Pergamon

Pattern Recognition, Vol. 29, No. 11, pp. 1919–1930, 1996

0031–3203/96 $15.00 + .00

PII: S0031–3203(96)00033–7

REAL-TIME AND AUTOMATIC TWO-CLASS CLUSTERING BY ANALYTICAL FORMULAS†

JA-CHEN LIN* and WU-JA LIN

Department of Computer and Information Science, National Chiao-Tung University, Hsinchu, Taiwan 30050, Republic of China

(*Received 5 October* 1995; *received for publication 28 February* 1996)

Abstract—Several feature-preserving two-class clustering methods are investigated in this paper. By preserving certain features of the input data, some formulas useful in calculating the two class representatives and population percentages are derived. The derived formulas are expressed in general forms suitable for any dimensionality higher than two. The complexities of the investigated methods are all of order N if the data size is N and hence are much faster than any other clustering method which uses $N \times N$ dissimilarity matrix. Additionally, all investigated methods use no initial guesses. Experimental results are included to make a comparison among the four investigated methods so that only two methods are recommended. Further comparisons with the k-means method and hierarchical clustering methods also are included. The proposed feature-preserving approach was found to be fast, automatic and suitable for any field requiring fast high-dimensional two-class clustering. Copyright © 1996 Pattern Recognition Society. Published by Elsevier Science Ltd.

Two-class clustering   High-dimensional space   Feature-preserving
Analytical fast clustering   General form k-means   Hierarchical agglomerative clustering

1. INTRODUCTION

Clustering is a data analysis technique that has seen wide application in many fields such as image segmentation,[1,2] image registration,[3,4] computer vision,[5] psychiatry[6] and politics.[7] A special branch of clustering methodology concerns two-class clustering techniques. These techniques can be applied in BTC image compression[8,9] and binary decision tree construction,[7] among others. In order to meet the practicability requirements among these applications, these techniques are expected to perform fast automatic two-class clustering. However, most of the existing clustering methods, such as the k-means method and the hierarchically methods using dissimilarity matrix, are either nonautomatic or time-consuming. Therefore, it is desirable to develop a real-time method which can partition input data into two classes automatically. The major concern of the design introduced in this paper thus focuses on two basic goals: *real-time* and *automation*.

In this paper we design and compare four new analytical methods that use analytical formulas to evaluate the two-class representatives and population percentages directly. We begin with three dimensions and then extend the derived formulas to include higher dimensions. All the formulas were obtained by solving equations that confined the class representatives and population percentages to preserve certain features of input data. For 3-D (three-dimensional) data, our approach not only preserves $\{p, \bar{x}, \bar{y}, \bar{z}\}$, where p denotes the population and $(\bar{x}, \bar{y}, \bar{z})$ denotes the data centroid, but also the features combined in one of the following four ways: $\{|\overline{xy}|, |\overline{xz}|, |\overline{yz}|, |\overline{xyz}|\}$, $\{\overline{xy}, \overline{xz}, \overline{yz}, \overline{xyz}\}$, $\{\overline{x^2}, \overline{y^2}, \overline{z^2}, \bar{r}\}$, $\{\overline{x^2}, \overline{y^2}, \overline{z^2}, \overline{xyz}\}$. (Throughout this paper, the "bar" always denotes the average function. For example, $|\overline{xy}| = \frac{1}{N}\sum_{k=1}^{N}|x_k y_k|$.) All of these features are easily calculated. Note that although readers might want to try many other features, for example, $\{\overline{|x|+|y|+|z|}, \overline{x^2 y}, \overline{\sin x}, \overline{\log z}, \ldots\}$, we found that many combinations do not yield reasonable experimental results and some combinations even yield no solution for almost every data set. For example, when we preserved $\{p; \bar{x}, \bar{y}, \bar{z}; \overline{x^2}, \overline{y^2}, \overline{z^2}; \overline{|x|+|y|+|z|}\}$, we found that (after certain derivations) preserving these features will constrain the input data to satisfy

$$\sqrt{\overline{x^2}\,\overline{y^2}} + \sqrt{\overline{x^2}\,\overline{z^2}} + \sqrt{\overline{y^2}\,\overline{z^2}} = \overline{|xy| + |xz| + |yz|}$$

which, of course, is not necessarily true for an arbitrarily given data set. After examining (and experimenting with) many methods that preserve different features, we introduce in this paper only four of them, namely, Methods 1–4.

The remainder of this paper is organized as follows. In Section 2, the 3-D case is investigated and the

† This study was supported by the National Science Council, Republic of China, under contract number: NSC.82-0408-E-009-314.

* Author to whom correspondence should be addressed.

corresponding formulas for each of the four methods are listed. In Section 3 these four methods are generalized to handle data with more dimensions. In Section 4 some experimental results are shown. In Section 5 some comparisons with the k-means and the hierarchical clustering methods are illustrated. Concluding remarks are presented in Section 6. Finally, a mathematical proof is collected in the Appendix to simplify the reading.

2. TWO-CLASS FEATURE-PRESERVING CLUSTERING FOR 3-D DATA

In this section the four feature-preserving two-class clustering methods for 3-D data are introduced. We first derive the formulas for Method 1. Let $H=\{(x_k, y_k, z_k)\}_{k=1}^{N}$ be a given data set that is to be clustered into two classes A and B. Let (x_A, y_A, z_A) and (x_B, y_B, z_B) denote the two corresponding class representatives. Similarly, let the population percentages be p_A and p_B, respectively. To solve for the eight unknowns:

$$\{p_A, x_A, y_A, z_A, p_B, x_B, y_B, z_B\}, \tag{1}$$

we need eight equations. The first equation we may use is the natural requirement:

$$p_A + p_B = 100\% = 1. \tag{2}$$

Also note that in order to avoid having the estimated class representatives be far away from the given data, we should also preserve the position of the centroid $(\bar{x}, \bar{y}, \bar{z})$ of the given N-point data set. That is, we should require that:

$$p_A x_A + p_B x_B = \frac{1}{N}\sum_{k=1}^{N} x_k = \bar{x} = 0, \tag{3}$$

$$p_A y_A + p_B y_B = \frac{1}{N}\sum_{k=1}^{N} y_k = \bar{y} = 0, \tag{4}$$

$$p_A z_A + p_B z_B = \frac{1}{N}\sum_{k=1}^{N} z_k = \bar{z} = 0. \tag{5}$$

(For simplicity, we always assume that the centroid of the given data set is the origin. If this is not the case, we just do a preprocessing step to translate all N data points to meet this assumption.) In addition to the four basic equations listed above, we still need $8-4=4$ more equations. We therefore try to preserve four other features of the given data. Here we use $\{\overline{|xy|}, \overline{|xz|}, \overline{|yz|}, \overline{|xyz|}\}$ for Method 1. Stated more completely, the remaining four equations are:

$$p_A|x_A y_A| + p_B|x_B y_B| = \overline{|xy|}, \tag{6}$$

$$p_A|x_A z_A| + p_B|x_B z_B| = \overline{|xz|}, \tag{7}$$

$$p_A|y_A z_A| + p_B|y_B z_B| = \overline{|yz|}, \tag{8}$$

$$p_A|x_A y_A z_A| + p_B|x_B y_B z_B| = \overline{|xyz|}. \tag{9}$$

Using the eight equations listed in (2)–(9), we solve for the eight unknowns mentioned in (1). Equations (3)–(5) imply that:

$$x_B = -\frac{p_A}{p_B} x_A, \tag{10}$$

$$y_B = -\frac{p_A}{p_B} y_A, \tag{11}$$

$$z_B = -\frac{p_A}{p_B} z_A. \tag{12}$$

Equations (10) and (11) together yield:

$$|x_B y_B| = \left(\frac{p_A}{p_B}\right)^2 |x_A y_A|,$$

and hence

$$p_B|x_B y_B| = \frac{p_A^2}{p_B}|x_A y_A|.$$

As a result:

