ELSEVIER

Pattern Recognition Letters 20 (1999) 1103–1111

Pattern Recognition Letters

www.elsevier.nl/locate/patrec

Multidimensional curve classification using passing-through regions

Mineichi Kudo *, Jun Toyama, Masaru Shimbo

Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University, Kita-13, Nishi-8, Kita-ku, Sapporo 060-8628, Japan

Abstract

A new method is proposed for classifying sets of a variable number of points and curves in a multidimensional space as time series. Almost all classifiers proposed so far assume that there is a constant number of features and they cannot treat a variable number of features. To cope with this difficulty, we examine a fixed number of questions like "how many points are in a certain range of a certain dimension", and we convert the corresponding answers into a binary vector with a fixed length. These converted binary vectors are used as the basis for our classification. With respect to curve classification, many conventional methods are based on a frequency analysis such as Fourier analysis, a predictive analysis such as auto-regression, or a hidden Markov model. However, their resulting classification rules are difficult to interpret. In addition, they also rely on the global shape of curves and cannot treat cases in which only one part of a curve is important for classification. We propose some methods that are especially effective for such cases and the obtained rule is visualized. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Multidimensional curve classification; Classification of sets of multidimensional points; Variable number of features; Subclass method; Binarization

1. Introduction

Almost all conventional classifiers assume that a sample is expressed by a fixed number of features. However, in some applications, the most natural expression of a sample is a set of feature points in a fixed dimension; for example, a handwritten character may be expressed by a set of strokes and one stroke may be expressed by four numerical values for the start and end positions of the stroke. In such a case, the number of feature points (strokes) can vary depending on the sample. In other applications such as classification of a time series, one sample may be measured over a long duration and another sample over a short duration, which results in a variable number of observations. Therefore, we have to devise a classification method applicable for such cases. Many syntactic approaches may be applicable, but they usually require the extraction of a special set of features and the expression of the set in a form suitable for their algorithms. In this study, we discuss a natural way to treat two typical expressions of a sample: a set of points in a high-dimensional space and a time series of points in a high-dimensional space.

* Corresponding author. Tel.: +81-11-706-6852; fax: +81-11-706-6852.

E-mail address: mine@main.eng.hokudai.ac.jp (M. Kudo)

PII: S0167-8655(99)00077-X

Classification of curves is very important in the fields of analysis of several natural signals, time-series like speech and biological signals (see Shiavi and Bourne, 1986). Many methods for classification of curves are based on Fourier analysis, a prediction method like the auto-regression model (e.g., Wu, 1994) and a hidden Markov model (see Rabiner and Juang, 1993). Particularly for analysis of time-series, many methods choose one source from a possible finite set of signal sources (e.g., Plataniotis et al., 1996; Kehagias and Petridis, 1997).

However, no approach gives an intuitive explanation of the resulting classification rule. In this study, we assume that the form of each curve is the most important cue for classification, and we try to find the regions that all or almost all curves of a class pass through and no curve of the other class does. Such regions are easy to view and interpret.

2. Rules to be found

First of all, we will present an outline of the proposed methods.

Assume that a sample $\boldsymbol{x}$ is a set of points in an n-dimensional space,

$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_m\}, \quad x_i(i = 1, \ldots, m) \in \mathbb{R}^n,$$

where m is variable depending on the sample. The expression is basically independent of the order of the elements.

We find a rule that, for example, requires k or more points to take a value greater than a threshold θ in some feature (dimension), or prohibits more than k points to take a value less than θ' in some feature. An example is shown in Fig. 1. Method A shows a rule requiring two or more points to have a larger value than a threshold (a solid arrow) and no point to have smaller value than another threshold (a dotted arrow), both cases in the ith feature f_i $(i = 1, \ldots, n)$. Method B shows the same rule, but an interval is used instead of a threshold. Method C finds a hyper-rectangle in the n-dimensional space instead of a set of thresholds or intervals defined in each feature. In Fig. 1, one hyper-rectangle (solid rectangle) requires that two or more points exist inside the area of the rectangle and another hyper-rectangle (dotted rectangle) requires that no point exists in the area of the rectangle.

