ELSEVIER

Analytica Chimica Acta 354 (1997) 263–274

ANALYTICA CHIMICA ACTA

Clustering data using a modified integer genetic algorithm (IGA)

Jian-Hui Jiang, Ji-Hong Wang, Xia Chu, Ru-Qin Yu*

Department of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China

Received 10 December 1996; received in revised form 8 July 1997; accepted 14 July 1997

Abstract

This paper developed a modified genetic algorithm with integer representation (IGA) for cluster analysis problem. The IGA method expands the basic concepts of conventional GAs to include fitness scaling, a modified selection operator, and three newly proposed genetic operators, i.e., competition, self-reproduction and diversification. Moreover, a new clustering criterion was introduced and compared with the commonly used square-error criterion. Clustering of simulated and real chemical data showed that IGA consistently outperformed conventional GAs both in search efficiency and in search precision, and the introduced criterion provided better performance than the square-error criterion. © 1997 Elsevier Science B.V.

Keywords: Cluster analysis; Integer genetic algorithm (IGA); Clustering criteria; Genetic algorithm (GA)

1. Introduction

Cluster analysis is a major problem in the areas of pattern recognition, unsupervised learning and data compression [1]. The objective of cluster analysis is to discover a sensible organization of data. Numerous methods have been reported for data clustering in the literature of diverse fields [2] including chemistry [3–5]. A variety of clustering methods are formulated as optimization problems by defining some criterion functions to be optimized. Apparently, the theoretical solution to a clustering problem is straightforward. One defines a criterion function, evaluates it over all possible partitions containing K clusters and picks the partition which gives the optimal criterion value. One difficulty arises from the fact that most of the clustering criterion functions are nonconvex and have quite a few local optima. Moreover, as there are approximately $K^N/K!$ possible ways to allocate N patterns among K clusters, the problems of clustering data have exponential complexity. These problems are known to be NP-complete (nondeterministic polynomial time complete) [6], and it is clear that exhaustive search of all possible partitions for the exact global optimum is impractical even for a relatively small data set. On the other hand, optimization of these clustering criteria using classic hill-climbing or downhill based techniques, though probably computationally efficient, frequently fall short after getting trapped into local optima, and the local optimal solutions thus obtained are usually insufficient for practical problem solving. To combat this difficulty, a viable way is to utilize an appropriate global optimization technique, such as simulated annealing (SA) [7] or a genetic algorithm (GA) [8]. These global optimization techniques provide practical approaches for finding the globally optimal or quasi-optimal solution with computational

*Corresponding author. Fax: +86 (731) 8824287.

PII S0003-2670(97)00462-5

requirement proportional to a small power of the sample size N.

Genetic algorithms (GAs) are a family of global optimization techniques, which have proven to be efficacious and efficient in tackling various complex and large-scale optimization problems, even those belonging to the class of NP-complete ones. The attractive power of GAs derives from the behavior of implicit exponential sampling of the search space during the genetic search process. Applications of GAs to the solution of chemical problems, recently have comprised growing interest from chemometricians [8–12]. A pilot study of using a GA for a data clustering problem has been presented by Lucasius et al.[13]. However, there are still many unsolved problems in GAs. The primary criticism toward GAs is the occurrence of premature convergence, which always induces GAs to produce an undesirable solution rather than the global optimum. Another problem with GAs is that, when the algorithm is terminated in limited time, one used to find the methodology exhibiting rather poor search precision. This unfavorable characteristic may to some extent indicate that the global optimality of the solution given by GAs is dubious.

Selection of a criterion for a practical clustering problem produces another difficulty to cluster analysis. A clustering criterion always reflects the investigator's intuitive notion about what a cluster looks like. In this sense, each clustering criterion imposes a certain structure on the data. The true clusters are difficult to be discovered, unless the data conform to the requirements of the selected criterion. This difficulty is especially serious for the criteria derived from some parametric models. For instance, the commonly used square-error criterion is biased towards equally-sized globular clusters, while another popular clustering criterion, the Friedman and Rubin's criterion [14], tends to impose a similar hyperellipsoidal configuration on each cluster. Both the square-error criterion and the Friedman and Rubin's criterion can be derived from some restricted Gaussian mixture models. An advantage of the parametric methods over the non-parametric ones is their high efficiency in the situations when the underlying data structures happen to be described by the parametric models. To keep up high efficiency while mitigate the tendency of imposing some structures, an appropriate way is to use a general parametric model, which can described most of data structures encountered in practice.

In this paper, attempts were made to attack both the difficulties inherent in the clustering problems, i.e., the local optima problem and the criterion selection problem, at the same time, to alleviate the problems with GAs, i.e., premature convergence and poor search precision. Our work was distinguished from the previous studies in two major aspects. First, a new clustering criterion has been introduced. This criterion, derived from a general Gaussian mixture model, has a very weak tendency of imposing a particular structure on the data, therefore, it is suitable for a vast of clustering problems encountered in practice. Second, a modified GA with integer representation (IGA) has been developed, which expands the basic concepts of conventional GAs to include fitness scaling and three new genetic operators, i.e., competition, self-reproduction and diversification. These new genetic operators, together with other conventional operators of selection, crossover and mutation in the modified form, integrate exploration and exploitation, global and local search in a balanced manner into the proposed IGA. Experimental results show that the developed IGA consistently outperforms conventional GAs both in search efficiency and in search precision, and the introduced clustering criterion provides a better performance than the commonly used square-error criterion.