$$\begin{aligned} p_A|x_A y_A| + p_B|x_B y_B| &= \left(p_A + \frac{p_A^2}{p_B}\right)|x_A y_A| \\ &= \frac{p_A(p_B + p_A)}{p_B}|x_A y_A| \\ &= \frac{p_A}{p_B}|x_A y_A| \end{aligned}$$

by equation (2). Therefore, equation (6) implies:

$$\frac{p_A}{p_B}|x_A y_A| = \overline{|xy|}. \tag{13}$$

By symmetry, we also have:

$$\frac{p_A}{p_B}|x_A z_A| = \overline{|xz|}, \tag{14}$$

$$\frac{p_A}{p_B}|y_A z_A| = \overline{|yz|}. \tag{15}$$

Cyclically multiplying any two of equations (13), (14) and (15), then dividing the product by the remaining one, we can finally obtain the analytical formula of class A's representative (x_A, y_A, z_A), namely:

$$x_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{|xy|}\,\overline{|xz|}}{\overline{|yz|}}}, \tag{16}$$

$$y_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{|xy|}\,\overline{|yz|}}{\overline{|xz|}}}, \tag{17}$$

$$z_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{|xz|}\,\overline{|yz|}}{\overline{|xy|}}}. \tag{18}$$

Once (x_A, y_A, z_A) is known, the representative (x_B, y_B, z_B) of class B can then be evaluated using equations (10)–(12). The remaining problem is therefore to determine the signs of $\{x_A, y_A, z_A\}$ and the values of p_A and p_B. To evaluate the values of p_A and p_B, note that equation (2) indicates that:

$$p_B = 1 - p_A. \tag{19}$$

Substituting equations (10)–(12) and (16)–(19) in (9), we obtain (see the Appendix for details):

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\sqrt{\delta^2 + 8\delta} - \delta - 2}{2}}, \tag{20}$$

where

$$\delta = \frac{\overline{|xyz|^2}}{\overline{|xy|}\,\overline{|xz|}\,\overline{|yz|}}. \tag{21}$$

Below we show how to determine the signs of $\{x_A, y_A, z_A, x_B, y_B, z_B\}$. Assume there are at least as many points in class A as there are in class B, that is, $p_A \geq p_B$. Then:

$$p_A = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{\sqrt{\delta^2 + 8\delta} - \delta - 2}{2}}. \tag{22}$$

To determine the signs of x_A, y_A, z_A, x_B, y_B and z_B, we proceed as follows: By equations (10)–(12), there are only four possible cases for the signs of (x_A, y_A, z_A) and (x_B, y_B, z_B), namely, Case I $((+,+,+)$ and $(-,-,-))$; Case II $((+,+,-)$ and $(-,-,+))$; Case III $((+,-,+)$ and $(-,+,-))$; case IV $((-,+,+)$ and $(+,-,-))$. Here, "+" indicates "≥ 0," and "−" indicates "≤ 0." Also note that each case contains two subcases, e.g. Case I includes not only the subcase "(x_A, y_A, z_A) being $(+,+,+)$ and (x_B, y_B, z_B) being $(-,-,-)$", but also "(x_A, y_A, z_A) being $(-,-,-)$ and (x_B, y_B, z_B) being $(+,+,+)$". To determine the signs we only have to inspect the data to check which case is most likely to happen. For example, if 52 points have been checked, and if (x_{53}, y_{53}, z_{53}) is (75, 43, −9), then Case II receives one more point, etc. When all N data points are checked, the case receiving the highest score gives the signs of (x_A, y_A, z_A) and (x_B, y_B, z_B). Assume Case II is the case. Then, if $(+,+,-)$ occurs more often than $(-,-,+)$ in the N data points, we set (x_A, y_A, z_A) to be $(+,+,-)$ and (x_B, y_B, z_B) to be $(-,-,+)$.

Having obtained the two class representatives (x_A, y_A, z_A) and (x_B, y_B, z_B), the decision boundary that separates the input data into two classes can be defined as a plane perpendicular to the line segment $\overline{AB}$ connecting $A = (x_A, y_A, z_A)$ and $B = (x_B, y_B, z_B)$ and cutting the 3-D space such that Np_A points fall in the halfspace containing A. However, if we notice that p_A and p_B are just some estimated values, and keep in mind the real-time requirement that we pointed out in the first paragraph of the Introduction, then a simpler and faster way to cluster data is to use the nearest-neighbor rule: a data point is assigned to class A if and only if the data point is closer to (x_A, y_A, z_A) than to (x_B, y_B, z_B). In other words, just take the decision boundary to be a plane perpendicular to $\overline{AB}$ and passing through its mid-point. Hereafter, we will use this simpler rule for Methods 1–4. As a result, the clustering result (i.e. the job of assigning each data point to a suitable class) is completely determined by the positions of the computed values of the class representatives.

Below, we list the formulas for the other three methods (Methods 2–4). To save space, we omit the proofs and just give the features they preserve and the resulting formulas. Method 1 is also summarized here for comparison.

Method 1. Preserving $\{p, \bar{x}, \bar{y}, \bar{z}, \overline{|xy|}, \overline{|xz|}, \overline{|yz|}, \overline{|xyz|}\}$ to obtain:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\sqrt{\delta^2 + 8\delta} - \delta - 2}{2}}$$

$$\left(\text{where } \delta = \frac{\overline{|xyz|^2}}{\overline{|xy|}\,\overline{|xz|}\,\overline{|yz|}}\right),$$

$$x_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{|xy|}\,\overline{|xz|}}{\overline{|yz|}}},$$

$$y_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{|xy|}\,\overline{|yz|}}{\overline{|xz|}}},$$

$$z_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{|xz|}\,\overline{|yz|}}{\overline{|xy|}}}.$$

Method 2. Preserving $\{p, \bar{x}, \bar{y}, \bar{z}, \overline{xy}, \overline{xz}, \overline{yz}, \overline{xyz}\}$ to obtain:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\overline{xyz}^2}{\overline{xyz}^2 + 4\overline{xy}\,\overline{xz}\,\overline{yz}}},$$

$$x_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{xy}\,\overline{xz}}{\overline{yz}}},$$

$$y_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{xy}\,\overline{yz}}{\overline{xz}}},$$

$$z_A = \pm\sqrt{\frac{p_B}{p_A}\frac{\overline{xz}\,\overline{yz}}{\overline{xy}}}.$$

Method 3. Preserving $\{p, \bar{x}, \bar{y}, \bar{z}, \overline{x^2}, \overline{y^2}, \overline{z^2}, \bar{r}\}$, where $\bar{r} = (\sum_{k=1}^{N}\sqrt{x_k^2 + y_k^2 + z_k^2})/N$, to obtain:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{(\bar{r})^2}{\overline{r^2}}} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{(\bar{r})^2}{\overline{x^2} + \overline{y^2} + \overline{z^2}}}$$

$$x_A = \pm\sqrt{\left(\frac{p_B}{p_A}\overline{x^2}\right)},$$

$$y_A = \pm\sqrt{\left(\frac{p_B}{p_A}\overline{y^2}\right)},$$

$$z_A = \pm\sqrt{\left(\frac{p_B}{p_A}\overline{z^2}\right)}.$$

Method 4. Preserving $\{p, \bar{x}, \bar{y}, \bar{z}, \overline{x^2}, \overline{y^2}, \overline{z^2}, \overline{xyz}\}$ to obtain:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\overline{xyz}^2}{\overline{xyz}^2 + 4\overline{x^2}\,\overline{y^2}\,\overline{z^2}}},$$

$$x_A = \pm\sqrt{\left(\frac{p_B}{p_A}\overline{x^2}\right)},$$

$$y_A = \pm\sqrt{\left(\frac{p_B}{p_A}\overline{y^2}\right)},$$

$$z_A = \pm\sqrt{\left(\frac{p_B}{p_A}\overline{z^2}\right)}.$$

The signs used in these formulas can be determined using the rule discussed earlier. Also note that Method 2 might not have a solution if:

$$\overline{xy}\cdot\overline{xz}\cdot\overline{yz} < 0, \tag{23}$$

because equation (23) will make one of the two population percentages negative. This problem, however, will not occur with Methods 3 and 4 (here $\bar{r}^2 \le \overline{r^2}$ is guaranteed by Schwartz's inequality). Method 1 might have no solution if $\delta < 1$ [see equation (20)]. However, we found that $\delta < 1$ rarely occurs (especially if the data-rotation preprocessing introduced in Section 4 was used first). In the following section we extend the proposed methods to handle data with more dimensions. The generalized formulas are also listed.