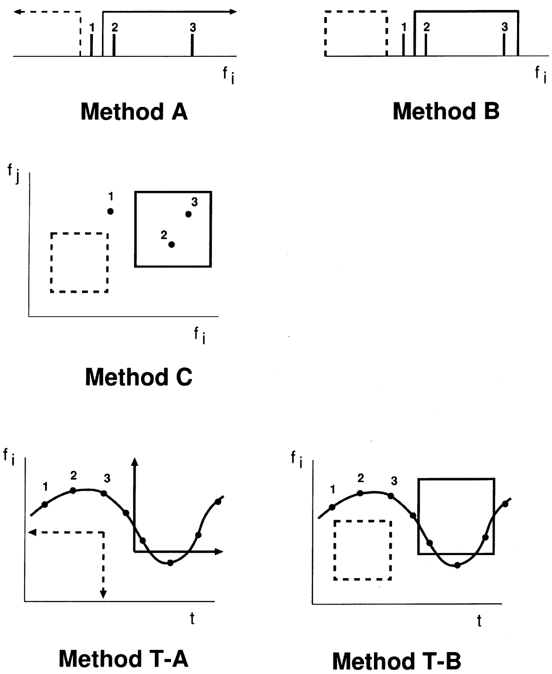

Fig. 1. Three methods for classification of sets of a variable number of points and two methods for classification of time series. One sample consisting of three points is shown for methods A, B and C, and one curve of points is shown for methods T-A and T-B.

Next, let us consider a curve in n-dimensional space. Let us assume that a curve is a discrete time series

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_T), \quad x_t(t = 1, \ldots, T) \in \mathbb{R}^n,$$

where the length T depends on the curve. A parameter t is referred to as time. This case can be regarded as a special case of the above, that is, the case that a set of points are ordered by a time index. Thus, the above-mentioned methods A, B and C are also applicable to this case by ignoring the order of appearance. For this special case, two more methods are examined (Fig. 1). In the relationship between t and each feature f_i $(i = 1, \ldots, n)$, method T-A shows a rule requiring that two or more points exist in the right-upper region (solid arrows) and no point exists in the

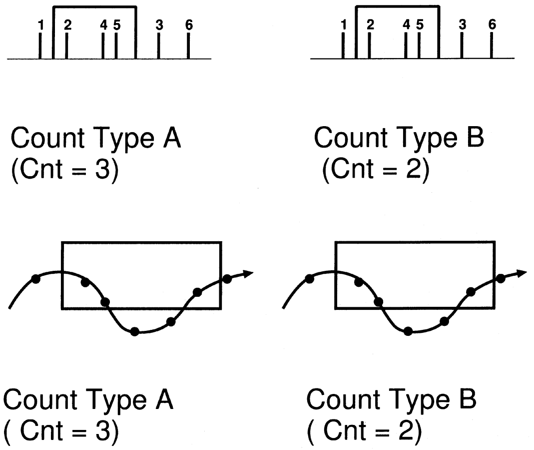

Fig. 2. Two types of counting.

left-lower region (dotted arrows). Right-lower and left-upper regions will be considered also in practice. Method T-B shows a rule using two rectangles with the same requirement.

In addition, we use two ways to count the number of points itself in a region (Fig. 2). One way is to count the number of points and the other way is to count how many times a sequence of points enters the region. Usually, the first way (Count Type A) is used for methods A, B and C and the second way (Count Type B) is for methods T-A and T-B.

3. Realization of rules

In this section, we describe our methodology to realize the above idea. We pay attention only to curves, that is, time series. A set of points can be treated in a similar way. Let us assume a training set $S_\omega = \{\boldsymbol{x}\}$ of N_ω curves for class $\omega \in \Omega$.

We find a set $\mathscr{R}_\omega = \{R\}$ of regions R (it can be a half-interval, an interval, or a hyper-rectangle depending on the method used) for which the maximum number of curves of S_ω pass through and no curve of $S_\mu \in \Omega$ $(\mu \neq \omega)$ passes through. The details will be given in the following section.

In methods T-A and T-B, a region is taken as a rectangle in a plane spanned by each dimension and time. For the ith $(i = 1, \ldots, n)$ feature (dimension), a rectangle R is specified by $R = [t_i, T_i] \times [m_i, M_i]$, where t_i and T_i are the two ends (can be $-\infty$ or ∞) in time, and m_i and M_i are the two ends in the ith dimension.

It can happen that n rectangles defined over different dimensions coincidently have the same time interval: $t_1 = t_2 = \cdots = t_n$, $T_1 = T_2 = \cdots = T_n$. In this case, the set of these rectangles corresponds to a hyper-rectangle defined in the original n-dimensional space. However, this does not usually happen.