2. Theory

2.1. Criterion functions for cluster analysis

The problem of hard clustering can be mathematically stated as follows: given a set of N patterns, x_1, $x_2,\ldots, x_N$ in d-metric space, the aim is to group the patterns into several disjoint subsets, i.e., clusters, C_1, $C_2,\ldots, C_K$ without the aid of category labels such that the patterns within the same clusters are somehow more similar to each other than patterns in different clusters. The term 'hard' means that each pattern belongs to only one cluster. Suppose that cluster C_k has N_k patterns, then one has

$$\sum_{k=1}^{K} N_k = N \tag{1}$$

One way to make this into a soundly formulated problem is to define a criterion function measuring the clustering quality of any partition of the data. Therefore, the clustering problem is to seek the partition which optimizes the criterion function, known as partitional clustering. The most widely used criterion function is the square-error criterion. Let $\boldsymbol{m}_k$ be the mean of cluster C_k , i.e.,

$$\boldsymbol{m}_k = \sum_{i=1}^{N_k} \frac{x_i^{(k)}}{N_k} \tag{2}$$

where $x_i^{(k)}$ is the ith pattern belonging to cluster C_k, then the square-error criterion is defined by

$$J_{se} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \|x_i^{(k)} - \boldsymbol{m}_k\|^2 \tag{3}$$

The objective of the square-error clustering is to find a partition that minimizes J_{se} for a fixed K. This criterion has been criticized for two reasons. One is that it tends to seek for equally-sized hyperspherical clusters, and the other is that this criterion is not invariant under nonsingular linear transformations. The Friedman and Rubin's criterion seeks a partition which minimizes the determinant of $\boldsymbol{S}_w$, the within-cluster scatter matrix, which is defined by

$$\boldsymbol{S}_w = \sum_{k=1}^{K} \sum_{i=1}^{N_k} (x_i^{(k)} - \boldsymbol{m}_k)(x_i^{(k)} - \boldsymbol{m}_k)^T \tag{4}$$

where the subscript T denotes the matrix transposition. This criterion has an advantage over the square-error criterion in that it is invariant to nonsingular linear transformations of patterns. One drawback with this criterion is the imposition of a similar hyperellipsoidal structure on each cluster. Both the square-error criterion and the Friedman and Rubin's criterion can be derived from the Gaussian mixture models

$$f(x) = \sum_{k=1}^{K} p_k \phi(x|\boldsymbol{\mu}_k, \Sigma_k) \tag{5}$$

where p_k , $\boldsymbol{\mu}_k$ and $\sum_k$ are the a priori probability, the mean vector and the covariance matrix, respectively, of the k th component normal distribution ϕ. In most clustering problems, p_k (k =1,..., K) are implicitly assumed to be equal. It has been pointed out that the square-error criterion can be derived from the restricted Gaussian mixture model with $\Sigma_k = \lambda \boldsymbol{I} (k = 1, \ldots, K)$, where $\boldsymbol{I}$ is the identity matrix and λ is a positive number, while the Friedman and Rubin's criterion can be derived from the restricted Gaussian mixture model with identical covariance matrix for each cluster [2]. Without these particular restrictions to the covariance matrices, the following general Gaussian mixture model

$$f(x) = \sum_{k=1}^{K} \phi(x|\boldsymbol{\mu}_k, \Sigma_k) \tag{6}$$

provides a more general model for practical clustering problems, which allows each cluster to have an arbitrary hyperellipsoidal structure. In fact, most data involved in chemical practice can be described by this general model. Therefore, a criterion, which can be derived from this general Gaussian mixture model (Eq. (6)), is introduced in the presented work, as defined by

$$J_{Gm} = \sum_{k=1}^{K} N_k \ln(|\boldsymbol{S}_k / N_k|) \tag{7}$$

where $\mathbf{S}_k$ is the scatter matrix of the kth cluster and $|\cdot|$ denotes the deteminant of a square matrix. Minimization of this criterion provides a flexible way for cluster analysis. This criterion has a very weak tendency to seek clusters of a particular structure and it is invariant to nonsingular linear transformations of patterns. Generally, this criterion can be optimized using the so-called CEM algorithm, a classification version of the EM algorithm. The detailed steps of the CEM algorithm for this criterion (Eq. (7)) are presented as follows:

(i) Select an initial partition containing K clusters of the patterns.
(ii) Compute $\boldsymbol{m}_k = \sum_{i=1}^{N_k} x_i^{(k)} / N_k$ and $\boldsymbol{S}_k = \sum_{i=1}^{N_k} (x_i^{(k)} - \boldsymbol{m}_k)(x_i^{(k)} - \boldsymbol{m}_k)^T$.
(iii) Compute $d_k(x_i) = (x_i - \boldsymbol{m}_k)\boldsymbol{S}_k^{-1}(x_i - \boldsymbol{m}_k)^T / N_k + \ln(|\boldsymbol{S}_k / N_k|)(1 \le i \le N, 1 \le k \le K)$.
(iv) Update partition by assigning each x_i to the cluster which gives the minimum $d_k(x_i)$.
(v) If current partition is distinguished from the previous one, return to (ii); otherwise, terminate the algorithm.

In passing, we have derived from the most general Gaussian mixture model (Eq. (5)) a clustering criterion, as defined by

$$J_{\mathrm{Gm}}{}' = \sum_{k=1}^{K} N_k \ln(|S_k/N_k|/N_k^2) \tag{8}$$

Since the a priori probabilities p_k ($k=1,\ldots,K$) are not assumed to be equal in deriving this criterion, this criterion may have a certain advantage over the aforementioned criterion (Eq. (7)) under the mixture sampling situations. i.e., the number of samples in different clusters possibly differs significantly. We have experienced many examples that, in usual clustering problems in which the number of samples in different clusters does not have great difference, both criteria (Eq. (7) and Eq. (8)) turn out to show a comparable performance.

