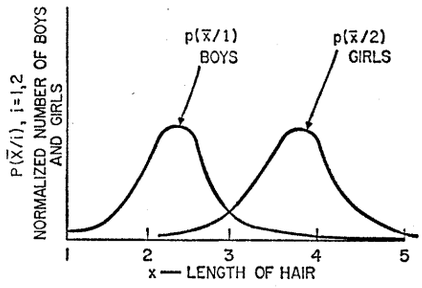

Fig. 1 (left). Two classes of Gaussian distributions.

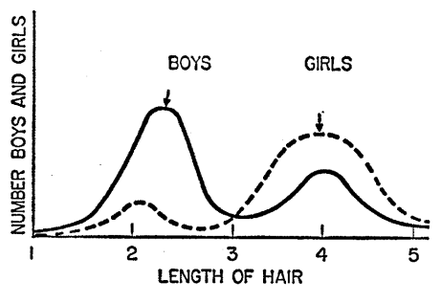

Fig. 2 (right). Two classes of non-Gaussian distributions.

Pattern Representation

Suppose our problem is to distinguish between boys and girls on the basis of a single measurement or feature, and we choose the average length of hair as the distinguishing feature. If we further restrict ourselves to youths between 8 and 16 years old, and take equal numbers of boys and girls as our samples, we might obtain the smoothed out curves shown in Fig. 1, which shows the distribution of boys and girls as a function of the average length of hair. In this highly idealized case the curves are familiar bell-shaped (Gaussian) distributions. Clearly, a classification scheme that would assign the label "boy" to any youngster with average hair length less than arbitrary length 3, and the label "girl" for hair length greater than 3, would yield a minimum number of errors (*9*).

A more realistic set of curves might be as shown in Fig. 2, which illustrates that a small number of boys today consider long hair fashionable, and some girls prefer short hair. Minimum number of errors can no longer be obtained by separating the classes by simple means and, although it is possible to find an optimum solution by statistical methods, this solution will yield a high number of errors. Clearly, more identifying features are necessary. Suppose we added another feature, for example, a number representing the ratio of shoulder to hip size. It would be difficult to illustrate this new situation in the same manner as shown in Figs. 1 and 2. A three-dimensional space is required in which two-dimensional surfaces are drawn. A more geometric representation is shown in Fig. 3, one which leads directly to a useful algebraic description of each pattern. Again, for simplicity this figure is a highly idealized representation. The pattern B is a solid dot representing one boy with hair 1.5 arbitrary units long and a ratio of shoulder to hip size of 1.3. Similarly the pattern G is an open circular dot representing a girl with hair length 4 arbitrary units, ratio of shoulder to hip size 1.1; these two samples can be represented algebraically: $B(1.5,1.3)$; $G(4,1.1)$. To represent patterns geometrically with more features would require higher dimensional spaces. To represent them algebraically merely requires adding corresponding sets of numbers for additional features. Thus if one added as a new identifying feature the pitch of voice on an arbitrary scale of 1 to 4 spread over a range from bass to alto, the boy

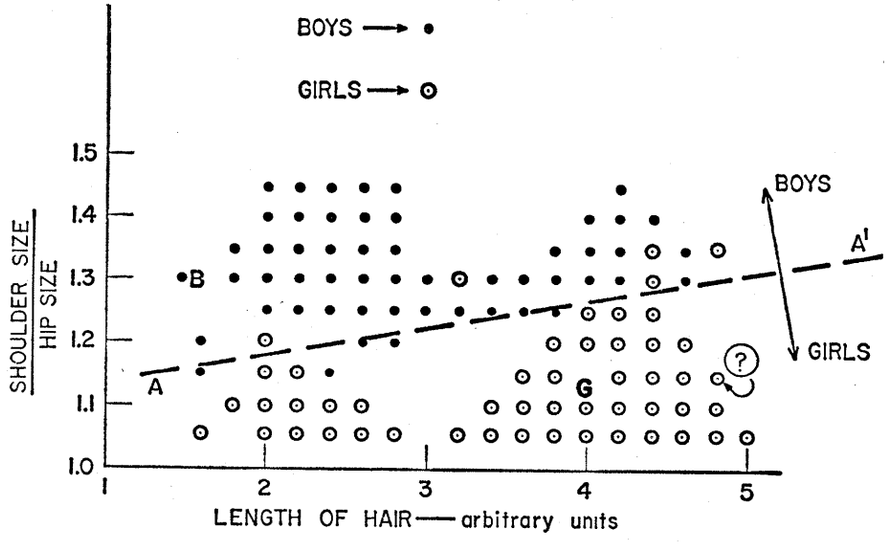

Fig. 3. Geometric representation of two features.

