Journal of the Korean Institute
of Industrial Engineers
Vol.24, N0.1, March, 1998

A New Type of Clustering Problem with Two Objectives

복수 목적함수를 갖는 새로운 형태의 집단분할 문제

Jae-Yeong Lee

이재영

Abstract

In a classical clustering problem, grouping is done on the basis of similarities or distances (dissimilarities) among the elements. Therefore, the objective is to minmize the variance within each group while maximizing the between-group variance among all groups. In this paper, however, a new class of clustering problem is introduced. We call this a laydown grouping problem (LGP). In LGP, the objective is to minimize both the within-group and between-group variances. Furthermore, the problem is expanded to a multi-dimensional case where the two-way minimization process must be considered for each dimension simultaneously for all measurement characteristics. At first, the problem is assessed by analyzing its variance structures and their complexities by conjecturing that LGP is NP-complete. Then, the simulated annealing (SA) algorithm is applied and the results are compared against that from others.

Key Words: Simulated Annealing, Clustering Problem, Analysis of Variance

1. Introduction

The *clustering*, or grouping of similar objects together, is a well-known topic in mathematical and engineering fields. The objective of a cluster analysis is to separate a set of objects into constituent clusters in such a way that according to specified criterion, the members within each cluster differ from each other as little as possible while at the same time members of distinct clusters differ from each other as much as possible (Späth 1980). The techniques developed in conventional cluster analyses, however, are aimed at minimizing the variance within each cluster while maximizing the variance of the cluster means (that is, the "between-group variance"), with no restrictions on the number of objects in each cluster.

In general, the clustering problem in large-scale systems is known to be NP-complete (Nondeterministic Polynomial-complete) (Park et al. 1988); and thus we must rely on a heuristic method to obtain a good solution. Simulated Annealing (SA) is widely used to solve several types of clustering problems. Park et al. (1988) applied SA to the clustering problem for finding a partition with minimum number of vertex sets among feasible partition sets. Klein and Dubes (1989) reported on experiments in projection and clustering using the SA algorithm. By projection, it means a nonlinear mapping of patterns in high-dimensional Euclidean space into two-dimensional Euclidean space. Selim and Alsultan (1991) also applied SA to the problem of clustering 약n data points (patterns) into C clusters.

In the textile field, there is a special case of the clustering problem that requires that the sizes of the clusters are identical to each other and both the within-group and between-group variances are small. In cotton spinning, the laydowns of cotton bales are formed in order to maintain uniform yarn quality with the following objectives

1. The means of g laydowns (clusters) must be as nearly equal as possible with respect to all characteristics (e.g., fineness, strength, elongation, length, and color, etc.).
2. Simultaneously, the within-group variance of each laydown must be as small as possible with respect to each characteristic; moreover the within-group variances for all clusters must be as nearly equal as possible.

This problem is practical and yet theoretically challenging since the entire U.S. cotton crop is now high volume instrument (HVI) tested; and the information on cotton characteristics is available to all buyers and users, with little or no guidance on how the numbers should be analyzed to achieve the desired goals.

To the best of our knowledge, few studies have been done to provide a good solution for the laydown grouping problem (LGP). Recently, an investigation was performed on the topic by Robin and Suh (1993) at North Carolina State University supported by grants from Cotton Incorporated and the National Textile Center. They used the *branch and bound* (BAB) algorithm with several heuristic rules for the one-dimensional case (handling only one characteristic, strength), and they extended it to multidimensional cases that simultaneously account for two additional characteristics (namely, micronaire and length) by using a decent algorithm in order to optimize the remaining characteristics based on the given solution obtained for the one-dimensional case.

In this paper, we develop the theories about LGP and apply SA algorithm to this LGP. First, LGP is carefully assessed and then appropriate SA algorithm for LGP is developed. This paper is organized as follows. In Section 2, we analyze the problem structure and the computational complexity of the laydown grouping problem, and we conjecture that this problem is NP-complete. In Section 3, we apply SA to the LGP. Finally in Section 4 we draw conclusions and recommend further research.