2.2. An integer GA for data clustering

GAs are a family of stochastic optimization techniques emulating the natural evolutional process in a numerical manner. The philosophy of GAs is to build the nature's principles of evolution into the algorithmic operators, such that the produced genetic operators are beneficial for search efficiency and implementation convenience. Based on such a philosophy, a modified GA with integer representation is proposed by the present authors with prior attempts to overcome the aforementioned problems in GAs.

2.2.1. Population representation

A salient characteristic of GAs is that they work with a population (usually of fixed size, say N_p) of string-represented candidate solutions. By maintaining a multipoint perspective on the search space with successive populations, GAs have relatively high chance of finding the global optimum. A candidate solution, i.e., an individual, in the proposed GA is an indicator vector whose components, i.e., genes, indicate the cluster membership ($k=1,\ldots,K$) of corresponding patterns. That is, a candidate solution for a clustering problem of N patterns is an N-dimensional vector, and the n th component of it is an integer equaling the ordinal number of the cluster to which the n th pattern belongs. As such an integer representation is used for candidate solutions, the proposed GA is referred to as integer GA (IGA). This integer representation for candidate solutions has two algorithmic advantages over the binary representation commonly used in conventional GAs. One is that, with this integer representation, the clustering criterion values can be calculated immediately without preliminary decoding of candidate solutions. On the other hand, with a binary representation which represents the cluster membership by several bits, unless $K+1$ happens to be a power of two, some illegal strings will unavoidably generated by common bit operations in conventional GAs. This incurs much inconvenience for the implementation of some genetic operators. Moreover, with the proposed integer representation, the developed IGA can be applied to hard partitional clustering problems with various criteria, besides the criterion (Eq. (7)) used in this work, merely through a minor modification.

2.2.2. Fitness scaling

As the clustering problem seeks a partition which minimizes a predefined clustering criterion (Eq. (7)), in principle, any non-increasing function of the criterion value with non-negative response can be used for fitness evaluation. However, because exploitation of useful information accumulated in past generations in GAs is based on the fitness values of individuals, fitness evaluation plays an important role in the genetic search process. Ideally, the fitness of individuals should be evaluated in such a manner that efficient exploitation takes place while sufficient diversity is maintained in the population for productive exploration. A feasible approach to this goal is to incorporate a fitness scaling transformation into the evaluation of fitness. There are two major concerns supporting the scheme of fitness scaling when individuals are reproduced or replaced probabilistically with expected rates proportional to their fitness values. One is the occurrence of a large offset value of fitness, which makes the selection operator simply serve as a random walk for non-optimal individuals. The other is the case, when the fitness values of a population tend to be homogeneous such that fitness scaling is needed to enhance the selection pressure on the population.

In this work, the fitness of an individual is evaluated by applying a sigmoidal scaling to its raw fitness

value, which is immediately taken to be the negative criterion value associated with this individual. That is, suppose the criterion value vector associated with current population is $\boldsymbol{J}_{\mathrm{Gm}} = (J_{\mathrm{Gm}}(1), J_{\mathrm{Gm}}(2), J_{\mathrm{Gm}}(N_{\mathrm{p}}))$ with the ith component being the criterion value of the ith individual, then for an individual with a criterion value J_{Gm}, the raw fitness value of it is $-J_{\mathrm{Gm}}$ and the evaluated fitness of this individual is given by

$$fitness\,(i) = 1/\{1 + \exp[-(aJ_{Gm} + b)]\} \tag{9}$$

where a and b are two constants controlling the slope and bias of sigmoidal function, and are determined in a manner such that the following set of equations hold:

$$aJ_{\mathrm{Gm},[N_{\mathrm{p}}/4]} + b = 3 \tag{10}$$

$$aJ_{\mathrm{Gm},N_{\mathrm{p}}-[N_{\mathrm{p}}/4]} + b = -3$$

where $[x]$ is the largest integer not larger than x, $J_{\mathrm{Gm},[N_{\mathrm{p}}/4]}$ and $J_{\mathrm{Gm},N_{\mathrm{p}}-[N_{\mathrm{p}}/4]}$ are the $[N_{\mathrm{p}}/4]$th and $(N_{\mathrm{p}} - [N_{\mathrm{p}}/4])$th smallest criterion values, respectively, in *JGm* of current population. Obviously, the fitness values such evaluated are always positive. This ensures the legality of fitness-proportional selection operations. Furthermore, with such a choice of a and b, the proposed scheme of fitness evaluation has two attractive characteristics. One is that the best performing $[N_{\mathrm{p}}/4]$ individuals have comparable fitness such that a mild competition among these individuals is guaranteed. This prevents fitness-proportional reproduction and replacement operators from generating a homogeneous population. Meanwhile, the worst performing $[N_{\mathrm{p}}/4]$ individuals have certain fitness values greater than zeros, which ensures that these individuals still have a certain probability to be selected to join future population such as to increase the diversity in population. The other is that, a certain competition pressure is maintained throughout the genetic search process.

2.2.3. Selection and competition

The selection operator has a slightly modified form in the proposed IGA. This operator generates a reproducible parent population of size N_{p}. Similarly to conventional GAs, each individual is selected to join the parent population with a probability proportional to its fitness, relative to the other individuals in current population. That is, the probability of individual j being selected is

$$p(j) = fitness\,(j) / \sum_{i=1}^{N_{\mathrm{p}}} fitness\,(i) \tag{11}$$

To generate a productive parent population, the $(2j-1)$th and the $(2j)$th parents ($j = 1,\ldots, N_{\mathrm{p}}/2$) are prohibited from being selected from the same individual. That is, if the $(2j-1)$th parent is generated according to Eq. (11) from current population, then the $(2j)$th parent should be produced according to Eq. (11) from those individuals in current population other than the $(2j-1)$th parent selected. In this sense, the parent population generated by the modified selection operator is considered to be an ordered population.