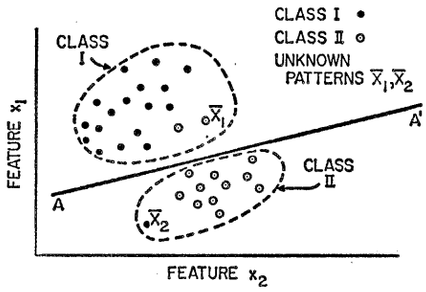

Fig. 4. Two classes with separable unimodal distributions.

would be represented by: $B(1.5,1.3,1)$. In general a boy to be represented by n features would be described: $B(x_1,x_2, \ldots x_n)$, where $x_1, x_2, \ldots x_n$ are numbers associated with the features, and the whole expression (a vector representation) would represent one pattern and its label (boy). The n-dimensional vector would then become a single point in an n-dimensional geometric space.

In Fig. 3, the pattern classification problem is to find some means of classifying a new point (indicated by ? in the figure) as boy or girl. One method that can be used would be to find a separating line, AA', which divided the plane into two regions, one region assigned to each class, with a minimum number of errors. In this case the *new* point can clearly be classified as a girl.

The Adaptive Categorizer

Assuming that we have selected and made some measurements which represent each of the patterns, let us examine more closely some of the problems of final classification. For simplicity we will deal with two classes of patterns to be distinguished from each other on the basis of only two features. Figure 4 shows the results of plotting the known (correctly labeled) patterns. Each pattern is shown as a single point in a two-dimensional space; the coordinates of the point are the measured values of the two features representing that pattern. Each class of pattern is represented by a cluster of scattered points, since the values of the distinguishing features are not identical for each pattern in that class. (In the boy-girl classification example, the average length of hair could well be different for each boy.) In the ideal case, the cluster representing each class would reduce to a single point, indicating that there existed a set of features which were identical for each member of that class. Such instances can be contrived, but they rarely occur in practice.

In this example the patterns (points) in each class, although scattered, form two fairly well-defined clusters; these clusters can evidently be separated by drawing the line AA', which divides the space into two regions. All points above this line belong to class one, all points below belong to class two. If two points representing two previously unclassified patterns $\bar{x}_1$, $\bar{x}_2$ are plotted, then $\bar{x}_1$ would be assigned to class one, and $\bar{x}_2$ assigned to class two. The line AA' is one of many that can be drawn to separate the clusters and in fact it need not be a straight line, but can be a segment of a higher order curve such as a circle, parabola, ellipse, or the like. However, the curve that can be represented most easily is a straight line, and if the cluster shapes are such as to require a curved line for separation, the curved line can be approximated by several straight line segments.

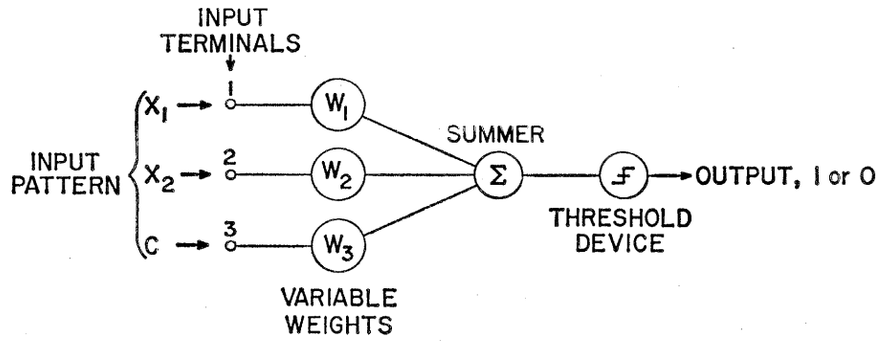

Fig. 5. Threshold logic network or unit.