3. GENERALIZATION TO d-DIMENSIONAL SPACE

We extend Method 1 first. Assume that the N-point two-class data are d-dimensional and hence can be expressed as $\{(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots, x_d^{(k)})\}_{k=1}^{N}$. Let the population percentages and class representatives of these two classes be $p_A, p_B, (x_1^{(A)}, x_2^{(A)}, \ldots, x_d^{(A)})$ and $(x_1^{(B)}, x_2^{(B)}, \ldots, x_d^{(B)})$, respectively. Therefore, $2d+2$ equations are needed because there are $2d+2$ unknowns to be solved for. As before, we may preserve the input data centroid. If we preserve these d features $\{\overline{x_i}\}_{i=1}^{d}$ and require that $p_A + p_B = 1$, there are still $(2d+2)-(d+1) = d+1$ features to be preserved. The natural extension of Method 1 from 3-D to d-dimensional space is to assign the remaining $d+1$ features to be $\{\Pi_{|x_1|}, \Pi_{|x_2|}, \ldots, \Pi_{|x_d|}, \overline{|x_1x_2x_3\cdots x_d|}\}$. Here, $\Pi_{|x_i|}$ is defined as:

$$\Pi_{|x_i|} = \frac{1}{N}\sum_{k=1}^{N}\left(|x_1^{(k)}x_2^{(k)}\cdots x_d^{(k)}/x_i^{(k)}|\right), \tag{24}$$

a value which is known once the N-point data set is given. Consequently, the set of features that we preserve is $\{p; \overline{x_1}, \ldots, \overline{x_d}; \Pi_{|x_1|}, \ldots, \Pi_{|x_1|}; \overline{|x_1x_2x_3\cdots x_d|}\}$. The resulting $2d+2$ equations are:

$$p_A + p_B = 1; \tag{25}$$

$$\begin{cases} p_A x_1^{(A)} + p_B x_1^{(B)} = \overline{x_1} = 0, \\ p_A x_2^{(A)} + p_B x_2^{(B)} = \overline{x_2} = 0, \\ \quad \vdots = \vdots \\ p_A x_d^{(A)} + p_B x_d^{(B)} = \overline{x_d} = 0; \end{cases} \tag{26}$$

$$\begin{cases} p_A|x_2^{(A)}x_3^{(A)}\cdots x_d^{(A)}| + p_B|x_2^{(B)}x_3^{(B)}\cdots x_d^{(B)}| = \Pi_{|x_1|}, \\ p_A|x_1^{(A)}x_3^{(A)}\cdots x_d^{(A)}| + p_B|x_1^{(B)}x_3^{(B)}\cdots x_d^{(B)}| = \Pi_{|x_2|}, \\ \quad \vdots = \vdots \\ p_A|x_1^{(A)}x_2^{(A)}\cdots x_{d-2}^{(A)}x_d^{(A)}| + p_B|x_1^{(B)}x_2^{(B)}\cdots x_{d-2}^{(B)}x_d^{(B)}| = \Pi_{|x_{d-1}|}, \\ p_A|x_1^{(A)}x_2^{(A)}\cdots x_{d-2}^{(A)}x_{d-1}^{(A)}| + p_B|x_1^{(B)}x_2^{(B)}\cdots x_{d-2}^{(B)}x_{d-1}^{(B)}| = \Pi_{|x_d|}; \end{cases} \tag{27}$$

$$p_A|x_1^{(A)}x_2^{(A)}\cdots x_{d-1}^{(A)}x_d^{(A)}| + p_B|x_1^{(B)}x_2^{(B)}\cdots x_{d-1}^{(B)}x_d^{(B)}| = \overline{|x_1\cdots x_d|}. \tag{28}$$

Note that we have assumed here for simplicity that $\overline{x_1} = \overline{x_2} = \overline{x_3} = \cdots = \overline{x_d} = 0$, just as we did in the previous section. After certain derivations, it can be proved that the generalized formula of class A's representative $(x_1^{(A)}, x_2^{(A)}, \ldots, x_d^{(A)})$ is:

$$\begin{aligned} x_1^{(A)} &= \pm\frac{\alpha}{\beta}\frac{\overline{|x_1\cdots x_d|}}{|x_2x_3\cdots x_d|} = \pm\frac{\alpha}{\beta}\frac{\overline{|x_1\cdots x_d|}}{\Pi_{|x_1|}}, \\ \vdots \quad & \vdots \qquad \vdots \\ x_i^{(A)} &= \pm\frac{\alpha}{\beta}\frac{\overline{|x_1\cdots x_d|}}{|x_1x_2\cdots x_{d-1}|} = \pm\frac{\alpha}{\beta}\frac{\overline{|x_1\cdots x_d|}}{\Pi_{|x_d|}}, \end{aligned} \tag{29}$$

and the population percentages p_A and p_B satisfy:

$$\frac{\overline{|x_1\cdots x_d|}^{d-1}}{\Pi_{|x_1|}\cdots\Pi_{|x_d|}} = \frac{\beta^{d-1}}{\alpha^d} = \frac{\left(p_A + \dfrac{p_A^d}{p_A^{d-1}}\right)^{d-1}}{\left(p_A + \dfrac{p_B^{d-1}}{p_B^{d-2}}\right)^{d}}. \tag{30}$$

Here, α and β are defined as:

$$\alpha = p_A + \frac{p_A^{d-1}}{p_B^{d-2}},$$

$$\beta = p_A + \frac{p_A^d}{p_B^{d-1}}.$$

Note that it is difficult to obtain the population percentages p_A and p_B by solving equation (30) directly, although the leftmost expression in equation (30) is a known value easily obtained from the given data. An alternative strategy is to build up a 50-entry table (see Table 1) of the computed values:

$$\frac{\left(p_A + \dfrac{p_A^d}{p_B^{d-1}}\right)^{d-1}}{\left(p_A + \dfrac{p_A^{d-1}}{p_B^{d-2}}\right)^{d}} \tag{31}$$

for $p_B \in \{1, 2, \ldots, 50\%\}$ (assume that $p_A = 1 - p_B \ge p_B$). Then, according to the value of:

$$\frac{\overline{|x_1\cdots x_d|}^{d-1}}{\Pi_{|x_1|}\cdots\Pi_{|x_d|}}, \tag{32}$$

we can use equation (30) to obtain the corresponding population percentages p_A and p_B easily by looking them up the table. (Due to space limitation, only $d = 9, 10, 11$ are shown in Table 1.)