Competition operator, as proposed by the present authors, serves as natural environments, in that it statistically selects the surviving individuals according to their fitness from the competing population. The competing population consists all individuals of the current population and the offspring populations generated by crossover and mutation. It can also considered as a probabilistically biased replacement operator. According to the fitness of individuals, the competition operator cyclically selects N_{p} surviving individuals from the competing population. The surviving probability of each individual in each cycle is also proportional to its fitness relative to other individuals in the competing population. However, to maintain sufficient diversity in the surviving population, in the proposed competition operator, each individual has merely one chance of being selected to join the surviving population. That is, when one individual survives in the previous cycle, its fitness is set to 0 in following cycles. To ensure that IGA will converge asymptotically toward the global optimum, the optimal individual thus far obtained survives and is retained in the surviving population with a probability of one.

2.2.4. Crossover and mutation

Crossover and mutation operators are two important sources of exploration. An efficient GA should feature a high exploratory power while maintaining a balance between exploration and exploitation. The proposed IGA deals with this trade-off using crossover and mutation operators with high exploratory power,

followed by a probabilistically biased replacement operator, i.e., competition. This enables IGA to seek a proper balance between exploration and exploitation without loss of exploratory power.

Crossover is a method of sharing useful information in two successful individuals. With a high exploratory power, the crossover operator has relative high potential to reproduce new building blocks, which ensures that the GA methodology can probe the attraction regions of all optima. The highest exploratory power is obtained by using a uniform crossover scheme in the presented work. In this crossover scheme, after a parent pair is drawn, each corresponding pair of genes exchange their values independently with the same probability of 0.5. As the modified selection operator generates an ordered population of individuals, a parent pair is obtained by pairing the $(2j-1)$th parent with the $(2j)$th parent immediately. Provided the population is sufficiently diverse, it can be expected that, together with such a pairing method, crossover scheme can be very productive. Another benefit of the uniform crossover scheme is that it has no position bias. As the patterns always have significantly correlated effect on the criterion value in the clustering problems, and no prior information about such correlation of patterns can be obtained, uniform crossover is a wise way to alleviate undesirable position bias.

The main role of mutation in conventional GAs is to ensure the total reach of genetic operators to cover the whole feasible region of the optimization problem. Mutation in IGA serves another potential source of exploratory power. During the mutation operation, each gene of all individuals has a probability of p_m to be set to a value randomly drawn from 1, 2,..., and K. Here p_m is the mutation probability, usually set to being larger than 0.2, such that mutation operator has high exploratory power.

2.2.5. Self-reproduction

Self-reproduction, as proposed by the present authors, provides a distinct approach to exploit the search space in IGA. The motivation of this new operator is to incorporate a local search heuristic into the search process of GAs. With this operator, IGA generally shows a significantly improved search efficiency compared with conventional GAs, at the same time without loss of the characteristic of convergence to the global optimum. Self-reproduction is accomplished by carrying out the following procedures on each individual with the same probability p_r, the self-reproduction probability.

(i) $age=0$.
(ii) $age=age+1$.
(iii) Call the steps (ii) to (iv) of the CEM algorithm for criterion (7), as presented before.
(iv) If $age \leq lifetime$ and the current individual is distinguished from the previous one, return to (ii); otherwise, calculate the value of criterion (7) of the current individual and terminate the self-reproduction procedure of the current individual.

where *lifetime* is a controlling parameter and *age* is a temporary variable. Note that the proposed reproduction operator is criterion-specific and should be modified according to the particular criterion used for partitional clustering. Nevertheless, generalization to many other criteria, say the square-error criterion, is straightforward. A general self-reproduction operator which is suitable for any clustering criterion can be developed in a manner similar to previous study [15]. In this work the above-described self-reproduction procedure is used since it is much more computationally efficient. Another advantage derived from self-reproduction is that a common drawback with conventional GAs, i.e., poor search efficiency, is remedied in IGA. Indeed, self-reproduction implements a parallel hybrid strategy in IGA.

2.2.6. Diversification

The objective of diversification operator, as proposed by the present authors, is to maintain the diversity in population such as to enhance the exploratory power of crossover. Diversification operator includes two algorithmic steps. First, individuals which differ from one of the other individuals in not more than one gene die away from the population. Subsequently, a certain number of individuals are randomly generated to join the population such that the population size remains N_p. This operator, followed by the modified selection operator described above, is useful for supporting a productive crossover operator. Moreover, the diversification operator ensures the legality of fitness scaling as described before, since $J_{Gm,[Np/4]}$ and $J_{Gm,Np-[Np/4]}$ in Eq. (10) will generally differ from each other.