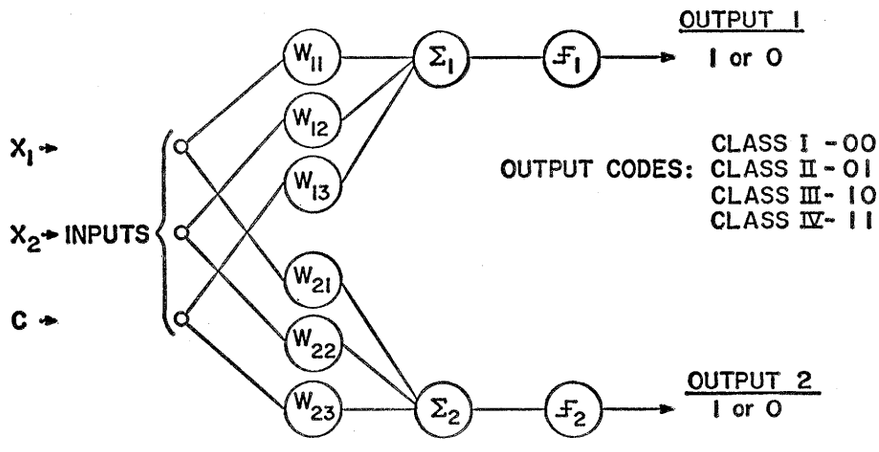

Fig. 6. Four-class threshold logic network.

Shown schematically in Fig. 5 is a threshold logic network, which, on the basis of samples of known and labeled patterns of each class, can use one of the many straight line solutions to the two-dimensional problem described in Fig. 4. The network is operated in three steps: (i) A pattern, represented by two numerical measurements x_1, x_2, and a constant, C, is presented to the network input terminals one, two, and three. (ii) These inputs are "weighted" by each terminal and multiplied by a variable quantity of weight W_1. They yield the products x_1W_1, x_2W_2, and CW_3, which are then added together algebraically to form a linear combination of weighted input values. These weights will be altered progressively as the network is "trained." (iii) The sum $(x_1W_1 + x_2W_2 + CW_3)$ is presented to a threshold device which determines whether or not it is greater than the threshold, 0—that is, whether this sum is positive or negative (*10*). If the sum is positive we can arbitrarily assign that pattern to class one; if the sum is negative, or 0, the pattern is assigned to class two. The output need therefore have only two values, which are here arbitrarily chosen as 1 or 0.

The object of the game is to find a set of weights W_1, W_2, and W_3, such that this network will give the output

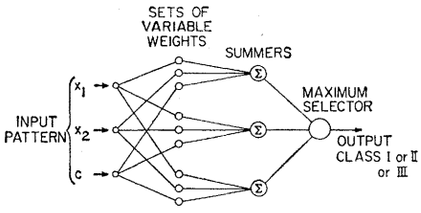

Fig. 7. Linear machine.

1, for all patterns known to be in class one, and the output 0 for all class two patterns. (Changing the values W_1, W_2, and W_3 is equivalent to rotating and translating the line AA' in Fig. 4).

One training procedure is the error-correction method. (i) Each pattern from a set of known patterns, called the training set, is presented sequentially to the network. (ii) If the network output correctly classifies the known pattern, no change is made and the network passes on to the next pattern. (iii) If the pattern is incorrectly classified then each weight is automatically changed by a small amount to correct the wrong response.

With these rules, the network is presented sequentially with all the patterns of the training set; corrections to the weights are made only on errors. The whole procedure is then repeated with the same training set. Repetition ceases either when there are no errors or when an acceptably small number of errors are produced in a complete pass through the training set. If there are no errors, then the process is said to have converged and the classes are described as linearly separable. Geometrically (Fig. 4), this means that the scatter of the two clusters permitted a straight line (AA') to be drawn which accurately separated all points into two desired classes.