The obtained values of p_A and p_B can then be used to evaluate α and β, and then substituted into equation

Table 1. The computed values of the left-hand side of equation (33)

p_B	$d=9$	$d=10$	$d=11$
1%	100.00000	100.00000	100.00000
2%	50.000000	50.000000	50.000000
3%	33.333333	33.333333	33.333333
4%	25.000000	25.000000	25.000000
5%	20.000000	20.000000	20.000000
6%	16.666666	16.666666	20.000000
7%	14.285713	14.285714	14.285714
8%	12.499996	12.500000	12.500000
9%	11.111103	11.111110	11.111111
10%	9.999983	9.999998	10.00000
11%	9.090877	9.090905	9.090908
12%	8.333276	8.333325	8.333332
13%	7.692208	7.692291	7.692305
14%	7.142691	7.142827	7.14852
15%	6.666397	6.666614	6.66666
16%	6.249575	6.249910	6.249981
17%	5.881698	5.882204	5.882320
18%	5.554576	5.555315	5.555498
19%	5.261693	5.262777	5.263060
20%	4.997864	4.999409	4.999838
21%	4.758836	4.761002	4.761642
22%	4.541104	4.544096	4.545035
23%	4.341734	4.345812	4.347167
24%	4.158229	4.163718	4.165646
25%	3.988437	3.995735	3.998442
26%	3.830458	3.840052	3.843805
27%	3.682592	3.695063	3.700203
28%	3.543274	3.559309	3.566266
29%	3.411036	3.431433	3.440740
30%	3.284465	3.310128	3.322438
31%	3.162175	3.194104	3.210198
32%	3.042781	3.082044	3.102835
33%	2.924890	2.972572	2.999100
34%	2.807101	2.864230	2.897630
35%	2.688027	2.755457	2.796902
36%	2.566346	2.644603	2.695200
37%	2.440883	2.529971	2.590605
38%	2.310736	2.409904	2.481020
39%	2.175445	2.282981	2.364276
40%	2.035196	2.148229	2.238358
41%	1.891034	2.005511	2.101784
42%	1.745040	1.855914	1.954167
43%	1.600395	1.702127	1.796912
44%	1.461282	1.548636	1.633888
45%	1.332564	1.401588	1.471757
46%	1.219294	1.268198	1.319616
47%	1.126143	1.155796	1.187777
48%	1.056928	1.070791	1.086000
49%	1.014359	1.017925	1.021876
50%	1.000000	1.000000	1.000000

(29) to obtain the representative of class A. The sign of class A's representative in each dimension can be determined using the method described in the previous section. The class representatives of class B can then be evaluated using equation (26). Below, we list the generalized formulas for Methods 2–4. The proofs are omitted to save space. For comparison, the generalized formulas from Method 1 are also included.

Method 1 (d-dimensional). Preserving the features $\{p; \overline{x_1}, \ldots, \overline{x_d};\ \Pi_{|x_1|}, \ldots, \Pi_{|x_d|};\ \overline{|x_1 x_2 \ldots x_{d-1} x_d|}\}$ to obtain:

$$\frac{\left(p_A + \dfrac{p_A^d}{p_B^{d-1}}\right)^{d-1}}{\left(p_A + \dfrac{p_A^{d-1}}{p_B^{d-2}}\right)^{d}} = \frac{\overline{|x_1 \ldots x_d|}^{d-1}}{\Pi_{|x_1|} \cdots \Pi_{|x_d|}}, \tag{33}$$

$$x_i^{(A)} = \pm \frac{\left(p_A + \dfrac{p_A^{d-1}}{p_B^{d-2}}\right)}{\left(p_A + \dfrac{p_A^{d}}{p_B^{d-1}}\right)} \frac{\overline{|x_1 \ldots x_d|}}{\Pi_{|x_i|}}, \quad \forall i = 1, \ldots, d. \tag{34}$$

Method 2 (d-dimensional). Preserving the features $\{p; \bar{x}_1, \ldots, \bar{x}_d;\ \Pi_{x_1}, \ldots, \Pi_{x_d};\ \overline{x_1 x_2 \cdots x_{d-1} x_d}\}$ to obtain:

$$\frac{\left(p_A + (-1)^d \dfrac{p_A^d}{p_B^{d-1}}\right)^{d-1}}{\left(p_A + (-1)^{d-1} \dfrac{p_A^{d-1}}{p_B^{d-2}}\right)^{d}} = \frac{\overline{x_1 \cdots x_d}^{d-1}}{\Pi_{x_1} \cdots \Pi_{x_d}}, \tag{35}$$

$$x_i^{(A)} = \frac{\left(p_A + (-1)^{d-1} \dfrac{p_A^{d-1}}{p_B^{d-2}}\right)}{\left(p_A + (-1)^d \dfrac{p_A^{d}}{p_B^{d-1}}\right)} \frac{\overline{x_1 \cdots x_d}}{\Pi_{x_i}}, \quad \forall i = 1, \ldots, d. \tag{36}$$

Here, Π_{x_i} is defined as in equation (24), except that the absolute value symbols on both sides of the equation are now taken away.

Method 3 (d-dimensional). Preserving the features $\{p; \overline{x_1}, \ldots, \overline{x_d}, \overline{x_1^2}, \ldots, \overline{x_d^2}; \bar{r}\}$ to obtain:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{(\bar{r})^2}{\overline{r^2}}} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{(\bar{r})^2}{\overline{x_1^2} + \cdots + \overline{x_d^2}}}, \tag{37}$$

$$x_i^{(A)} = \pm \sqrt{\frac{p_B}{p_A} \overline{x_i^2}}, \quad \forall i = 1, \ldots, d. \tag{38}$$

Method 4 (d-dimensional). Preserving the features $\{p; \overline{x_1}, \ldots, \overline{x_d}; \overline{x_1^2}, \ldots, \overline{x_d^2};\ \overline{|x_1 x_2 \cdots x_{d-1} x_d}\}$ to obtain:

$$\left[\frac{p_B^{d/2}}{p_A^{(d/2)-1}} + (-1)^d \frac{p_A^{d/2}}{p_B^{(d/2)-1}}\right]^2 = \frac{\overline{x_1 x_2 \cdots x_{d-1} x_d}^2}{\overline{x_1^2}\,\overline{x_2^2} \cdots \overline{x_{d-1}^2}\,\overline{x_d^2}}, \tag{39}$$

$$x_i^{(A)} = \pm \sqrt{\left(\frac{p_B}{p_A} \overline{x_i^2}\right)}, \quad \forall i = 1, \ldots, d. \tag{40}$$

The signs of $x_i^{(A)}$ in equations (34), (38) and (40) can be determined in the manner proposed in Section 2. Furthermore, the p_A in equations (33), (35) and (39) can also be obtained from the (50-entry) table look-up technique just described because the right-hand sides of equations (33), (35) and (39) are known values.