2. Analysis of the Laydown Grouping Problem (LGP)

2.1 Variance Structure of LGP

Since both objectives in Section 1 are concerned with variances of the bale characteristics in LGP, it is necessary to analyze the variance structure of LGP. Suppose we have N bales to be divided into g clusters (laydowns), and each laydown has n bales. Therefore, $N = n \times g$. Let m_l the lth laydown mean, let $\overline{m}$ be the grand mean, let V_l be the lth laydown within-group variance, let V_{bet} betbe the between-group variance of all laydowns, and let C_{lj} be the value of one characteristic (for example, strength) of the jth bale in the lth laydown. Then we can easily obtain the following statistics from HVI data --

$$m_l = \frac{1}{n}\sum_{j=1}^{n} C_{lj}, \quad \overline{m} = \frac{1}{gn}\sum_{l=1}^{g}\sum_{j=1}^{n} C_{lj} \qquad (1)$$

$$V_l = \frac{1}{n-1}\sum_{j=1}^{n}(C_{lj}-m_l)^2, \quad V_{bet} = \frac{1}{g-1}\sum_{l=1}^{g}(m_l-\overline{m})^2. \qquad (2)$$

Using the additive property of the sum of squares in the analysis of variance (ANOVA), we obtain the total sum of squares, TSS, by adding the sum of squares for treatments, SSTR, and the error sum of squares, SSE, as follows:

$$\text{TSS} = \text{SSTR} + \text{SSE}$$

$$\begin{aligned} &= n\sum_{l=1}^{g}(m_l-\overline{m})^2+\sum_{l=1}^{g}\sum_{j=1}^{n}(C_{lj}-m_l)^2 \\ &= n(g-1)V_{bet}+(n-1)\sum_{l=1}^{g}V_l. \end{aligned}$$

Let AV = $\frac{1}{g}\sum_{l=1}^{g}V_l$ and BV = V_{bet}, then we have the linear relationship between AV and BV as follows:

$$\begin{aligned} \text{TSS} &= n(g-1)\,\text{BV}+g(n-1)\text{AV}. \\ \text{AV} &= -\frac{n(g-1)}{g(n-1)}\,\text{BV}+\frac{\text{TSS}}{g(n-1)} = -a\text{BV}+b. \end{aligned}$$

This relationship is shown in Figure 1. Thus, if we minimize AV then BV is maximized, and vice versa. Therefore, if either AV or BV is given, our task is to seek the best set of clusters at the point p on the line in figure 1.

to minimize the largest within-group variance among all laydowns. To formalize this, we arbitrarily number the bales 1, 2, ..., N; and we let C_i denote the characteristic of interest for bale $i(i = 1, ..., N)$ so that

$C_{lj} = C_i$ if and only if bale i is the jth bale in the lth laydown,

for $l = 1, ..., g$, $j=1, ..., n$, and $i=1, ..., N$. We let $X=[X_{il}]$ be a $N\times g$ binary matrix defining an assignment of bales to laydowns so that

$$X_{il} = \begin{cases} 1, & \text{if bale } i \text{ is assigned laydown } l, \\ 0, & \text{otherwise,} \end{cases}$$

for $1\le i\le N$ and $1\le l\le g$. Thus we can formulate the quantities m_l and V_l (for $l = 1, ..., g$) together with $\overline{m}$, V_{bet}, and AV as function of the bale-assignment matrix X