The basic ideas incorporated in the above procedure were first proposed and analyzed by Rosenblatt (*7*) and later elaborated on by Widrow (*11*). If the classes are linearly separable this procedure is guaranteed to converge in a finite number of repetitions. In other words, a solution is guaranteed, if one exists.

Multiple Classes

This procedure may be extended to problems that involve more than two classes. Figure 6 shows an arrangement whereby the addition of another threshold logic network, operating in parallel at the input with the original network, enables classification of as many as four classes. The two sets of outputs (0 or 1 from each of the networks) can be arbitrarily arranged to form four distinct output codes, namely 00, 01, 10, and 11, each of which may represent a class. For example the class represented by code 01 means that network one has a desired output of 0, and network two has simultaneously a desired output of 1. The training procedure (error-correction) is unchanged; each of the two networks is treated in precisely the same way as for the two-class case. A network's weights are altered if and only if its particular binary output is wrong. Thus if a training pattern produces an output of 01, while the correct output should be 00, the output of network one is correct and its weights will remain unchanged; the output of network two is incorrect and its weights will be adjusted to make its output correct.

The maximum number of assignable code words, and therefore classes, increases exponentially, as 2^n with the number, n, of threshold logic networks. There is no known analytical procedure to assign a particular code word to a particular class, and the difficulty in separating many classes may be increased enormously by a poor choice of code words for the different classes. A number of schemes use modern coding techniques, which generally improve performance by using less than the maximum number of possible codes for a given number of threshold logic networks. Other machine organizations avoid this problem.

The Linear Machine

The linear machine eliminates this coding problem entirely and has shown excellent performance in multiclass classification problems (*8, 13*). Figure 7 shows a linear machine composed of three parallel networks of adaptive elements designed for the categorization of three classes of patterns. Each network accepts the input pattern composed of two measured features, x_1 and x_2, and a constant, C. These values are multiplied by three sets of variable weights, and the sums are compared in a device which selects the largest sum. Each network is assigned to one class. A presented pattern is identified as belonging to the class of the network with the maximum sum. The combination of multiplicative weights and the summer is called a dot-product unit (*12*).

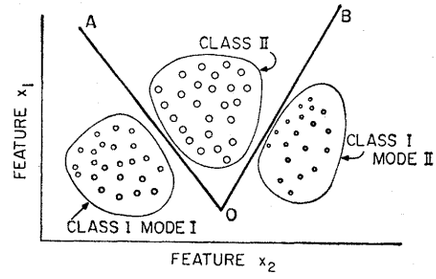

Fig. 8. Multimodal classification problem.

The training procedure of this machine is similar to that of the threshold logic machine. A set of training patterns is presented one at a time. If a correct classification is made, the weights associated with the wrong dot-product unit are each changed to reduce its sum output, and simultaneously the weights associated with the desired (or right) dot-product unit are changed to increase its sum output. This procedure is repeated until all or almost all of the patterns in the training set are correctly identified by the "trained" networks.

A linear machine with n dot-product units acting in parallel becomes an n-class categorizer, with one network assigned to each of the n classes. There is no output coding problem; the class is determined by the summer with the greatest output. The simplification and increased performance has, apparently, been achieved at the price of using many more networks. The same number of networks arranged as threshold logic units, with ideal coding, would probably handle far more classes. It is an open question whether such ideal codes exist.

As with the threshold logic machine, if a solution exists (that is, if the problem is linearly separable) the error-correction procedure will converge to a solution for the n-class linear machine (*13*).

Multimodal and Other Problems

Figure 8 shows a representation of a problem containing two classes, with one class having two modes or clusters (*14*). This is a simple version of a multimodal problem. The surrounding hulls roughly define the boundaries of the clusters. A two-class problem is shown in Fig. 9 where one class has a complex nonspherical distribution. In both these cases the two

classes cannot be separated by a single straight line. One could use curved lines or, as shown, several intersecting single lines, *AO* and *BO*. These lines would define the boundaries of several regions, each of which could be assigned to a given class.