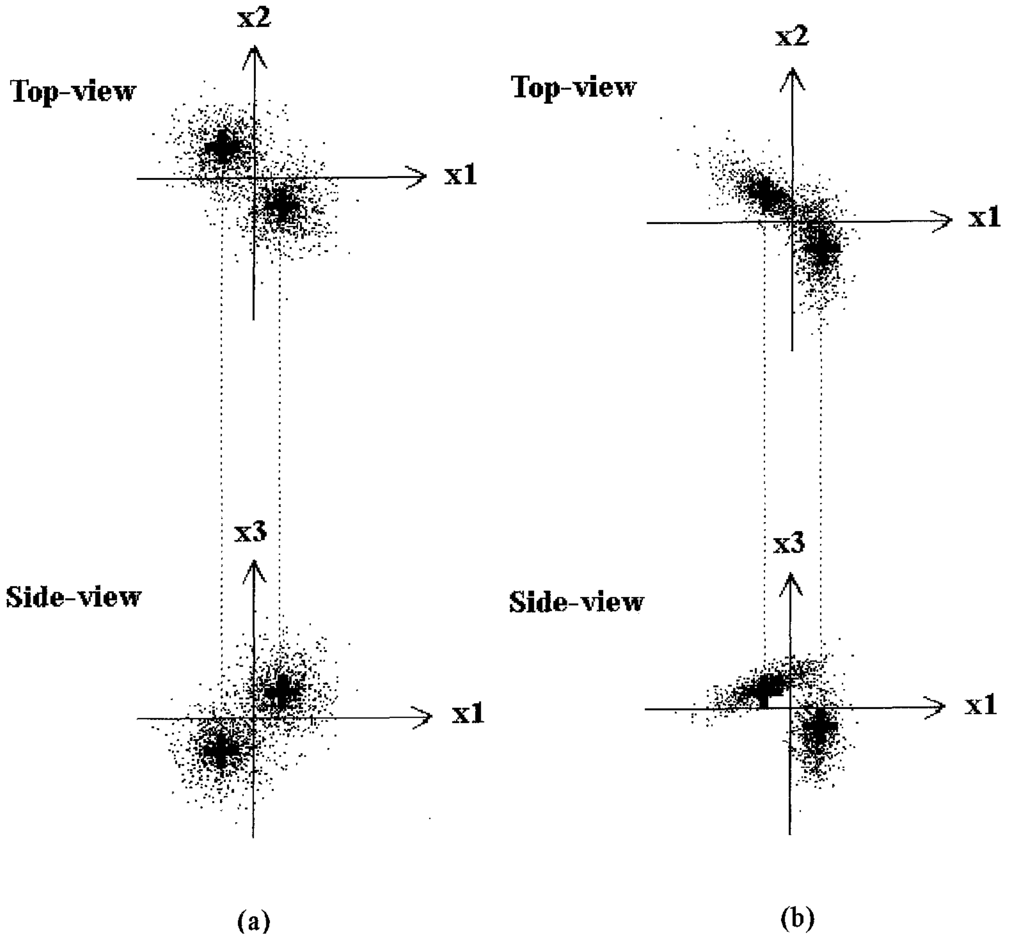


Fig. 1. The computed class representatives of Method 1 for two data sets (a) and (b) in 3-D. Each class representative, when projected on a 2-D coordinate plane, is denoted by the centroid of a cross.

4. EXPERIMENTAL RESULTS

In this section we compare experimental results from the four proposed methods, remarking first that before we apply the methods to the data, we always preprocess it by rotating the data set so that $\overline{x_1^2} = \overline{x_2^2} = \cdots = \overline{x_d^2}$, because we found that the preprocessing gives a general improvement in clustering results no matter which of the four methods is used. The reason this preprocessing can improve the performance is probably that: all clustering formulas (33)–(40) are symmetric among the d variable $x_1, x_2, \ldots, x_d$ and hence no variable should be more dominant than the others; on the other hand, all d variables having identical variations indicates that the ability to distinguish the clusters is similar among the d variables (for example, if $\overline{x_1^2} \approx 0$, then the data are less distinguishable on the x_1 axis); therefore requiring $\overline{x_1^2} = \overline{x_2^2} = \cdots = \overline{x_d^2}$ seems to fit the unbiased nature (unbiased among variables) of our formulas (33)–(40). It can be proved that the preprocessing can be carried out using two-dimensional rotations $d-1$ times and the rotation angles can be calculated explicitly.

Figure 1 shows the experimental results from Method 1 for two data sets in 3-D space. The two classes in data set (a) are spherical, while the two classes in data set (b) are oval-like. In this figure, to sketch 3-D data, we first project each 3-D point (x_1, x_2, x_3) on the x_1–x_2 plane to yield a "top-view", then on the x_1–x_3 plane to yield a "side-view". There are 2000 points in each data set illustrated in Fig. 1 and the running time, including the time to rotate data, was about 0.1 s on the average using a Pentium-90 PC. Each data set's computed class representatives (x_A, y_A, z_A) and (x_B, y_B, z_B) are marked by crosses. [For example, in Fig. 1(a), the centroids of the two crosses in the x_1–x_2 plane are (x_A, y_A) and (x_B, y_B), whereas the centroids of the two crosses in the x_1–x_3 plane are (x_A, z_A) and (x_B, z_B).] We can see that the "crosses" are close to the actual positions of the class centroids.

Figure 2 shows the mis-assignment rates when Methods 1–4 were used for 10 designed 3-D data set. Each data set contained 2000 points and the population ratios of these data sets, from left to right, were designed to be 1:9, 2:8, 3:7, 4:6, 5:5, 1:9, 2:8, 3:7, 4:6 and 5:5. Furthermore, in each data set the two classes

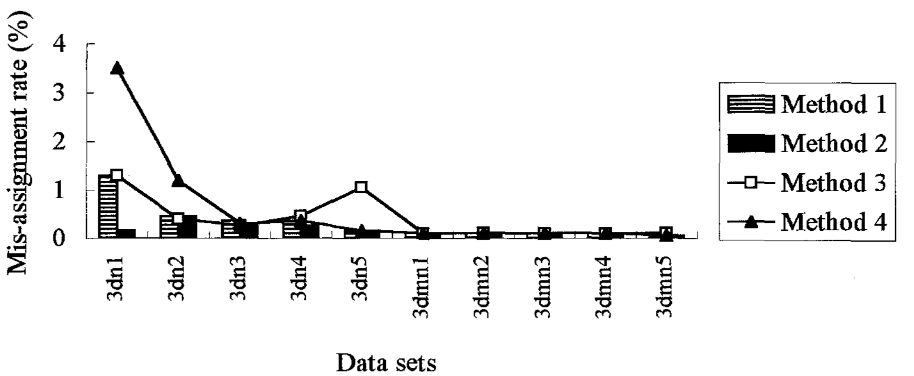


Fig. 2. The mis-assignment rates of Methods 1–4 for 10 tested 3-D data sets.

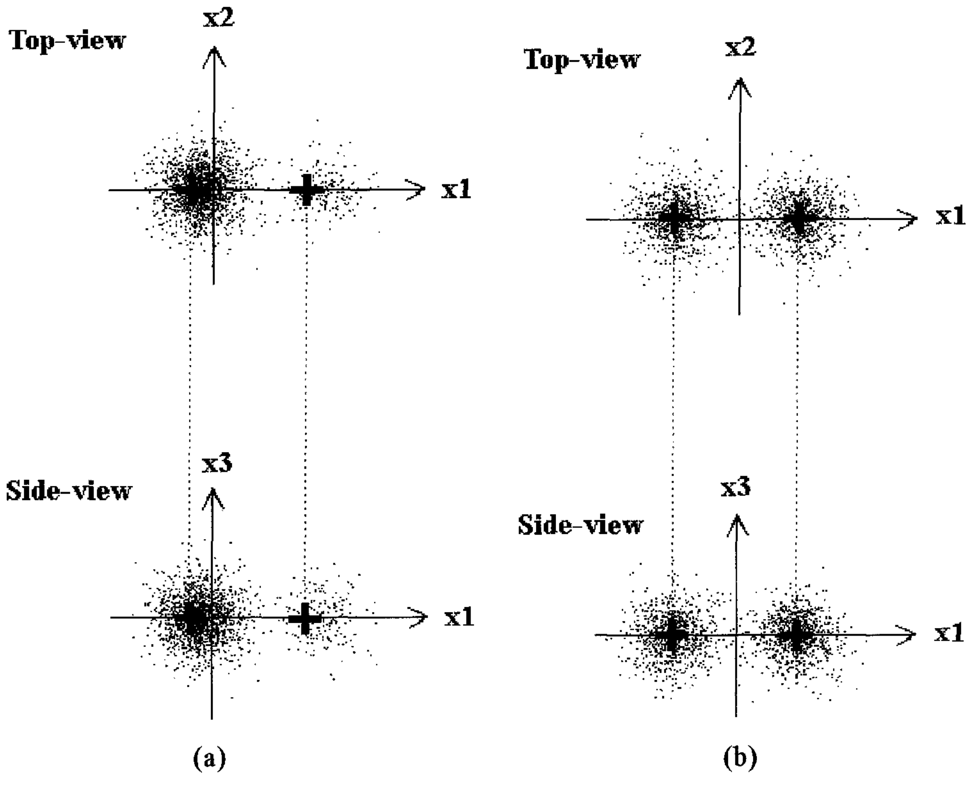


Fig. 3. The computed class representatives of Method 1 for the data sets 3dn1 [shown in (a)] and 3dn5 [shown in (b)] mentioned in Fig. 2.

were designed to be closer in Category 3dn than in Category 3dmn. The distributions of the two designed classes in each data set are both Gaussian distribution. The mis-assignment rates used in this paper were evaluated by counting the percentage of mis-classified data points for a specified data set when a specified method was applied. For example, the 2000-point data set 3dn1 was originally created by merging a 200-point set S_A and a 1800-point set S_B. When a method was applied to cluster the data set 3dn1, if 5 out of the 200 points in S_A were mis-assigned to Class B and if 35 out of the 1800 points in S_B were mis-assigned to Class A, then there were $5 + 35 = 40$ points mis-assigned and the mis-assignment rate is thus $40/2000 = 2\%$ for the data set 3dn1 and the method in question. Figure 3 shows the computed class representatives of Method 1 for the data sets "3dn1" and "3dn5" mentioned in Fig. 2. Although "3dn1" was the most troublesome data set for Method 1 to handle (see Fig. 2), it can be seen from Fig. 3(a) that the two computed class representatives are still not far away the actual ones. Fig. 3(b) illustrates the good estimated class representatives of Method 1 for the data set "3dn5".