2.2.7. Algorithm

Having described various genetic operators in detail, the algorithmic steps for IGA can be presented as follows:

(i) Select an appropriate parameter setting for IGA.
(ii) Initialize a population of candidate solutions of size N_p randomly such that each gene of an individual takes a value from 1 to K with equal probability.
(iii) Calculate the values of clustering criterion (Eq. (7)) of each individual in the population.
(iv) If the maximal generation admissible g_m expires or the optimal performance has not been improved for some number of generations g_n, the algorithm terminates; otherwise, it continues.
(v) Evaluate the fitness of individuals in current population using Eq. (9).
(vi) Select a parent population composed of N_p ordered copies of individuals in current population.
(vii) Pair the $(2j-1)$th parent with the $(2j)$th parent (j $=1,\ldots, N_p/2$) in the parent population.
(viii) Crossover the $N_p/2$ parent pairs such as to produce an offspring population (offspring population 1) of size N_p.
(ix) Mutate each gene in offspring population 1 with a probability p_m such as to produce another offspring population (offspring population 2).
(x) Calculate the criterion values of individuals in offspring populations 1 and 2 and evaluate their fitness using Eq. (9).
(xi) All individuals in current population and these two offspring populations are subjected to the competition operator such as to generate a surviving population.
(xii) Let each individual in the surviving population self-reproduce with a probability p_r, and subsequently update the individuals which have self-reproduced by their offsprings produced by self-reproduction.
(xiii) The updated surviving population is subjected to the diversification operator such as to generate a population of size N_p of next generation.
(xiv) Return to (iv).

To verify the performance of the proposed IGA, another GA is used in this work. This GA is essentially conventional except for an integer representation, and the algorithm is presented as follows:

(i) Select a parameter setting and initialize a population randomly as IGA.
(ii) Calculate criterion values for all individuals.
(iii) If the terminating criterion of IGA is satisfied, the algorithm terminates; otherwise, it continues.
(iv) Generate a parent population by rank-based selection.
(v) Generate $N_p/2$ parent pairs as IGA.
(vi) Uniformly crossover each parent pair with a probability p_c.
(vii) Mutate each gene with a small probability p_m.
(viii) If the best performing individual obtained thus far is not retained in the offspring population, the worst individual in the offspring population is replaced by the best performing one.
(ix) Return to (iii).

3. Experimental

Four data sets, two simulated and two real chemical data sets, were used for demonstrating the performance of the clustering criterion introduced and the IGA proposed. For the sake of validating the clustering results visually, both of the simulated data sets are two-dimensional.

3.1. Simulated data 1

This data set is composed of two Gaussian clusters, each consisting of 50 independent and identically distributed patterns. Cluster 1 has an expected mean of $\boldsymbol{\mu}_1$ and an expected covariance matrix Σ_1, and cluster 2 has an expected mean of $\boldsymbol{\mu}_2$ and an expected covariance matrix Σ_2. Here,

$$\boldsymbol{\mu}_1 = (3.5,\ 0)^T \quad \boldsymbol{\mu}_2 = (0,\ 0)^T \quad \Sigma_1 = 0.3^2 \times \mathrm{diag}(1,\ 1) \quad \Sigma_2 = \mathrm{diag}(1,\ 1)$$

where $\mathrm{diag}(x_1, x_2)$ is a diagonal matrix with diagonal elements of x_1 and x_2.

3.2. Simulated data 2

This data set is composed of three Gaussian clusters, each consisting of 50 independent and identically

distributed patterns. The three clusters have the same covariance matrix Σ but have different expected means, i.e., μ_1, μ_2 and μ_3, respectively. Here,

$$\mu_1 = (3,\ 0)^T \quad \mu_2 = (0,\ 0)^T \quad \mu_3 = (1.5,\ 1.5 \times 3^{1/2})^T$$
$$\Sigma = \mathrm{diag}(0.3,\ 1)$$

3.3. Chinese tea data

The Chinese tea data consist of 31 Chinese tea samples belonging to three categories: green tea, black tea and oolong tea [16]. Each sample is represented by a set of measurements on six chemical components: cellulose, hemicellulose, lignin, polyphenols, caffeine and amino acids. The dimensionality of these data was reduced to two by applying principal component analysis (PCA) to the autoscaled data in this work. That is, each pattern was represented by the first two principal component scores of corresponding six-dimensional data point. The aim was to investigate whether the reduced data formed obvious clustering structure according to their categories.

3.4. Iris data

The *Iris* data consist of 150 samples, each represented by a four-dimensional pattern [17]. These samples have their origin from three categories of Iris flower: *Iris Setosa*, *Iris Versicolor* and *Iris Virginica*. The aim was to investigate whether the patterns formed distinguished clusters according to their categories.

4. Results and discussion

To start IGA, an appropriate parameter setting should be selected to keep a balance between exploitation and exploration. These parameters include the population size N_p, the mutation probability p_m, the self-reproduction probability p_r and the controlling parameter, *lifetime*, in self-reproduction operator. Since IGA deals with the trade-off between exploitation and exploration using separated genetic operators, these parameters have no significant effect on the convergence rate of IGA. A consistent parameter setting was used in this work for all the data sets studied, where N_p was set to 20, p_m was set to 0.3, p_r was set to 0.1 and *lifetime* was set to 10. Similarly, a consistent parameter setting was used for the conventional GA, including the population size set to 20, the crossover probability set to 0.9 and the mutation probability set to 0.01. Such a parameter setting has been extensively used in GA literature. Both in IGA and in the conventional GA, the controlling parameter, g_m and g_n, for terminating criterion were set to 600 and 100, respectively. To make a rational comparison between the proposed IGA and the aforementioned conventional GA, each algorithm was repeated for 10 times and the convergence rates were averaged. In all clustering problems in this work, the number of clusters, K, was assumed to be known a priori. When such a priori information is not available, one should run the clustering algorithm repeatedly for different values of K, and the best K can be found using a cluster validation method. However, the presented work would skip the studies of cluster validation for the sake of simplicity. To compare the behavior of the introduced clustering criterion (Eq. (7)) and the square-error criterion (Eq. (3)), a variant of IGA, was developed for minimizing the square-error criterion. This could be immediately accomplished by modifying the self-reproduction operator based on the K-means algorithm [2].