Problems with these properties can often be handled quite adequately by machine organizations that are variants of the threshold logic and linear machines already described—for example, the Madaline (*15*) and Piecewise linear machines (*8, 13*). Figure 10 shows a schematic drawing of a Piecewise linear machine that is designed to handle two classes with two modes in each class. As in the linear machine the basic networks are dot-product units, with one unit assigned to each known or suspected mode. In the diagram, two dot-product units are assigned to each class. When a pattern is presented to the input terminals common to all the dot-product units, it is recognized as belonging to the class represented by the dot-product unit with the largest output sum. The error-correction training is essentially the same as for the linear machine, that is, as the training patterns are presented in sequence, no corrections are made if any one of the dot-product units associated with the correct class has a maximum output. If the wrong dot-product unit has a maximum, then its weights are changed by equal amounts to reduce the total output sum. Simultaneously one of the right dot-product units which has the largest output sum for that pattern (but not larger than that of the wrong dot-product unit) has each of its weights changed to increase its output sum.

This organization and training procedure does not always converge to a possible solution, but it often converges closely enough to be useful in many practical cases. A method guaranteed to converge is still being sought.

Prototype and Cluster Seeking

Other interesting methods for the design of the categorizer are being developed. An important group of methods stems from one of the first techniques used in pattern classification—that of template matching. This method consisted of storing all the known patterns, comparing the unknown pattern with each of the stored patterns according to some similarity measure, and assigning the unknown to the class of that known pattern to which it is most similar. In terms of our geometric representations, a useful similarity criterion could be the "nearness" between the point representing the unknown pattern and each of the points representing all the known patterns. "Nearness" is usually measured by the Euclidean distance between points. In many problems the large number of distance computations makes this method prohibitively expensive. Since patterns of different classes tend to aggregate together into distinguishable clusters of points, each cluster of points can be replaced by one or more prototype points which are representative of the cluster. The distance between the unknown pattern point and each of the prototype points can be measured, the unknown point can be assigned to the class of the closest prototype. Since the number of classes (and therefore prototype points) is quite small compared to the necessary number of training patterns, the computation time which is saved is highly significant.

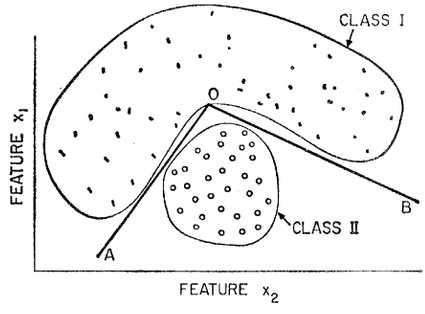

Fig. 9. Unimodal classification problem with nonuniform distribution.

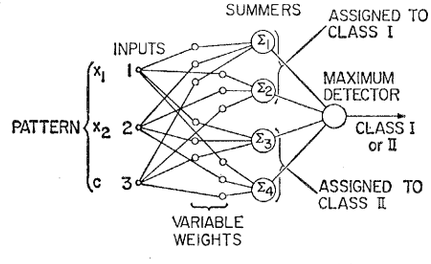

Fig. 10. Piecewise linear machine.

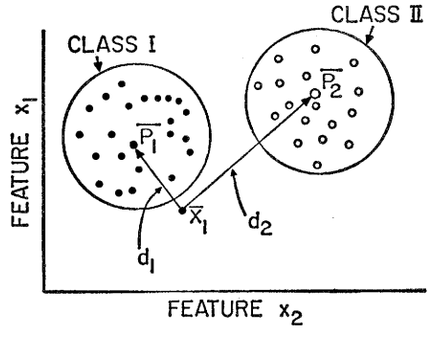

Fig. 11. Classification by prototype in unimodal distributions.

Figure 11 shows an example of a simple classification problem where the two classes are separated and clustered in an orderly manner. In this case the prototype points are $\bar{P}_1$ and $\bar{P}_2$, the centers of gravity of the two clusters of known patterns. An unknown pattern $\bar{X}_1$ with distances d_1 and d_2, respectively, to prototype points $\bar{P}_1$ and $\bar{P}_2$ would be classified as belonging to $\bar{P}_1$ (class one) because d_1 is smaller than d_2.