Similarly, a set of regions which any curve of S_ω should not pass through is denoted by $\bar{\mathscr{R}}_i = \{\bar{R}\}$. As an extension, we consider how many times a curve passes through the region. By $R^{(k)}$, we denote a region which curves of S_ω must pass through at least k times. Also, by $\bar{R}^{(k)}$, we denote a region which these curves must not pass through k times or more. Now, R and $\bar{R}$ are denoted by $R^{(1)}$ and $\bar{R}^{(1)}$, respectively. We use the notations of $\mathscr{R}_\omega^{(K)} = \{R^{(k)}\ (1 \leqslant k \leqslant K)\}$ ($R^{(k)}$ can be more than one) and $\bar{\mathscr{R}}_\omega^{(K)} = \{\bar{R}^{(k)}\ (1 \leqslant k \leqslant K)\}$. An example is shown in Fig. 3. The curves of class ω must pass through $R^{(k)}$ k times or more and must not pass through $\bar{R}^{(k)}$ k times or more.

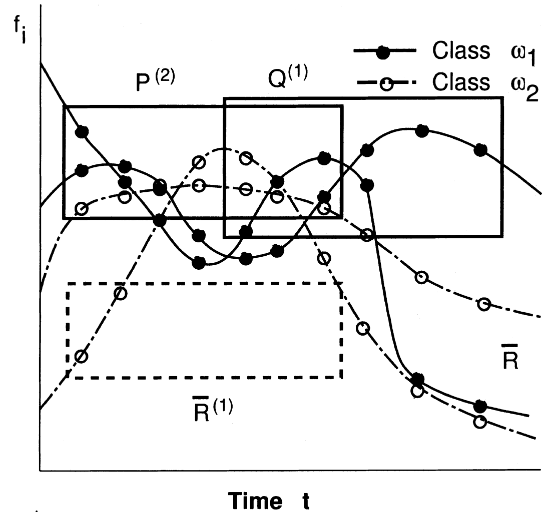

Fig. 3. Three rectangular regions specifying class ω_1. The curve of ω_1 should pass through rectangle $P^{(2)}$ twice or more and rectangle $Q^{(1)}$ at least once, but should not pass through $\bar{R}^{(1)}$ even once.

As shown in the next section, lastly, we find a set of rules $\Phi_\omega = \{\mathscr{R}_\omega^{(K)} \cup \bar{\mathscr{R}}_\omega^{(K)}\}$ and a rule is of an element $\mathscr{R}_\omega^{(K)} \cup \bar{\mathscr{R}}_\omega^{(K)}$.

4. Subclass method

To realize this idea, we use a subclass method (Kudo and Shimbo, 1989). In the subclass method, every pattern is assumed to be a binary pattern. We find a family $\mathscr{X}$ of subsets X, called "subclasses", of a training set S^+ of a class satisfying the following conditions:

1. Let $\boldsymbol{X}$ be the binary pattern obtained by AND operation over all elements of X in each bit; then at least one bit of 1 in $\boldsymbol{X}$ is 0 in any sample of S^-, another training set consisting of the samples of the other classes. (Exclusiveness)
2. X is maximal with respect to the inclusion relation in the set of subsets satisfying the above condition. (Maximalness)

For example, let us consider the following data:

	Sample	Features
S^+	$\boldsymbol{x}_1$	00011001110011100011
	$\boldsymbol{x}_2$	00001011110001100111
	$\boldsymbol{x}_3$	00111000110000101111
	$\boldsymbol{x}_4$	00111000110001100111
	$\boldsymbol{x}_5$	01111000010001100111
	$\boldsymbol{x}_6$	00111000110111100001
S^-	$\boldsymbol{y}_1$	00001011110011100011
	$\boldsymbol{y}_2$	00001011110000101111
	$\boldsymbol{y}_3$	11111000000000101111
	$\boldsymbol{y}_4$	01111000010011100011
	$\boldsymbol{y}_5$	00011001111111100000

The subclass method finds two subclasses:

$$X_1 = \{\boldsymbol{x}_2, \boldsymbol{x}_4, \boldsymbol{x}_5\} \quad \text{and} \quad X_2 = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_6\},$$

and their corresponding binary vectors $\boldsymbol{X}_1, \boldsymbol{X}_2$ are

$$\boldsymbol{X}_1 = 00001000010001100111,$$
$$\boldsymbol{X}_2 = 00011000110000100001.$$

The position of 1's shows the commonness of the positive class. Due to its high computational cost, a semi-optimal set of subclasses is usually found by a randomized version (Kudo et al., 1996) of the subclass method. The most desirable situation is that where only one subclass, S^+ itself, is found; then the class can be thought of as "being totally isolated from the other classes".