The first simulated data set consists of two spherical clusters of different size. The clustering structure can be visualized straightforwardly from Fig. 1. One can discover that the patterns forms two obviously separated clusters. The clustering results of this data set obtained using the IGA for the square-error criterion (Eq. (3)) is shown in Fig. 1(a), where the cluster boundary is a straight line. It can be seen that four patterns belonging to the second cluster are misclassified into the first cluster. This shows the aforementioned defect of the square-error criterion that it imposes an equally-sized spherical clustering structure on the data. In contrast, the results given by the IGA for the introduced criterion (Eq. (7)) are much favorable. As shown in Fig. 1(b), the clustering results are perfectly consistent with the actual clustering structure, and the cluster boundary obtained is approximately a circle. On the other hand, it was discovered that IGA showed much higher efficiency than the conventional GA. IGA always located the same optimal solution in different runs with a com-

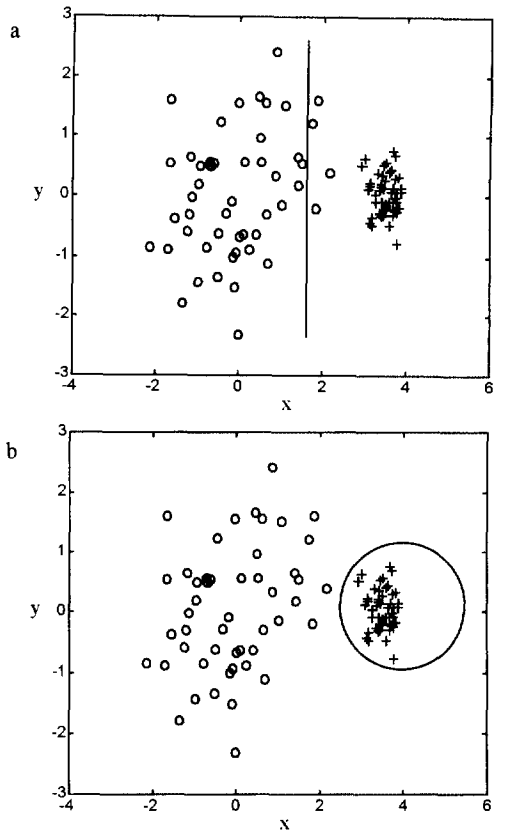

Fig. 1. Clustering results of simulated data set 1 (Cluster 1, +; Cluster 2, ○). The solid line or curve is the approximate cluster boundary given by the algorithm used. a. Clustering results obtained using the IGA for the square-error criterion. b. Clustering results obtained using the IGA for the introduced criterion.

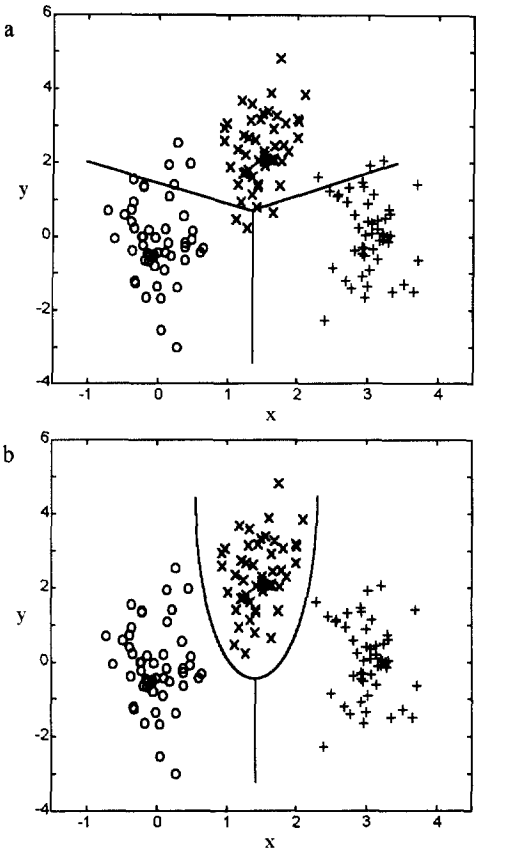

Fig. 2. Clustering results of simulated data set 2 (Cluster 1, +; Cluster 2, ○; Cluster 3, ×). The solid line or curve is the approximate cluster boundary given by the algorithm used. a. Clustering results obtained using the IGA for the square-error criterion. b. Clustering results obtained using the IGA for the introduced criterion.

putational requirement of 9 generations on the average. The conventional GA, however, took 461 generations to converge to the optimal given by IGA or its neighborhood.

The second simulated data set is composed of three identically-shaped ellipsoidal clusters with centroids located at the vertices of an equilateral triangle with intervertex distance of 3. As can be seen from Fig. 2 that all patterns form three obvious separated clusters. The clustering results given by the IGA for the square-error criterion are shown in Fig. 2(a). Ten patterns are grouped incorrectly. This is a further evidence of the fact that the square-error criterion tends to look for clusters of equally-sized spherical clusters. The criterion (Eq. (7)) introduced by the present authors is not subject to this defect. As is shown in Fig. 2(b), the clustering results have the actual clustering structure and all patterns are grouped correctly. In this experiment, it took merely 16 generations for IGA to reach the same optimum in different runs, but the conventional GA required 542 generations on the average to converge to this optimal solution or its neighborhood.