Figure 4 illustrates the mis-assignment rates of Methods 1–4 for some 10-D data sets. The data tested in Fig. 4 were designed in a way similar to that of

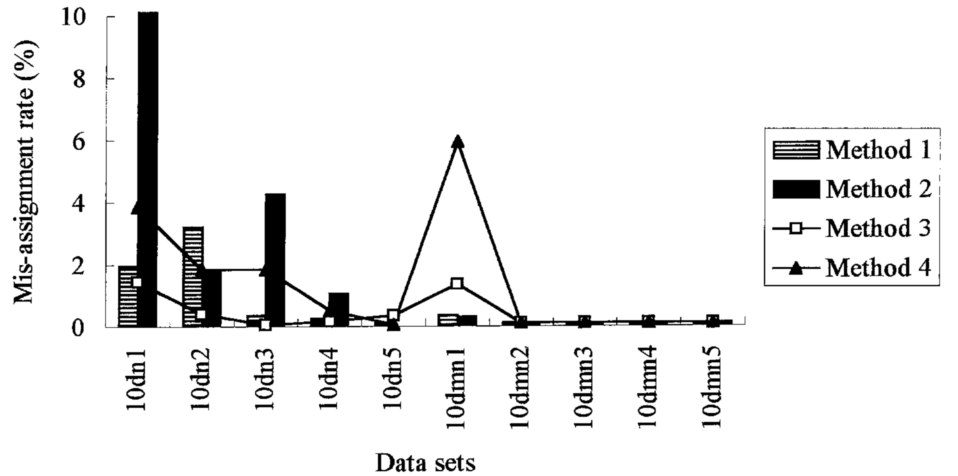


Fig. 4. The mis-assignment rates of Methods 1–4 for 10 10-D data sets. Note that Method 2 fails to yield solutions for the data sets 10dn5.

Fig. 2. In other words, the two classes are closer in Category 10dn than in Category 10dmn and the population ratios between the two classes were designed to be 1:9 for the data set 10dn1, 2:8 for the data set 10dn2, and so on. Note that the experimental results of Method 2 for 10-D data are poor. (Moreover, Method 2 did not yield a solution for the data set "10dn5" because the data made the right-hand side of equation (35) negative, whereas the left-hand side of equation (35) was found to be positive for all p_A and p_B.) Also note that, relatively speaking, Method 4 did not handle the 3-D and 10-D tests well enough. As for Method 3,

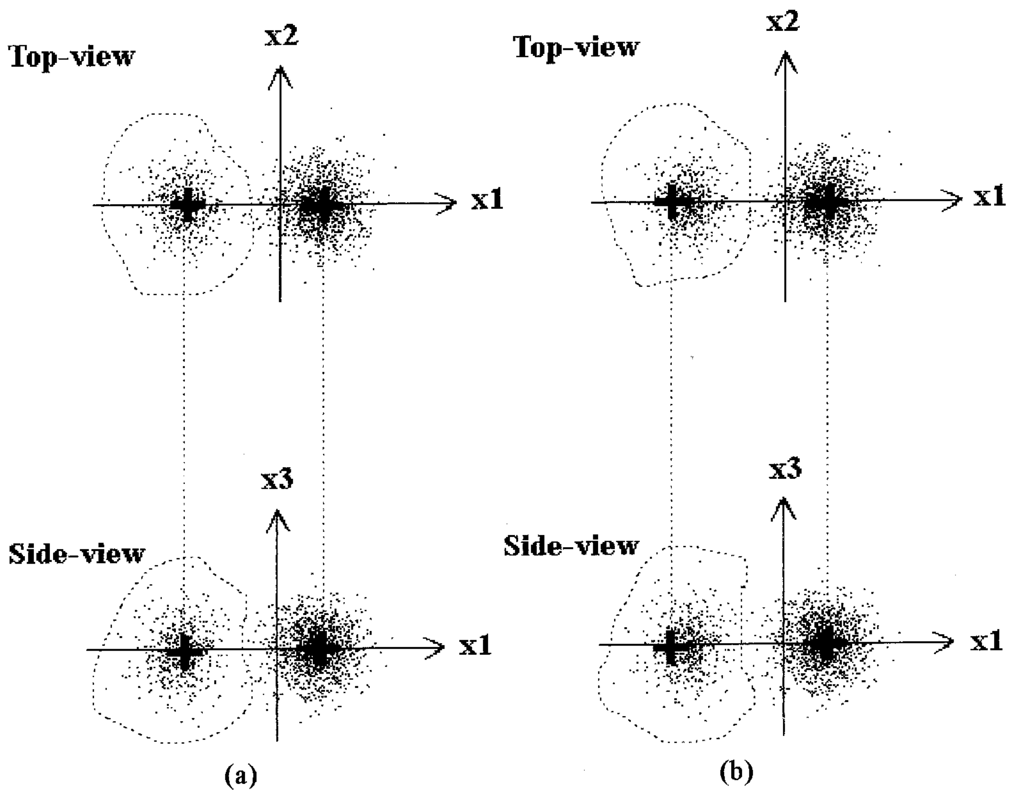


Fig. 5. A comparison of (a) Method 1 and (b) Method 3. The 3dn3 data set was used in both (a) and (b). The computed class representatives are denoted by crosses. The data points which are assigned to class A (class B) are those data points enclosed (not enclosed) by the dotted closed curves.

although its mis-assignment rate was not high, the detected class representatives are in general worse than those detected by Method 1 (one such example is illustrated in Fig. 5). We therefore think that, on average, Method 1 is better than the other three. To convince ourselves, we tested many other data sets from 3-D to 11-D. On average, Method 1 outperformed the other methods. Therefore, we recommend Method 1 as a "first-choice" method. The other three methods, especially, Method 3, should be used as "second-choice" auxiliary methods in case Method 1 fails to generate a solution [this very rare case occurs if and only if there is no p_A and p_B that satisfy equation (33) for a specified right-hand side]. In our experiments, with the data-rotation preprocessing technique mentioned above, only two of the 180 tests caused Method 1 to fail. (In fact, these two data sets still caused Method 1 to fail if we did not apply the data-rotation technique.) Of course, if a reader wishes to stop worrying about whether a solution exists or not, he may always use Method 3 directly without trying Method 1. However, after all points are assigned, he should re-calculate the two class centroids by averaging the coordinates of all points assigned to class A (class B). For example, averaging all points enclosed (not enclosed) by the dotted closed curves in Fig. 5(b). This is because, as stated above, Method 3 gives good point-assignment but poor class representatives.