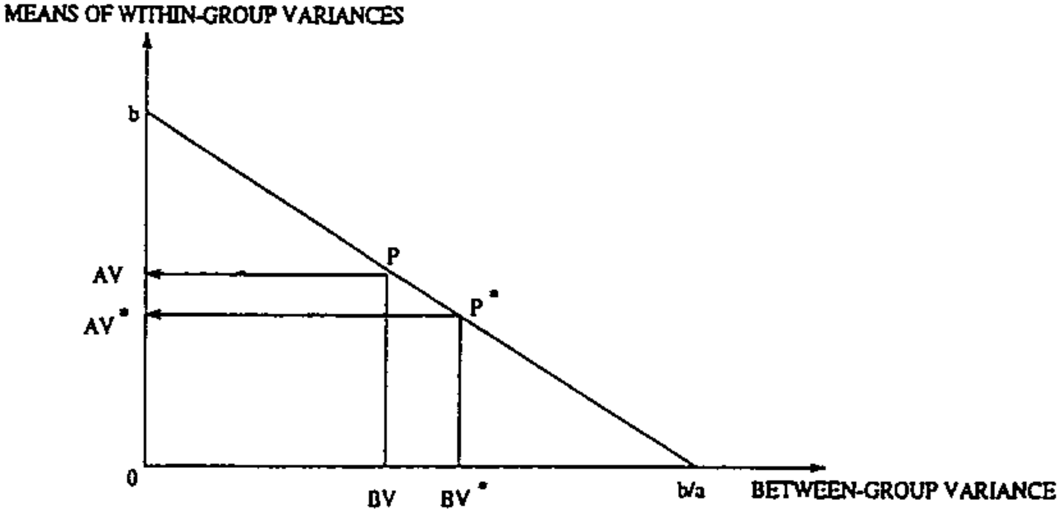

Figure 1. Linear relationship between AV and BV

2.2 Objective Function of LGP

Because of the linear relationship in Figure 1. with a given value of BV, AV is immediately determined. That is, if a between-group variance is fixed, the first objective of LGP in Section 1 is fixed. Therefore,

what we have to do is to satisfy the second objective in Section 1. One way to meet this second objective is

as follows:

$$m_l = m_l(X) = \frac{1}{n}\sum_{i=1}^{N}X_{il}C_i \quad \text{for } l = 1, ..., g, \tag{3}$$

$$V_l = V_l(X) = \frac{1}{n-1}\sum_{i=1}^{N}[X_{il}C_i-m_l(X)]^2, \quad \text{for } l = 1, ..., g, \tag{4}$$

$$\overline{m} = \overline{m}(X) = \frac{1}{g}\sum_{l=1}^{g}m_l(X) = \frac{1}{gn}\sum_{i=1}^{N}\sum_{l=1}^{g}X_{il}C_i, \tag{5}$$

$$V_{bet} = V_{bet}(X) = \frac{1}{g-1}\sum_{l=1}^{g}[m_l(X) - \overline{m}(X)]^2 \qquad (6)$$

$$AV = AV(X) = \frac{1}{g}\sum_{l=1}^{g}V_l(X). \qquad (7)$$

Thus, by using the notation in (3)-(7), we can write a mathematical model of LGP as follows:

$$\text{(Q1)} \quad \text{Minimize} \quad Z(X) = \max_l V_l(X) \qquad (8)$$

$$\begin{aligned} \text{subject to } & \sum_{i=1}^{N} X_{il} = n = N/g, && l = 1,\dots, g, \\ & \sum_{l=1}^{g} X_{il} = 1, && i = 1,\dots, N \\ & V_{bet}(X) \le BV^*, && \\ & X_{il} \in \{0, 1\} && \text{for } i = 1,\dots, N \text{ and } l = 1,\dots, g, \end{aligned}$$

where BV* is an upper limit on the between-group variance and its corresponding value, AV*, is the associated lower limit on the average within-group variance. Note that the inequality constraint $V_{bet}(X) \le BV^*$ allows us to trade some portion of total variance from the between-group variance to the within-group variances (for example, in Figure 1, p^* moved to p). Now we set up a definition of *better solutionto meet* both objectives in Section 1, simultaneously.