The reduced Chinese tea data are shown in Fig. 3. One can perceive that the patterns form three clearly separated clusters, each associated with one category of tea samples. The clustering results given by the IGA for the square-error criterion are shown in Fig. 3(a), where the sample C7 is grouped into the cluster associated with oolong tea. However, it is discovered that the nearest three neighbors of the pattern corre-

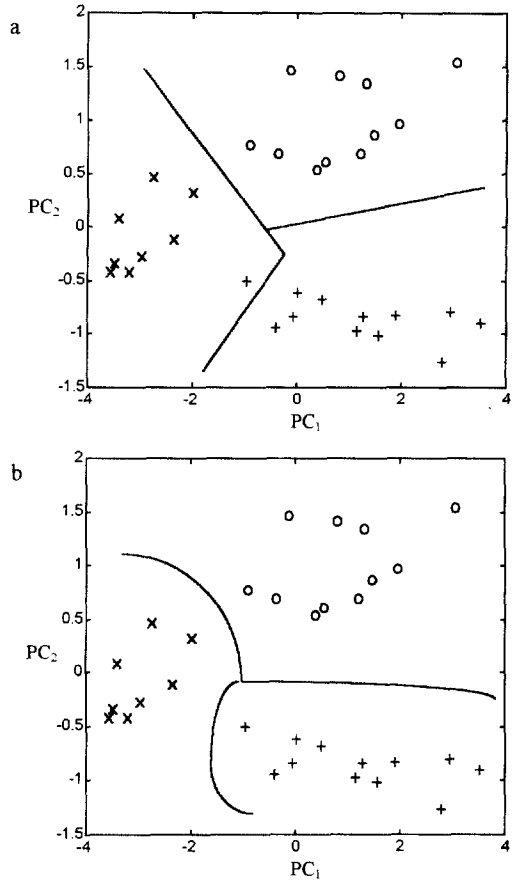

Fig. 3. Clustering results of the Chinese tea data (Green tea, +; Black tea, ○; Oolong tea, ×). The solid line or curve is the approximate cluster boundary given by the algorithm used. a. Clustering results obtained using the IGA for the square-error criterion. b. Clustering results obtained using the IGA for the introduced criterion.

sponding to sample C7 in the reduced space belong to the cluster corresponding to green tea. Therefore, it might be more reasonable to allocate the pattern of sample C7 into the cluster associated with green tea. The clustering results obtained using the IGA for the introduced criterion are shown in Fig. 3(b), where all patterns are clustered in a manner perfectly consistent with one's perception. The reduced data set also indicates that the restrict Gaussian mixture model deriving the square-error criterion is not a general description of practical data sets, in contrast, the general Gaussian mixture model (Eq. (6)) provides a satisfactory description for practical data sets. In this case, IGA always located the same optimal solution in different runs with a computational requirement of merely 11 generations, while the conventional GA required 476 generations on the average to reach this optimum or its neighborhood.

The Iris data have been extensively studied in cluster analysis [3]. It has been discovered that the patterns of Iris Setosa form a cluster clearly separated from the patterns of Iris Versicolor and Iris Virginica, and the clusters of Iris Versicolor and Iris Virginica are to some extent overlapped. In clustering these data, the IGA for the square-error criterion and the IGA for the introduced criterion both correctly allocate the patterns of Iris Setosa into one cluster separated from the patterns of Iris Versicolor and Iris Virginica, however, the clustering results for the patterns of Iris Versicolor and Iris Virginica are distinguished significantly. To validate visually the clustering results of the patterns of Iris Versicolor and Iris Virginica, a two-dimensional representation of the patterns of Iris Versicoor and Iris Virginica is obtained by projecting these data onto the first two principal components (PCs), as is shown in Fig. 4. The variance accounted for by these two PCs is 99.85% of the original total variance, therefore, the resulting two-dimensional projections could be a fine representation of the original clustering structure. Using such a two-dimensional representation, one could present and validate the clustering results in a straightforward way. It is noteworthy that, if a two-dimensional representation with sufficient precision could not be obtained using PCA, one could take advantage of nonlinear mapping techniques [18] to produce an improved two-dimensional representation. The clustering results given by the IGA for the square-error criterion are shown in Fig. 4(a). It is observed that patterns 51, 53 and 78, which belong to Iris Versicolor, are grouped into the cluster majorly associated with Iris Virginica, and patterns 102, 107, 114, 115, 120, 122, 124, 127, 128, 134, 139, 143, 147 and 150, which belong to Iris Virginica, are allocated into the group majorly associated with Iris Versicolor. The results given by the IGA for the new criterion are shown in Fig. 4(b), where only five samples (69, 71, 74, 84 and 134) are grouped into the clusters inconsistent with their actual category labels. Since the new criterion seeks a quadratic or piece-wise quadratic boundary for clusters, the results obtained give evidence of the fact that the patterns of Iris Versicolor and