The situation is quite different when there are more than one cluster in one class, and the distribution of the patterns in this class is not known, that is, when one does not know which patterns belong to which of the clusters in that class. A case of this kind is shown in Fig. 12 where class one has two modes and it is not known beforehand which patterns (points) of class one are associated with each of the modes. If one chooses for the prototype of class one the center of gravity of all the points in class one, one would obtain the point $\bar{P}_4$, which might almost coincide with the prototype $\bar{P}_2$, the center of gravity of class two points. Clearly $\bar{P}_4$, in this case, is more representative of class two than of class one, and performance on new patterns would not be satisfactory. One has to identify the two modes and the known patterns associated with each mode if the prototype is to be represented adequately. Thus $\bar{P}_1$ and $\bar{P}_3$, the center of gravity prototypes for modes one and two of class one, would, together with $\bar{P}_2$ of class two, make a good representation of the data.

Serious problems arise in even simple cases, such as shown in Fig. 13, where each class is unimodal but one or more of the clusters may not be very regular. In this case the center of gravity $\bar{P}_1$ of the elongated cluster of class one would poorly represent this class. A better representation would be obtained by using two prototype points, $\bar{P}_3$ and $\bar{P}_4$.

It is occasionally possible, by detailed analysis of the data, to determine the distributions well enough to choose suitable prototype points. With multidimensional, multiclass problems this may be too difficult and expen-

sive. There are several cluster-seeking techniques available with adaptive or learning procedures which are effective in finding suitable prototypes, for example, one for each regularly-shaped cluster or several for irregular clusters (*5, 16, 17*). Although more calculations are performed to effect those methods than in the boundary-seeking methods, the results may justify the cost. A combination of cluster-seeking and boundary-seeking methods would probably be more effective than either; such combinations are being sought.

Applications and Results

Some classification techniques based on adaptive or trainable structures are matching and in some cases exceeding the performance of more conventional statistical methods. The problem areas range from weather forecasting, through medical diagnosis, into the more familiar area of recognition of alphanumeric characters. Nilsson (*2*) provides many up-to-date references. Casey *et al.* (*18*) conducted a large-scale experiment with 6000 typewritten alphanumeric characters of many fonts. They used an adaptive technique and reported about 20 errors, a performance better than other methods tried on the same data. Brain and Munson (*19*) show results of less than 1 percent error in the classification of complex hand-printed graphical symbols. Both of these problems are multimodal and multiclass, with poorly known distributions. Stark (*20*) has successfully attacked the problem of electrocardiograph interpretation with adaptive techniques. Gerdes (*21*) reports results on the recognition of man-made and other objects displayed in aerial photographs, a problem area of particular difficulty in view of the overwhelming mass of irrelevant information which is usually recorded and from which a relatively small selection must be made.

In each of the successful experiments, much effort was expended in the selection and extraction of suitable features in the data-filtering stage. There is no common thread running through either the type of features or the method by which they were selected. In the future more can be gained by improving the data filter than in finding more efficient ways to adaptively "learn" to perform satisfactory categorization.

Pattern Recognition: The Future

The problems remaining in the broader field of pattern recognition are formidable indeed. For example, the recognition by machines of hand-printed characters with performance equalling that of trained humans requires not only pattern classification, but the use of context. Because of considerable similarity of shape, some letters and numbers (*6* and *G*, *5* and *S*, *U* and *V*, for example) are often confused. Some poorly formed characters have lost all or most distinguishable features. Ambiguous or wrong classification decisions are highly probable unless contextual cues are used. Cues can be sought in the relations between the character in question and its neighboring characters, the words in which it is embedded, sentences, and so forth, until, if necessary, the whole text is used in the search. Similar but far more difficult problems in this class are the recognition of hand-written script and the recognition of the words in flowing speech. With the aid of programmed tests based on known contextual relations, the digital computer can be used to augment pattern-classification techniques.

As with feature selection in data-filtering systems, there are no general methods available for the development or selection of good contextual relations and tests. As yet, the value of these relations can only be judged empirically.