Let AV* be the average within-group variance corresponding to a given upper limit BV* Consider two arbitrary solution matrices $X^\dagger$ and X of the problem Q1 so that

$$V_{bet}(X) \le BV^* \text{ and } V_{bet}(X^\dagger) \le BV^*.$$

However, it is *not* necessarily true that

$$V_{bet}(X) = V_{bet}(X^\dagger) \text{ or } AV(X) = AV^* \text{ or } AV(X^\dagger) = AV^*.$$

Definition 1. Let $V^\dagger$ and V be the vectors of within-group variances obtained by $X^\dagger$ and X, respectively: $V \equiv [V_1(X),\dots, V_g(X)]$ and $V^\dagger \equiv [V_1(X^\dagger),\dots, V_g(X^\dagger)]$. If

$$V_{bet}(X) \le BV^*, \qquad V_{bet}(X^\dagger) \le BV^*,$$

and

$$\sum_{l=1}^{g}[V_l(X^\dagger) - AV^*]^2 \ < \ \sum_{l=1}^{g}[V_l(X) - AV^*]^2,$$

then $X^\dagger$ is a *better solution to* problem Q1 than X. □

This definition is based on a preference for more consistent (that is, less variable) within-group variances relative to the target value AV*, even if the average within-group variance $AV^\dagger$ for the bale assignment $X^\dagger$ is larger than the average within-group variance for the bale assignment X.

According to the definition above, we introduce an alternative objective function of Q1 as follows:

$$\text{(Q1)}^\dagger \quad \text{Minimize} \quad Z^\dagger(X) = \sum_{l=1}^{g}[V_l(X) - AV^*]^2 \qquad (9)$$

$$\begin{aligned} \text{subject to } & \sum_{i=1}^{N} X_{il} = n = N/g, && l = 1,\dots, g, \\ & \sum_{l=1}^{g} X_{il} = 1, && i = 1,\dots, N, \\ & V_{bet}(X) \le BV^*, && \\ & X_{il} \in \{0, 1\} && \text{for } i = 1,\dots, N \text{ and } l = 1,\dots, g. \end{aligned}$$

The new objective function (9) in problem $Q1^\dagger$ avoids extreme disparities among the within-group variances. Note that the objective function (8) of problem Q1 does not prevent such disparities since it only tends to minimize the largest within-group variance. However, (9) is not guaranteed to yield a solution to the LGP that is always preferable in practice to the solution obtained with (8).

Suppose there exists a *perfect solution* whose all within-group variances are the same while satisfying $V_{bet} = BV^*$. Then, this perfect solution is the only optimal solution considered as the best for both problems Q1 and $Q1^\dagger$. Therefore, as long as the perfect solution is concerned, both objective functions (8) and (9) are asymptotically the same. However, since there is no such a perfect solution in a real problem, which objective function is better is the matter of user's preference in practical point of view.

2.3. Complexity of LGP

Note that the objective function in Q1 contains the variance terms $V_l(X)$ which make it extremely difficult to compare the complexity of Q1 with other optimization problems. In order to find the complexity of LGP, we establish a lemma which tells us that any sample variance term can be represented as the sum of squared distances between all (distinct) pairs of the sample values.

Lemma 2. For any set of n sample values $\{x_i \in \Re$, i = 1,..., n$\}$ with sample mean $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$, we have

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \equiv (1/2n) \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2.$$

Proof. Let I_n denote the $n \times n$ identity matrix, and let U_n denote the $n \times n$ matrix with every element equal to 1. Let

$X \equiv [x_1, x_2, \ldots, x_n]$

denote the $1 \times n$ vector where ith element is x_i, i = 1, 2,...,n. It is easy to check that

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = x(I_n - \frac{1}{n} U_n)\, x' \tag{10}$$

and that

$$\begin{aligned} \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2 &= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i^2 - 2x_i x_j + x_j^{2)} \\ &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j \\ &= xI_n x' - \frac{1}{n} xU_n x' \\ &= x(I_n - \frac{1}{n} U_n) x' \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad by\,(10), \end{aligned}$$

which completes the proof. ◻

This lemma allows us to reexpress any variance $V_l(X)$ in terms of a quadratic form whose complexity is easy to analyze. Using Lemma 2, both variance terms $V_l(X)$ (in the objective function) and $V_{bet}(X)$ (in the third constraint) of problem Q1 can be replaced as follows:

$$(Q1^{\ddagger}) \quad \text{Minimize} \quad Z^{\ddagger}(X) = \max_{l} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij}^2 X_{il} X_{jl} \tag{11}$$