Before using this method for a certain application, a binning method has to be carried out. In the binning method, a bit corresponds to a characteristic to be examined; in our study, a characteristic is associated with $R^{(k)}$ or $\bar{R}^{(k)}$. Once the original feature vectors have been converted into binary vectors, we can know which characteristics are important for distinguishing one class from others, by the position of 1's remaining in a subclass $\boldsymbol{X}$.

5. Binning

Taking method T-B as an example, we explain our binning method. A similar procedure is applied to the remaining methods.

We consider a 1-dimensional curve defined over discrete time t,

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_T), \quad x_t \in \mathbb{R}.$$

Let us assume that class ω is a positive class and let the training set S_ω be referred to as a positive sample set S^+ and all other training sets be combined into a negative sample set S^-.

We set up equally spaced U thresholds θ_u $(u = 0, \ldots, U-1)$ in time, whose ends are located at 0 and the maximum length of curves in $S^+ \cup S^-$. Similarly, in each feature, equally spaced V thresholds ϕ_v $(v = 0, \ldots, V-1)$ are set up such that both ends are located at the minimum and maximum values of x_t of every curve $\boldsymbol{x} \in S^+ \cup S^-$.

The binning is carried out as follows. A curve $\boldsymbol{x} \in S^+$ is transformed into a binary vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_B)$ by the following procedure. For all possible $\binom{U}{2} \cdot \binom{V}{2}$ rectangles in the discrete domain, let $b_{uu'vv'}^{(k)} = 1$, if $\boldsymbol{x} \in S^+$ passes through $R_{uu'vv'} = [\theta_u, \theta_{u'}] \times [\phi_v, \phi_{v'}]$ more than or equal to k times $(k = 1, \ldots, K)$, otherwise $b_{uu'vv'}^{(k)} = 0$ (Fig. 4). In addition, all of the inverted bits are added to the vector because 0's have the same importance as 1's. By multiplying by the dimensionality n, the total length in bits becomes

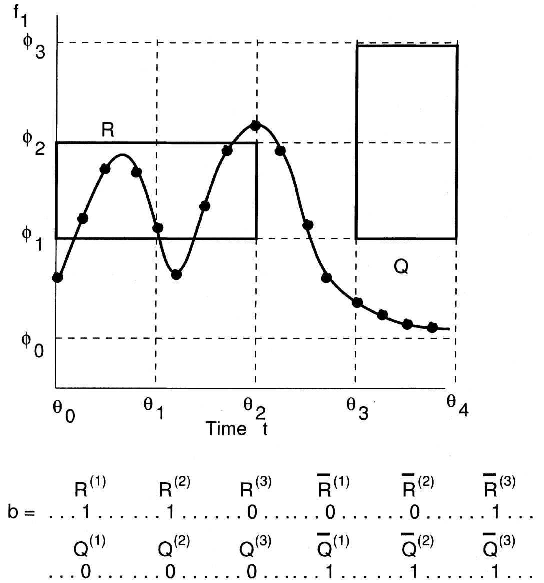

Fig. 4. Bits related to $K = 3$ for R and Q.

$$B = 2nK\binom{U}{2}\binom{V}{2}.$$

The bit length of each method is shown in Table 1.

After this binning, the subclass method is applied to the converted binary vectors. It should be noted that a subclass indicates many regions simultaneously because a bit of 1 corresponds to one region and many bits of 1 can exist in one subclass, and furthermore a class is characterized by a set of subclasses.

For classification of a novel curve $\boldsymbol{x}$ without a class label:

1. First we count the number of subclasses in each class for which $\boldsymbol{x}$ satisfies its corresponding rectangle rules. We choose the class for which the largest percentage of subclasses is satisfied and assign $\boldsymbol{x}$ to that class.
2. If no such subclass is found, we use the average of the nearness of subclasses to $\boldsymbol{x}$ in each class, where "nearness" is measured by the number of bits that is 1 in the subclass $\boldsymbol{X}$ and 0 in $\boldsymbol{x}$. The nearest class is chosen for the assignment.