The visual recognition of three-dimensional objects presents many new problems. Objects that are to be recognized are usually embedded in an environment of many other distinguishable objects and backgrounds. The separation or isolation of individual objects preparatory to classification becomes a formidable task in its own right and may require considerable analysis based on such cues as color, three-dimensional measurements, and texture. If, in addition, one desires to recognize objects that are partially obscured by portions of other objects, a new hierarchy of tests may be necessary. It is generally infeasible to increase the number of categories to include all possible combinations of overlapping objects. As in the context-related problems, it appears necessary to discover, list, and test for relations between parts of objects, between objects and objects, and between objects and the environment in the field of view of the sensing device. In fact, the size of the field of view itself is an important variable, and areas of adjustable size should be sampled when a scene of interest is being scanned.

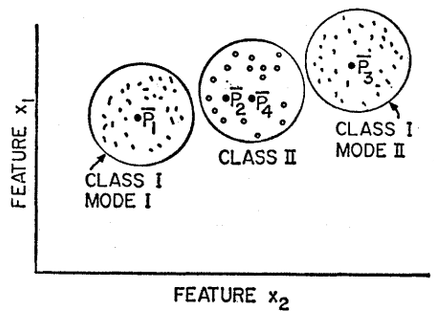

Fig. 12. Classification by prototype in multimodal distributions.

Again, as with advanced character-recognition systems, the programmed digital computer is essential in augmenting the categorization part of the recognition problem. Among other functions the computer program must provide for the manipulation and selection of the sensory data (*19*), and storage of relevant relations in a manner which simplifies search, retrieval, and logical questioning (*22*); the program must also provide instructions that govern the sequence of operations.

Summary

Man's intelligent behavior is due in part to his ability to select, classify, and abstract significant information reaching him from his environment by way of his senses. This function, pattern recognition, has become a major focus of research by scientists working in the field of artificial intelligence.

At the lowest level, pattern recognition reduces to pattern classification,

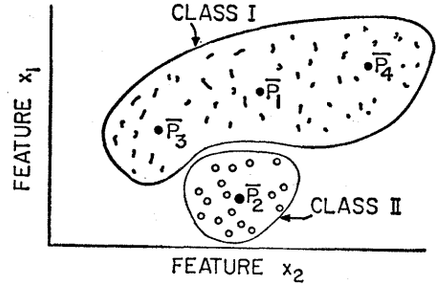

Fig. 13. Classification by prototype in nonuniform clusters.

which consists of the separation, into desired classes, of groups of objects, sounds, odors, events, properties, and the like; the separations are based on sets of measurements made on the entities being classified. The pattern classifier is composed of a data filter and a categorizer. The data filter selects the distinguishing features and represents them as sets of real numbers; each set is termed a pattern. The categorizer assigns each pattern to one of several desired classes.

Patterns can be represented geometrically as points in an n-dimensional space; the n coordinates of each point are the numerical values of the features selected to represent the pattern. A pattern classification system separates an n-dimensional space into regions, each of which ideally contains points of only one class. One method to effect this separation is by means of "trainable" categorizers—major components of adaptive machines. They consist of networks whose internal parameters are varied according to a set of fixed rules during a training cycle. A statistically large sample of known patterns are presented, one at a time, to the networks; internal corrections are made each time a pattern is erroneously classified. Classification performance tends to improve as the set of known patterns is cycled repetitively through the machine. Finally, the adequacy of adaptation is tested by a separate set of similar patterns which have not been used in the training process.

A number of different machine organizations and training rules have been developed and are being applied successfully to numerous classification problems. More difficult recognition problems requiring the aid of logical tests and analysis, search and association, use the digital computer programmed to supplement the functions of the adaptive classifier.