5. COMPARISON WITH k-MEANS AND HIERARCHICAL METHODS

In this section we compare Method 1 with two other types of widely used clustering methods, the k-means method and the hierarchical agglomerative clustering methods. Figure 6 shows the computed class representatives of the k-means ($k = 2$) method for the two data sets used in Fig. 1. It can be seen that the clustering results are very similar to that of Fig. 1. However, for the data shown in Fig. 7(a), the results are completely distinct. In Fig. 7(b) we list the mis-assignment rates

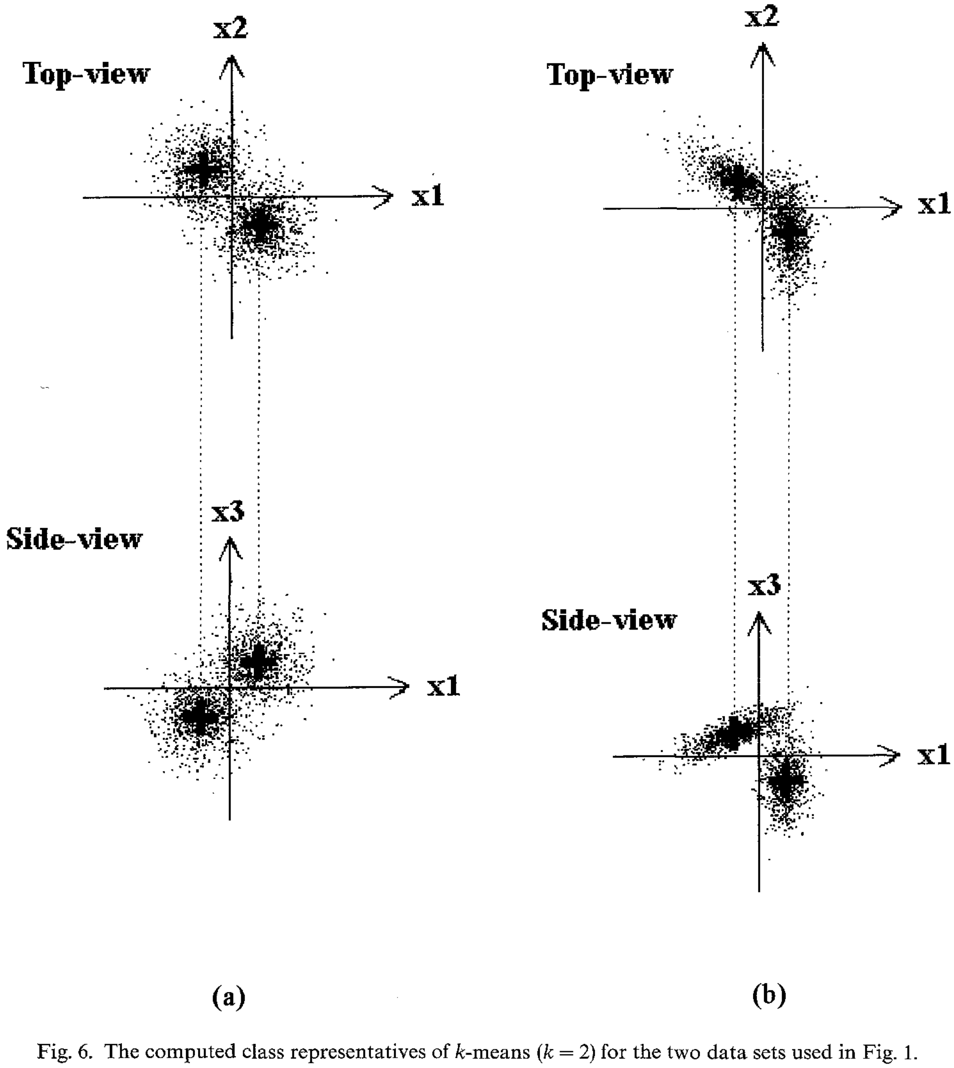
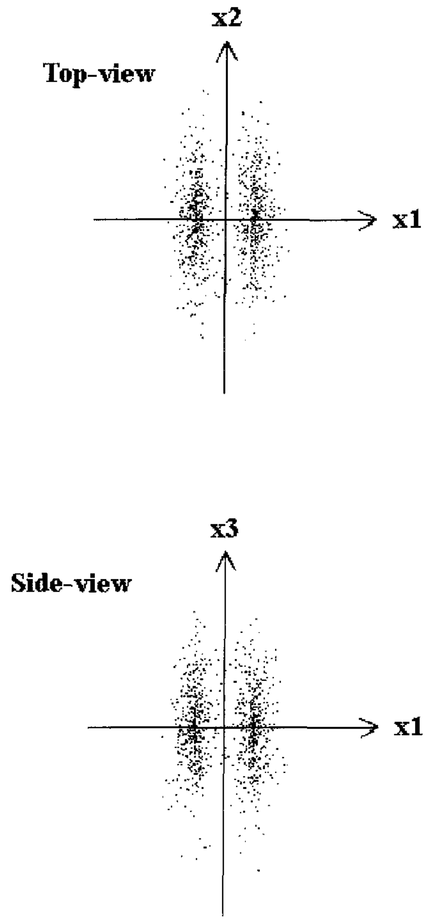


Fig. 6. The computed class representatives of k-means ($k = 2$) for the two data sets used in Fig. 1.



(a)

	Mis-assignment rates	Computation time(in sec.)
our Method 1	3.1%	0.11
k-means method(k=2)	45.2%	0.27
Ward's hierar.	17%	32.95
single linkage hierar.	49.9%	968
complete linkage hierar.	43.1%	160

(b)

Fig. 7. The mis-assignment rates and computation time for the k-means and several hierarchical methods. The 3-D data used is shown in (a).

and computation time of the proposed Method 1, the k-means and four distinct versions of hierarchical agglomerative clustering methods. The number of points in the data set is 1000 and the machine used was a Pentium-90 PC. The reason the k-means showed a very poor mis-assignment rate is that the k-means tried to minimize the total square sum and hence cut the set into upper and bottom halves (along with the x_3 axis) rather than the proper cut of splitting the set into left and right halves (along with the x_1 axis). As for the extremely poor mis-assignment rate of the single linkage hierarchical method, it is because one of the two detected clusters contains one point only. It can also be seen that the computation time of the k-means method was about twice as long as ours and the hierarchical agglomerative clustering methods were several hundred- to several thousand-times slower than ours. In general, the k-means and hierarchical methods do not guarantee better clustering results than ours (sometimes theirs are even worse) and the k-means always requires some initial guesses, while the hierarchical method always has extremely slow computation speed. Our (fast) methods therefore deserves a try. A special note is that the memory requirement of the hierarchical agglomerative methods is proportional to the square of the number of data points; this often makes it impossible to hierarchically cluster large size data on a PC. (Just 7000 points will cause PC problems if hierarchical methods are used.)

An interesting comparison can also be made between the natures of the proposed approach and the k-means method (using $k = 2$). The clustering result and the computation time of the k-means method strongly depend on the initial guess of the partitions (or cluster centroids). This does not happen to our approach because the inputs of our formulae are completely determined by the "whole" given data set rather than of a "partial" data set (which will need to guess which data points should be used). As for the computation time, the k-means and ours were found to be competitive (we used several initial guesses to run the k-means method several times to obtain an average computation time, although the computation time used by our approach was a fixed number once the number of points and the dimensions were fixed).

6. CONCLUDING REMARKS

In this paper we investigated several feature-preserving two-class clustering methods for 3-D or higher dimensional data. The proposed methods operate analytically even when applied to cluster data in a great many dimensions. Among the four investigated methods. Method 1 is the one we recommend most highly because of its low mis-assignment rate and acceptable estimation of class representatives. On the other hand, Method 3 can also be used if the reader remembers that, once the data points are all assigned, the two-class representatives should be replaced by the two-class centroids (see the end of Section 4). As for the other two, Methods 2 and 4, although they are also unsupervised, and have general forms in any space with more than two dimensions, their mis-assignment rates, on average, were found to be not as good as that of Methods 1 and 3.