6. Experiments

6.1. Vowel recognition by formants

This dataset is for recognition of the five Japanese vowels $\{/a/, /e/, /i/, /o/, /u/\}$. In total, 6000 vowel samples were taken from the database ETL-WD-I-1 (Hayamizu et al., 1985). We divided all of the samples into two sets: a training set consisting of 1000 samples (200 for each vowel), and a test set of 5000 samples (1000 for each vowel). Through frequency analysis, peak frequencies [Hz] characterizing each vowel, called "formants", are calculated for a sample. In this analysis, some false formants can appear because six peak frequencies are forcibly calculated regardless of their actual existence. We removed such false formants by referring to their widths, i.e., by removing formants with a width larger than a reasonable value. As a result, samples are expressed by four to six formants.

We applied method A with $V = 200$ and $K = 3$. In total, 12, 29, 27, 19 and 31 subclasses were respectively obtained for five classes.

The largest subclass of class /a/ is shown in Fig. 5. In this figure, for example, the dotted line in the left direction at level 1 shows the region for which formants are prohibited to exits, and the solid line in the left direction at level 1 shows the region for which at least one formants should exit. In fact, all ten sets of points of /a/ satisfy these requirements and have therefore been recognized correctly. On the other hand, for example, nine of ten samples of /e/ violate this subclass rule, because their first formants appear in the prohibited region, and the remaining one sample violates one of the limitations of the rule.

Table 1
Required bit length (time has U thresholds and every feature has V thresholds)

Method A	Method B	Method C	Method T-A	Method T-B
$2nKV$	$2nK\binom{V}{2}$	$2K\binom{V}{2}^n$	$8nKUV$	$2nK\binom{U}{2}\binom{V}{2}$