References and Notes

1. Minsky's view is that intelligence "is more of an aesthetic question, or one of a sense of dignity, than a technical matter . . . a complex of performances which we happen to respect but do not understand," M. Minsky, *IRE Inst. Radio Engrs.* **49**, 8 (1961). Pollard offers as "a cynical definition of Artificial Intelligence [that] it is a property of a system or machine which appears to give the capability of decision-making in a manner which is not readily apparent to an audience; however, when this property is explained in a logical manner it is deemed not to be intelligence because it is explainable." From a preface by B. W. Pollard, presented at Artificial Intelligence Sessions, Winter General Meeting, Institute of Electronic and Electrical Engineers, New York, January 1963. Turing's operational definition of a "thinking" machine was that it was one which could answer questions posed by a human questioner sufficiently well to deceive him into believing it was human. A. M. Turing, in *The World of Mathematics*, J. R. Newman, Ed. (Simon & Schuster, New York, 1956), vol. 4, pp. 2099–2133.
2. N. J. Nilsson, in *Proc. Bionics Symp. 1966*, in press.
3. D. H. Hubel and T. N. Wiesel, *J. Physiol.* **160**, 106 (1962); J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, W. H. Pitts, *IRE Inst. Radio Engrs.* **47**, 1940 (1959).
4. L. Uhr and C. Vossler, *Proc. Western Joint Computer Conf. 1961*, p. 555.
5. G. H. Ball and D. J. Hall, *Proc. Int. Communications Conf.*, 1966.
6. G. Sebestyen, *Decision-Making Processes in Pattern Recognition* (Macmillan, New York, 1962); N. Abramson and D. Braverman, *IRE Trans. Inform. Theory* **8**, 588 (1962).
7. F. Rosenblatt, *Principles of Neurodynamics; Perceptions and the Theory of Brain Mechanisms* (Spartan Books, Washington, D.C., 1961).
8. N. J. Nilsson, *Learning Machines: Foundations of Trainable Pattern Classifying Systems* (McGraw Hill, New York, 1965).
9. In Fig. 1, i represents the category, $\bar{X}$ represents the pattern, $p(\bar{X}/i)$ represents the probability density functions for boys and girls shown in the two curves, and $p(i)$ for $i = 1, 2, \ldots$, represents the a priori probability that a pattern belongs to category i. (In this case with assumed equal populations of boys and girls, $p(i)$ is always 0.5). For any new pattern $\bar{X}$, one computes the quantities $p(\bar{X}/i)\ p(i)$ for each i and assigns that pattern to that class i for which this quantity is maximum. This method yields the optimum result, but requires complete prior knowledge of the distributions.
10. Formerly, this type of machine used thresholds which were variable quantities, not fixed to be the constant zero. The term CW_3 in the sum is the equivalent, functionally, to the variable threshold.
11. B. Widrow and M. E. Hoff, *IRE Inst. Radio Engrs. WESCON Conv. Record* **1960**, 96 (1960).
12. The dot-product unit computes the scalar or dot product of the pattern vector $\bar{X}$ and weight vector $\bar{W}$, namely, $\bar{X} \cdot \bar{W}$.
13. R. O. Duda and H. Fossum, *IREE Inst. Elect. Electron. Engrs. Trans. Electron. Computers* **15**, 220 (1966).
14. The term "mode" is often used loosely in pattern recognition literature; frequently its usage is synonymous with the term "cluster." Statistically, modes are local maxima in probability distributions.
15. B. Widrow, in *Self-Organizing Systems 1962*, M. C. Yovits, G. T. Jacoby, G. D. Goldstein, Eds. (Spartan Books, Washington, D.C., 1962).
16. C. A. Rosen and D. J. Hall, *IEEE Inst. Elect. Electron. Engrs. Trans. Electron. Computers*, in press.
17. G. Sebestyen, *IRE Inst. Radio Engrs. Trans. Inform. Theory* **8**, 582 (1962).
18. R. G. Casey, Ed. *IBM Res. Rep. No. RC 1500* (1955).
19. A. E. Brain and J. H. Munson, *Stanford Res. Inst. Rep. No. 22*, April 1966; *ASTIA* No. AD632563, contract DA 36-039 AMC-03247(E) (U.S. Army Electronics Command).
20. L. M. Stark, M. Okajina, G. H. Whipple, *Comm. Ass. Computing Machinery* **5**, 527 (1962).
21. J. W. Gerdes, W. B. Floyd, G. O. Richards, *Tech. Rep. No. RADCTR-65-438*, Rome Air Development Center, Griffiss Air Force Base, N.Y. (1966).
22. B. Raphael, *Proc. Fall Joint Computer Conf.*, 1964.

23. I thank my colleagues at Stanford Research Institute, especially N. J. Nilsson and R. O. Duda, for helpful criticism. Supported by the Office of Naval Research, U.S. Army Electronics Command, and Rome Air Development Center.