$$\text{subject to} \quad \sum_{i=1}^{N} X_{il} = n = N/g, \qquad l = 1,\ldots, g, \tag{12}$$

$$\sum_{l=1}^{g} X_{il} = 1, \qquad i = 1,\ldots, N, \tag{13}$$

$$\frac{1}{2g(g-1)} \sum_{u=1}^{g} \sum_{v=1}^{g} D_{uv}^2(X) \leq BV^*, \tag{14}$$

$$X_{il} \in \{0, 1\} \qquad \text{for all } i, l, \tag{15}$$

where d_{ij} and $D_{uv}(X)$ are the distances of the characteristic values of two bales C_i and C_j -- i.e.,

$$d_{ij} = |C_i - C_j| \quad \text{for } i = 1,\ldots,N \text{ and } j = 1,\ldots,N;$$

and of the two laydown means $m_u(X)$ and $m_v(X)$ -- i. e.,

$$\begin{aligned} D_{uv}(X) &= |m_u(X) - m_v(X)| \\ &= \frac{1}{n} \left| \sum_{i=1}^{N} C_i X_{iu} - \sum_{j=1}^{N} C_j X_{jv} \right| \quad \text{for } u = 1,\ldots, g \text{ and } v = 1,\ldots, g. \end{aligned}$$

Note that the constant term $1/[2N(N-I)]$ in $Z^{\ddagger}(X)$ is deleted and the LHS of the constraint (14) is nothing but between-group variance $V_{bet}(X)$. Note also that d_{ij}^2 is counted only if both X_{il} and X_{jl} variables are not equal to zero (i.e., when both the ith and the jth bales are in lth laydown).

Now, based on the linear relationship between AV and BV in Figure 1, we build a following assumption.

Assumption 3. If we interchange the objective function (11) and the constraint (14) in $Q1^{\ddagger}$ such that

$$\begin{aligned} (Q1^{*}) \quad \text{Minimize } Z^{*}(X) &= \sum_{u=1}^{g} \sum_{v=1}^{g} D_{uv}^2(X) \\ &= \frac{1}{n^2} \sum_{u=1}^{g} \sum_{v=1}^{g} [\sum_{i=1}^{N} C_i X_{iu} - \sum_{j=1}^{N} C_j X_{jv}]^2 \end{aligned}$$

$$\text{subject to } \sum_{i=1}^{N} X_{il} = n = N/g, \qquad l = 1,\dots, g$$

$$\sum_{l=1}^{g} X_{il} = 1, \qquad i = 1,\dots, N$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^{2} X_{il} X_{jl} \leq WV^{*} \qquad l = 1,\dots, g$$

$$X_{il} \in \{0, 1\} \qquad \text{for all } i, l,$$

where WV* is a upper limit of all within variances, then both problems $Q1^{\ddagger}$ and Q1* have the same computational complexity. □

The Assumption 3 is based on the following reasons. First, both problems $Q1^{\ddagger}$ and Q1* have the same structure of feasible solution spaces. That is, in $Q1^{\ddagger}$, we seek the configuration of laydown grouping which has the smallest UV (where UV is the largest within-group variance of that configuration) given by an upper limit of BV as $BV^{\ddagger}$ that defines immediately the lower limit of AV by $AV^{\ddagger}$. On the other hand, in Q1*, we seek the configuration of laydown grouping which has the smallest EV given by an upper limit of V_l as WV* for all l. Then by the Samuelson's inequality (Samuelson 1968)

$$|V_l - AV| \leq s\sqrt{g-1} \textit{ where } s = \sqrt{1/(g-1) \sum_{l=1}^{g} (V_l - AV)^2},$$

we have the lower limit of AV as AV* that defines immediately the upper limit of BV by BV*. Therefore, when $BV^{\ddagger}$ = BV* and UV = WV* we have exactly the same feasible solution space for both problems.

Secondly, the objective functions of both problems seek a perfect solution (if it exists) whose within-group variances are all equal. That is, the perfect solution minimizes both objective functions.

Based on the Assumption 3, we claim the following.