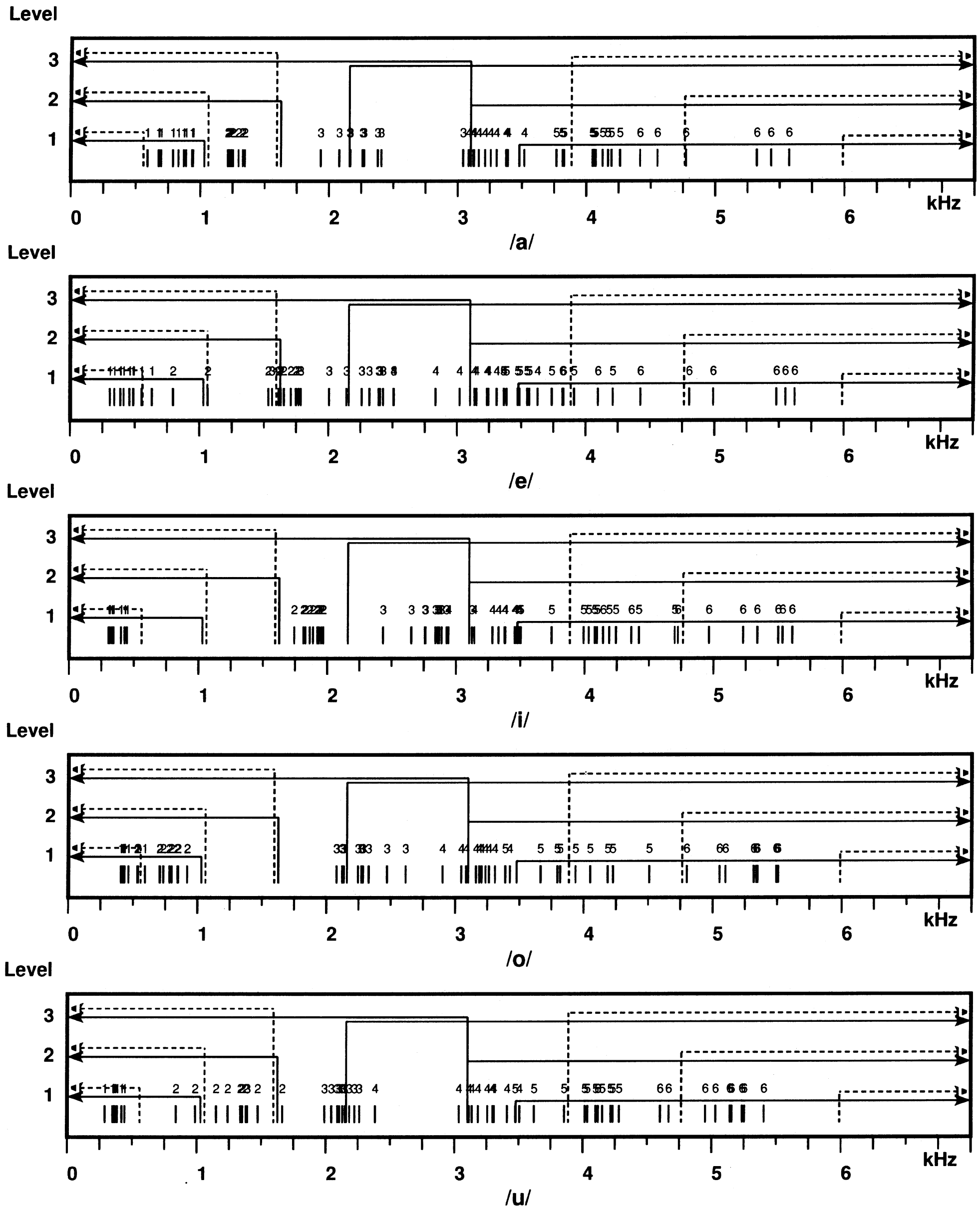

Fig. 5. A subclass obtained for vowel /a/ using method A. Ten samples in each class are shown. The same subclass is shown in all figures. Small numbers with a bar show formants: 1 is the 1st formant, 2 is the 2nd formant, and so on. Solid lines at level k $(R^{(k)})$ show that at least k formants should be in the range, and dotted lines at level k $(\bar{R}^{(k)})$ show that only fewer than k formants are permitted.

Method B was also applied to the same dataset. The number of subclasses obtained were 3, 6, 13, 8 and 16, respectively. The recognition rates of methods A and B are shown in Table 2 together with those of some conventional classifiers (a linear classifier, a quadratic classifier, 5-nearest neighbor classifier) for which six forcibly calculated formants are used.

From Table 2 we can see that method A has the highest recognition rate. One possible explanation for the reason why method A worked better than method B is that method B found too much redundant commonness; in other words, some sort of over-fitting to the training samples was done.

6.2. Polygon recognition

Next, we dealt with the problem of distinguishing three kinds of polygons with 3–5 vertices: (1) polygons whose vertices are all located in a circle with a radius of 0.5, (2) polygons for which

Table 2
Results of recognition of five Japanese vowels

	Classifier				
	Linear	Quadratic	5-NN	Method A	Method B
Recognition rate (%)	82.9	85.3	85.8	86.9	84.2