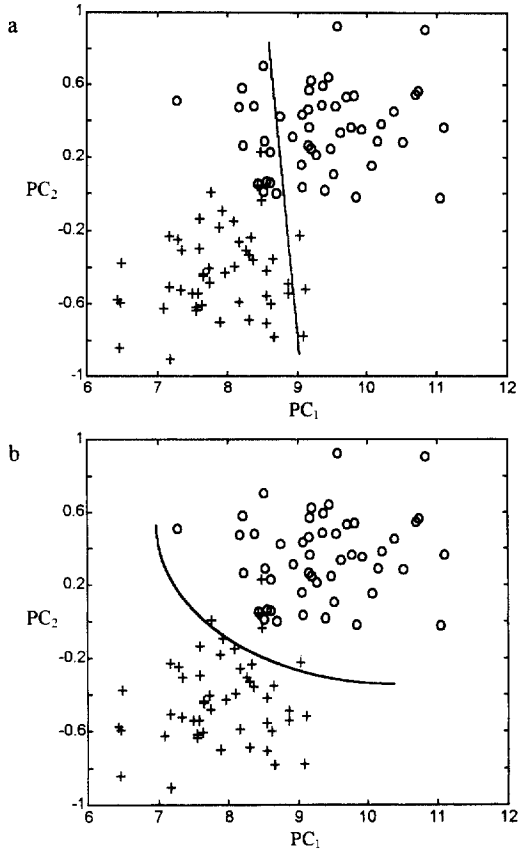

Fig. 4. Clustering results of the *Iris* data (*Iris Versicolor*, +; *Iris Virginica*, ○). The solid line or curve is the approximate cluster boundary given by the algorithm used. a. Clustering results obtained using the IGA for the square-error criterion. b. Clustering results obtained using the IGA for the introduced criterion.

Iris Virginica are essentially divided into two clusters by a quadratic surface. Compared with the misleading results given by the square-error criterion, these results are more consitent with the actual data structure. The advantage of the new clustering criterion is that it can adapt itself to the clustering structure inherent in the data involved, while the square-error criterion always imposes a certain structure on the data. In this example, IGA always located the same optimal solution in different runs with a computational requirement of merely 127 generations on the average, while the conventional GA required more than 600 generations to reach this optimum. The typical curves of the criterion values versus the generations for the proposed IGA and the conventional GA are shown in

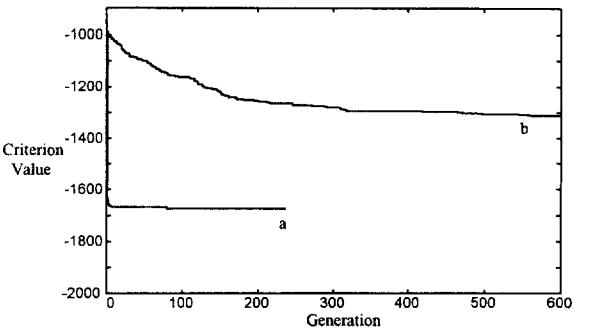

Fig. 5. Typical curves of the criterion values versus the generations for the developed IGA (a) and the conventional GA (b).

Fig. 5. One can observe that IGA shows much higher efficiency than the conventional GA.

It is noteworthy that the introduced clustering criterion (Eq. (7)) is more suitable for cluster analysis of data sets of which the ratio of sample number to the dimensionality is relatively large, since this criterion implicitly requires that all clusters of interest have nonsingular covariance matrices. In using this criterion for clustering data meeting such a requirement, the partitions which produce clusters having singular covariance matrices are inadmissible and should be assigned a sufficient large criterion value. In the presented work, therefore, the individuals which produce clusters having singular covariance matrices are assigned a very large criterion value, say 10 000, and the self-reproduction operation of these individuals is prohibited. In the case of applying this criterion to clustering data not meeting such a requirement, a preliminary dimensionality-reduction of data should be conducted to alleviate this problem. One could also take advantage of those methods, which have been proposed for the Friedman and Rubin's criterion to combat a similar problem [2].

5. Conclusions

This paper developed an improved GA with integer representation (IGA) and introduced a new clustering criterion for cluster analysis problems. The results presented show that the introduced criterion has an important advantage over the commonly used square-error criterion in that it can adapt itself to the data structure involved and has a very weak tendency to

look for clusters of particular structures. Compared with the conventional GAs, the developed IGA shows a much higher search efficiency and a better search precision.

Acknowledgements

The authors are grateful to the National Natural Science Foundation of China for financial support.

References

[1] R.-Q. Yu, Introduction to Chemometrics, Hunan Education Publishing House, Changsha, 1991.
[2] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ, 1988.
[3] D. Coomans, D.L. Massart, Anal. Chim. Acta 132 (1981) 225.
[4] D.L. Massart, L. Kaufman, D. Coomans, Anal. Chim. Acta 122 (1980) 347.
[5] P.K. Hopke, K. Kaufman, Chemom. Intell. Lab. Syst. 8 (1990) 195.
[6] H.P. Friedman, J. Rubin, J. Am. Stat. Assoc. 62 (1972) 1159.
[7] J.H. Kalivas, Chemom. Intell. Lab. Syst. 15 (1992) 1.
[8] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 19 (1993) 1.
[9] D.B. Hibbert, Chemom. Intell. Lab. Syst. 19 (1993) 277.
[10] C.B. Lucasius, G. Kateman, Chemom. Intell. Lab. Syst. 25 (1994) 99.
[11] M. Bos, H.T. Weber, Anal. Chim. Acta 247 (1991) 97.
[12] E. Fontain, Anal. Chim. Acta 265 (1992) 227.
[13] C.B. Lucasius, A.D. Dane, G. Kateman, Anal. Chim. Acta 282 (1993) 647.
[14] F. Maffioli, in: N. Christofides, A. Mingozzi, P. Toth, C. Sandi (Eds.), Combinational Optimization, Wiley, New York, 1979.
[15] J.-H. Jiang, J.-H. Wang, X.-H. Song, R.-Q. Yu, J. Chemometrics 10 (1996) 253.
[16] X. Liu, P.V. Espen, F. Adams, S.H. Yan, M. Vanbelle, Anal. Chim. Acta 282 (1993) 647.
[17] R.A. Fisher, Ann. Eugenics 7 (1936) 178.
[18] J.W. Sammon, IEEE Trans. Comput. 18 (1969) 401.