Conjecture 4. The LGP is NP-complete. □

In order to prove that a problem is NP-complete, we must show two things (Papadimitriou and Steiglitz 1982, page 353):

(a) That the problem is in NP.

(b) That a known NP-complete problems is polynomially transformable to the problem at hand.

For (a), it is clear that LGP is in NP (which is short for Nondeterministic Polynomial bounded) since a given proposed solution matrix X can be checked quickly (in polynomial time) to see if it satisfies all the requirements of the problem. That is, the number of steps to check if X is in feasible region of constraints (12), (13), and (14) in $Q1^{\ddagger}$ is $Ng + Ng + 4Ng(g-1)$ which is bounded above by $4Ng^2-2Ng = O(ng^3)$. For (b), we use the 2-Partition problem that is NP-complete(Garey and Johnson 1979, page 47) to see if the 2-Partition can be polynomially transformable to the LGP represented in problem Q1*. Note that we consider the problem Q1* instead of the problem $Q1^{\ddagger}$ based on the Assumption 3. Thus, we now minimize the between-group variance instead of minimizing the maximum within-group variance. Consider the 2-Partition problem which asks, given N integers $A_1,\dots, A_N$, if there is a set $S \subseteq \{1, 2,\dots, N\}$ such that

$$\sum_{j\in S} A_j = \sum_{j\in \overline{S}} A_j = \frac{1}{2}\sum_{j=1}^{N} A_j \tag{16}$$

where $\overline{S} = \{1, 2,\dots, N\} - S$.

Since the 2-Partition problem is a decision problem we also changed Q1* into a decision problem by setting $Z^*(X) \leq K$, where K is an arbitrary constant. More specifically, if $g=2$ and $Z^*(X) \leq K$, then we have a decision problem asking that, given N bales whose characteristic values are $\{C_1, C_2,\dots, C_N\}$, is there a subset $S_c \subseteq \{1, 2,\dots, N\}$ such that

$$\sum\sum_{i,j\in S_C} d_{ij}^2 = \sum\sum_{i,j\in \overline{S_C}} d_{ij}^2 = AV \tag{17}$$

where $\overline{S_C} = \{1, 2,\dots, N\} - S_c$ and AV is the average of two within-group variances. The reason we seek a solution to satisfy (17) is that the best solution of breaking N bales into two clusters (with a fixed value of Z^*) is dividing them in such a way that both within-group variances are identical (that is, we seek a perfect solution). Note that the case of $g=2$ is the simplest problem of LGP. It is clear that (16) is much simpler than (17), but it is not easy to show that 2-Partition is polynomially trasformable to LGP with $g=2$. In other words, a main task we have to do is to answer the following question.

Question 5. Given a set $S \subseteq \{1, 2, ..., N\}$, can we construct a set $S_c \leq \{1, 2, ..., N\}$ within polynomial time such that S and $\overline{S}$ are a solution of 2-Partition if and only if S_c and $\overline{S}_C$ are a solution of LGP with $g=2$ and $Z^*(X) \leq K$? □

If the answer is yes, then 2-Partition problem is nothing but a special case of Q1* when $g=2$, AV = $\sum_{j=1}^{n} A_j / 2$ and $Z^*(X) \leq K$.

Furthermore, the original optimization problem Q1* can be solved in logarithmic time (which is faster than polynomial time) by changing K value in the decision problem of Q1*. For example, let K^* be the optimal integer value of $Z^*(X)$ such that $K^* \in [0, \overline{K}]$. Then if $\overline{K} = 16$, we can point out exact value of K^* at most four times of comparisons by using the bisection search method. Consequently, based on the assumption that the Question 5 can be answered, we conjecture that LGP is NP-complete.

As we discuss for the conjecture above, LGP is interpreted as the partitioning problem because we have to partition N bales into g laydowns. LGP can also be interpreted as another type of knapsack problem because we have to minimize $Z(X)$ by allocating N bales to g laydowns with respect to the upper limit on BV*. Note that a 0-1 Knapsack problem can be polynomially transformed to 2-Partition problem (Papadimitriou and Steiglitz 1982, page 375), which means that a 0-1 Knapsack problem is a special case of 2-Partition problem (of course, a 0-1 Knapsack problem is also NP-complete).