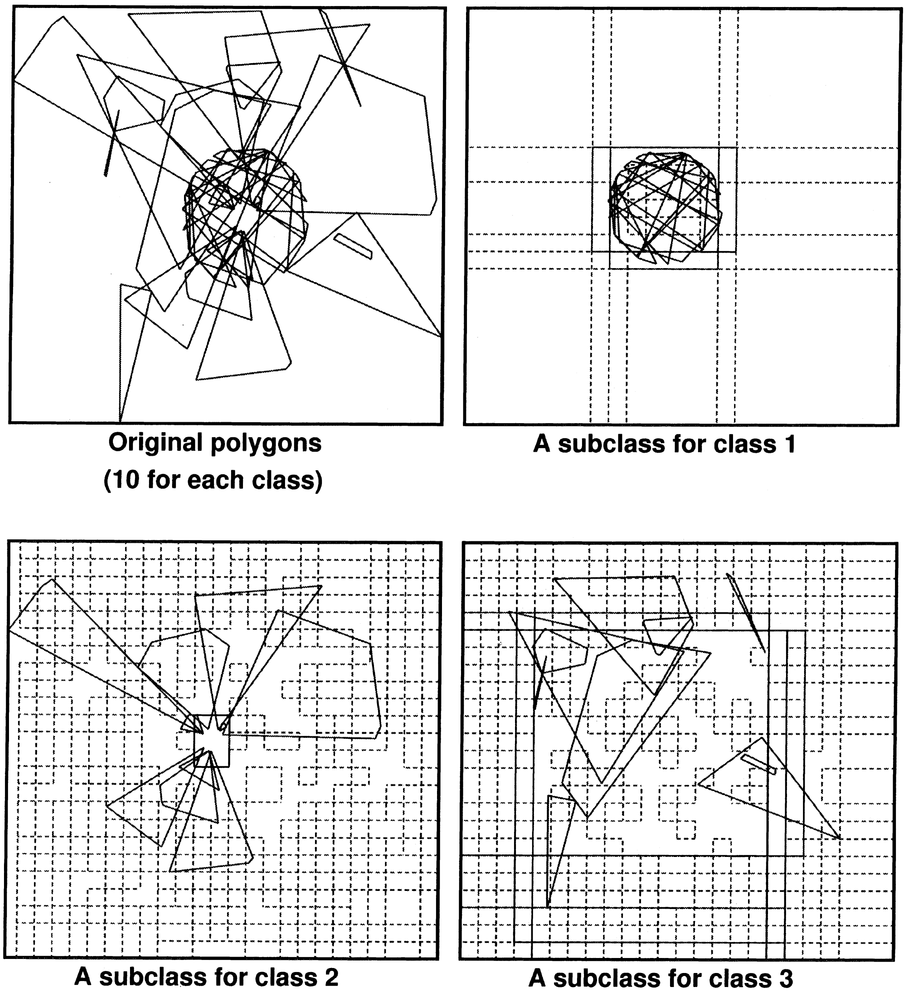

Fig. 6. Results of polygon classification. One subclass and ten samples for each class are shown.

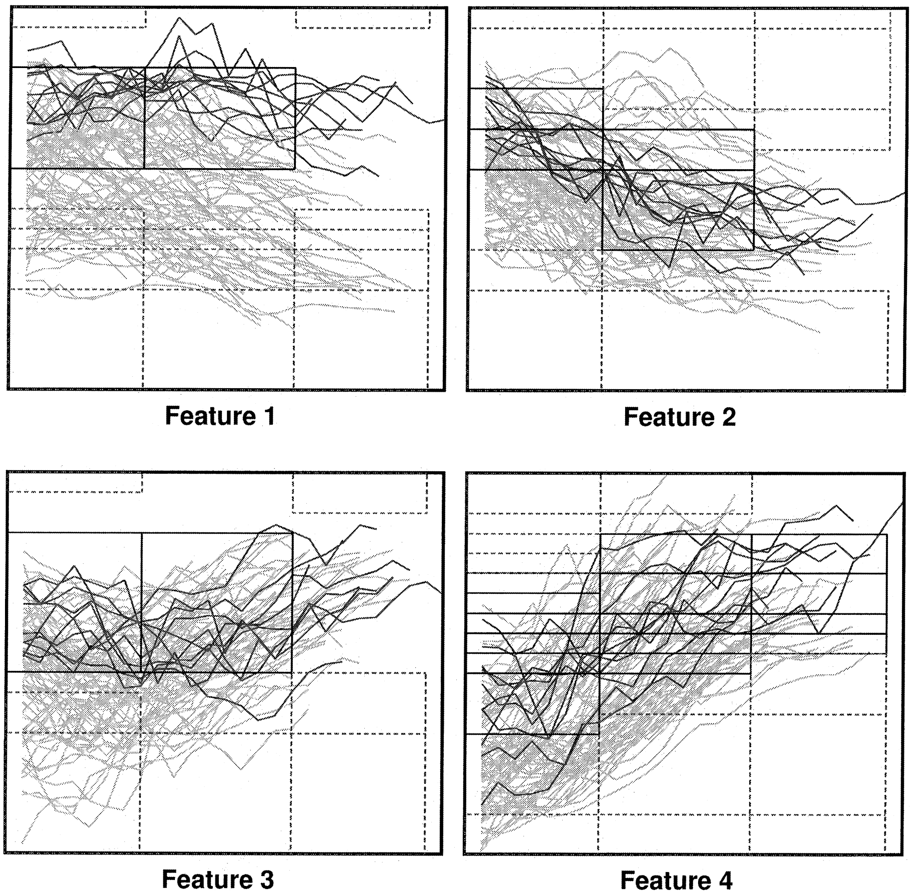

Fig. 7. A subclass for the first speaker in the first four features. The subclass and ten curves of each class (dark lines for the first speaker and light lines for the other eight speakers) are shown. The horizontal axis expresses time and the vertical axes feature values. The solid rectangles show the regions $R^{(1)}$ through which curves of the first speaker should pass, and the dotted rectangles are the regions $\bar{R}^{(1)}$ through which curves of the first speaker should not pass.

one vertex is located in a circle with a radius of 0.1, and (3) polygons whose vertices are located at arbitrary positions. A hundred samples for each class were used.