Also note that Methods 1–4 are all automatic and thier processing speeds are satisfactory. According to our experiments, to cluster 2000 points using a Pentium-90 PC, none of the four methods took more than 0.5 s to do the computation (including the data-rotation preprocessing) if the data are 10-D. This computation speed is several hundred- to several thousand-times faster than that of the hierarchical agglomerative clustering methods. Furthermore, comparing to the k-means,[10] none of the proposed methods needs initial guesses.

In summary, our methods are fast, storage-saving, free from choosing initial guesses, easy to cluster high-dimensional data, etc. Their direct applications include the code-book design for VQ, color image sharpening,[11] color image compression,[12] color palette, etc. Also note that when the methods are repeatedly used to partition a data set into several subsets, and the distances between the subsets are then checked so that adjacent subsets can be merged back to form a cluster, then the proposed methods can also be used in a split-and-merge manner to cluster data formed of more than two classes. This approach of using a two-class clustering method in a split-and-merge manner to solve multiple-classes problem can be found in reference (13).

REFERENCES

1. A. Khotanzad and A. Bouarfa, Image segmentation by a parallel, non-parametric histogram based clustering algorithm, *Pattern Recognition* **23**(9), 961–973 (1990).
2. Y. W. Lin and S. V. Lee, On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques, *Pattern Recognition* **23**(9), 935–952 (1990).
3. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, New Jersey (1988).
4. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Addison-Wesley Publishing Co., New York (1992).
5. J. M. Jolion, P. Meer and S. Bataouche, Robust clustering with applications in computer vision, *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(8), 791–802 (1991).
6. D. Lecompte, L. Kaufman and P. J. Rousseeuw, Hierarchical cluster analysis of emotional concerns and personality characteristics in a freshman population, *Acta Psych. Belg.* **86**, 324–333 (1986).
7. L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*. John Wiley & Sons, New York (1990).
8. H. B. Mitchell, N. Zilverberg and M. Avraham, A comparison of different block truncation coding algorithms for image compression, *Sign. Process. Image Commun.* **6**, 77–82 (1994).
9. A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, New Jersey (1989).
10. J. A. Hartigan and M. A. Wong, A k-means clustering algorithm, *Appl. Statist.* **28**, 100–108 (1979).
11. C. K. Yang, T. C. Wu, J. C. Lin and W. H. Tsai, Color image sharpening by moment-preserving technique, *Sign. Process.* **45**, 397–403 (1995).
12. C. K. Yang, J. C. Lin and W. H. Tsai, Color image compression by moment preserving and block-truncation technique, *Proc. 1994 IEEE Intl. Conf. Image Process.* Austin, Texas, U.S.A. (1994).
13. D. Chaudhuri, B. B. Chaudhuri and C. A. Murthy, A new split-and-merge clustering technique, *Pattern Recognition Lett.* **13**, 399–409 (1992).

APPENDIX

Preserving $\{p, x, y, z, |xy|, |xz|, |yz|, |xyz|\}$ yields:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\sqrt{\delta^2 + 8\delta} - \delta - 2}{2}}. \quad \left(\text{Here, } \delta = \frac{\overline{|xyz|}^2}{\overline{|xy|}\,\overline{|xz|}\,\overline{|yz|}}.\right)$$

Proof. Taking absolute values in equations (10)–(12), we can rewrite equation (9) as:

$$\left(p_A + \frac{p_A^3}{p_B^2}\right)|x_A y_A z_A| = \overline{|xyz|}. \tag{41}$$

By squaring the two sides of equation (41), then using equations (16)–(18), we have:

$$\left(p_A + \frac{p_A^3}{p_B^2}\right)^2 \left(\frac{p_B^3}{p_A^3}\right) \overline{|xy|}\,\overline{|xz|}\,\overline{|yz|} = \overline{|xyz|}^2. \quad (42)$$

Since:

$$\left(p_A + \frac{p_A^3}{p_B^2}\right)^2 \left(\frac{p_B^3}{p_A^3}\right) = \left(\frac{p_A(p_B^2 + p_A^2)}{p_B^2}\right)^2 \left(\frac{p_B^3}{p_A^3}\right)$$

$$= (p_A^2 + p_B^2)^2 \frac{p_A^2 p_B^3}{p_B^4 p_A^3} = \frac{(p_A^2 + p_B^2)^2}{p_A p_B}$$

$$= \frac{((p_A + p_B)^2 - 2p_A p_B)^2}{p_A p_B} = \frac{(1 - 2p_A p_B)^2}{p_A p_B},$$

equation (42) can be simplified as:

$$\frac{(1 - 2p_A p_B)^2}{p_A p_B} = \frac{\overline{|xyz|}^2}{\overline{|xy|}\,\overline{|xz|}\,\overline{|yz|}} = \delta \quad (43)$$

by the definition of δ in equation (21). Since $1 - 4p_A p_B + 4p_A^2 p_B^2 = \delta p_A p_B$, we have $4(p_A p_B)^2 - (\delta + 4)(p_A p_B) + 1 = 0$. Therefore:

$$p_A p_B = \frac{(\delta + 4) \pm \sqrt{\delta^2 + 8\delta}}{8}. \quad (44)$$

Let:

$$\gamma = \frac{(\delta + 4) \pm \sqrt{\delta^2 + 8\delta}}{8}, \quad (45)$$

then equation (44) can be rewritten as:

$$p_A p_B = \gamma. \quad (46)$$

Substituting equations (2) into (46), we obtain:

$$p_A(1 - p_A) = \gamma, \quad \text{i.e.}$$

$$p_A^2 - p_A + \gamma = 0.$$

Consequently:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 4\gamma}. \quad (47)$$

Although γ has two possible values in equation (45), $\gamma = ((\delta + 4) + \sqrt{\delta^2 + 8\delta})/8$ should never be used. (Otherwise, p_A becomes non-real because equation (47) implies:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{(\delta + 4) + \sqrt{\delta^2 + 8\delta}}{2}}$$

$$= \frac{1}{2} \pm \frac{1}{2}\sqrt{-\frac{\delta + 2 + \sqrt{\delta^2 + 8\delta}}{2}}.$$

$$= \frac{1}{2} \pm \frac{1}{2}\sqrt{\text{a negative number}},$$

for the δ defined in equation (21) is always non-negative.) We therefore always use $\gamma = ((\delta + 4) - \sqrt{\delta^2 + 8\delta})/8$. Substituting this γ into equation (47), we can obtain:

$$p_A = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{\delta + 4 - \sqrt{\delta^2 + 8\delta}}{2}}$$

$$= \frac{1}{2} \pm \frac{1}{2}\sqrt{\frac{\sqrt{\delta^2 + 8\delta} - \delta - 2}{2}}. \quad (49)$$

Note that equation (49) requires the δ defined in equation (21) satisfy $\sqrt{\delta^2 + 8\delta} - \delta - 2 \geq 0$, which occurs if and only if $\delta \geq 1$. In other words, $\overline{|xyz|}^2 \geq \overline{|xy|}\,\overline{|xz|}\,\overline{|yz|}$ is required.

About the Author—JA-CHEN LIN was born in 1955 in Taiwan, Republic of China. He received his B.S. degree in computer science in 1977 and M.S. degree in Applied Mathematics in 1979, both from National Chiao Tung University, Taiwan. In 1988 he received his Ph.D. degree in Mathematics from Purdue University, U.S.A. In 1981–1982 he was an instructor at National Chiao Tung University. From 1984 to 1988, he was a graduate instructor at Purdue University. He joined the Department of Computer and Information Science at National Chiao Tung University in August 1988 and is currently an Associate Professor there. His recent research interests include pattern recognition, image processing and parallel computing. Dr Lin is a member of the Phi-Tau-Phi Scholastic Honor Society.

About the Author—WU-JA LIN was born in 1969 in Taiwan, Republic of China. He received his B.S. degree from Tam Kang University in 1991 and M.S. degree from National Chiao Tung University in 1993. He is now working for his Ph.D. degree in the Computer and Information Science Department of Chiao Tung University. His recent research interests include pattern recognition, image processing and neural networks.