Method C ($V = 25, K = 1$) was applied to this problem. The result is shown in Fig. 6. The numbers of subclasses are 1, 3 and 9. This is reasonable because the last class is uniform and some instances of the class can belong to the first or second class. As a result, we cannot specify the last class in a simple way.

From Fig. 6, it can be seen that the subclass rules for class ω_1, using the dotted rectangles, limits the range where polygons of the class can exist to a small region with radius of 0.5. Similarly, the subclass of class ω_2 succeeded to find a rule requiring that all polygons of the class should have at least one vertex located in a circle with radius of 0.1 (a solid rectangle). However, we can see also that there is much redundant commonness which is not general in the class but is common among training samples of the class. Many dotted regions of classes ω_2 and ω_3 in Fig. 6 are such examples.

6.3. Speaker recognition

This dataset aims to distinguish two Japanese vowels /ae/ continuously uttered by nine male speakers (a 9-class problem). By 12° linear prediction analysis, 12 LPC cepstrum coefficients were calculated from each frame of the data. As a result, a sample is a time series with length 7–29 in a 12-dimensional space. A total of 270 time series (30 per class) were used for training and the remaining 370 time series were used for testing.

Method T-B with $(U = 4, V = 20, K = 1)$ was applied to this dataset. A subclass for the first speaker is shown in Fig. 7. The recognition rate was 94.1%. For comparison, the 5-state continuous Hidden Markov Model (HMM) (Rabiner and Juang, 1993) was also used. The recognition rate using HMM was 96.2%, which is higher than that using method T-B. This might be because a variance of samples within a class appears as a change in the global form of curves and does not meet the conditions of our technique, which is more effective when the most important cue to distinguish one class from the others lies in a certain local part.

7. Discussion

Some methods were proposed for the recognition of sets of a variable number of points such as figures in a plane, and for the recognition of curves over discrete time such as speech time series. It is very difficult to determine which of the proposed methods is the best for a given problem. Methods C and T-B are good for finding many characteristics that only samples of one class have in common, but sometimes they find also redundant commonness, especially when only a small number of training samples is available. Such redundant commonness is harmful to classification of novel samples. Although other simpler methods such as methods A, B and T-A, are not so good for this purpose, fewer redundant common characteristics are obtained by these methods. As a result, we have to choose a method whose complexity is neither too small nor too large. Some model selection criteria like MDL (Rissanen, 1989) may be useful for choosing an appropriate method. Further study is needed to establish effective ways to choose an appropriate method for a given problem.

8. Conclusion

We proposed a methodology for classifying sets of data points in a multidimensional space or multidimensional curves. The classification rule is based on the common regions through which only curves of one class pass. Based on the results of experiments, this approach was confirmed to be applicable for a wide range of applications and to work well for certain applications.

References

Hayamizu, S. et al., 1985. Generation of VCV/CVC balanced word sets for a speech data base. Bul. Electrotech. Lab. 49 (10) 803–834 (in Japanese).

Kehagias, A., Petridis, V., 1997. Predictive modular neural networks for time series classification. Neural Networks 10, 31–49.

Kudo, M., Shimbo, M., 1989. Optimal subclasses with dichotomous variables for feature selection and discrimination. IEEE Trans. Systems, Man and Cybern. 19, 1194–1199.

Kudo, M., Yanagi, S., Shimbo, M., 1996. Construction of class regions by a randomized algorithm: a randomized subclass method. Pattern Recognition 29, 581–588.

Plataniotis, K.N., Androutsos, D., Venetsanopoulos, A.N., Lainiotis, D.G., 1996. A new time series classification approach. Signal Process. 54, 191–199.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Chiffs, NJ.

Rissanen, J., 1989. Stochastic Complexity in Statistical Inquiry. Volume 15 of Comp. Sci., World Scientific, Singapore, 1989.

Shiavi, R.G., Bourne, J.R., 1986. Methods of biological signal processing. In: Young, T.Y., Fu, K.S., (Eds.), Handbook of Pattern Recognition and Image Processing. Academic Press, New York, Chapter 22.

Wu, B., 1994. Pattern recognition and classification in time series analysis. Appl. Math. Comput. 62, 29–45.