More analysis for the computational complexity of LGP is done by Lee (1995). Let ν be the number of distinct groupings in LGP, then Lee (1995) drives that ν can be expressed as an explicit function of N and n as follows:

$$\nu = \frac{N!}{(N/n)!(n!)^{N/n}}. \tag{18}$$

The function ν grows faster than exponential growth when $N \geq 20$ and n = 2, 4. Theoretically, Lee (1995) also showed the exponential growth of ν in LGP when g is fixed and let $n \to \infty$ so $N = ng \to \infty$.

3. Application of SA to LGP

3.1 Simulated Annealing (SA) Algorithm

The Simulated Annealing (SA) algorithm, introduced independently by Kirkpatrick et al. (1983) and Cěrny (1985), has performed successfully as a general heuristic algorithm for the solution of large, complex combinatorial optimization problems. A combinatorial optimization problem is designated as a pair (Ω, f) where Ω is the solution space and the cost function $f: \Omega \to \Re^+$ *assigns to each element of* Ω *a nonnegative real-valued cost. The minimum of* f over the set Ω is sought. An optimal solution of the problem is obtained once we find an element $i^* \in \Omega$, called a *global minimum*, with the property $f_{i^*} \leq f_i$ for *all* $i \in \Omega$. The strategy implemented by SA consists of exploring the solution space starting from an arbitrary selected solution, or state, and generating a neighboring state j by perturbing the current state i. Every time a new solution j is generated, its cost f_j is evaluated; and the new solution is either accepted with probability A_{ij} or rejected with probability $(1 - A_{ij})$ according, say, to the *Metropolis criterion* proposed by Metropolis et al. (1953),

$$A_{ij} = \begin{cases} 1, & \text{for } \Delta f_i \leq 0 \\ \exp(-\Delta f_i / T), & \text{for } \Delta f_i > 0, \end{cases}$$

where Δf_i is the difference between the costs at the new state j and the current state i (i.e., $\Delta f_i = f_j - f_i$), and T is a positive control parameter called the temperature.

Thus, there is a nonzero probability of continuing with a state having higher cost than the current state, which provides the reason why the SA algorithm can avoid being trapped at local minima. This sequence of trials is repeated until equilibrium at each temperature T is reached -- i.e., until the probability distribution of the configurations (= states) in Ω approaches the Boltzmann distribution, which is given by

$$\Pr\{\text{current configuration} = i\} \overset{\text{def}}{=} \pi_i(T) = \frac{1}{Z(T)}\exp(-f_i / T) \quad \text{for } all\ i \in \Omega,$$

where $Z(T) = \sum_{i \in \Omega} \exp(-f_i / T)$ is a normalizing factor depending upon the temperature T. The temperature T is lowered in steps until it approaches zero, with the system being allowed to reach equilibrium for each step by generating a sequence of trials in the previously described way. After termination, the final *frozen* configuration is taken as the optimal solution of the combinatorial optimization problem at hand. See Eglese (1990) for general applications of the SA algorithm.

3.2 Objective Function

Based on the linear relationship between AV and BV, it is easy to see that if we represent a given solution by the two coordinates BV and UV, where UV defines the upper limit (maximum) for all within-group variances, all feasible solutions must belong to a region that includes and is above the line AB in Figure 2, which shows the solution structure of laydown grouping problem (LGP).

Note that the vertical axis of Figure 2 represents UV, which is different from Figure 1. The point C is a extreme case that has maximum within-group variance when BV = 0 (if it exists). For any point that belongs to the line AB, the within-group variances must be all equal (i.e., UV = V_l for all l) since UV = AV on the line AB. Thus, by the definition of *better solutionin* Section 2.2, any solution on the line AB will be the ideal case. However, it is very difficult to have such an ideal solution with a large number of bales in a practical situation. Therefore, we define this ideal solution as a *lower boundfor* LGP. Line BC represents the worst case where the largest UV can possibly be made with a given between-group variance BV. Note that BC is not necessarily a straight

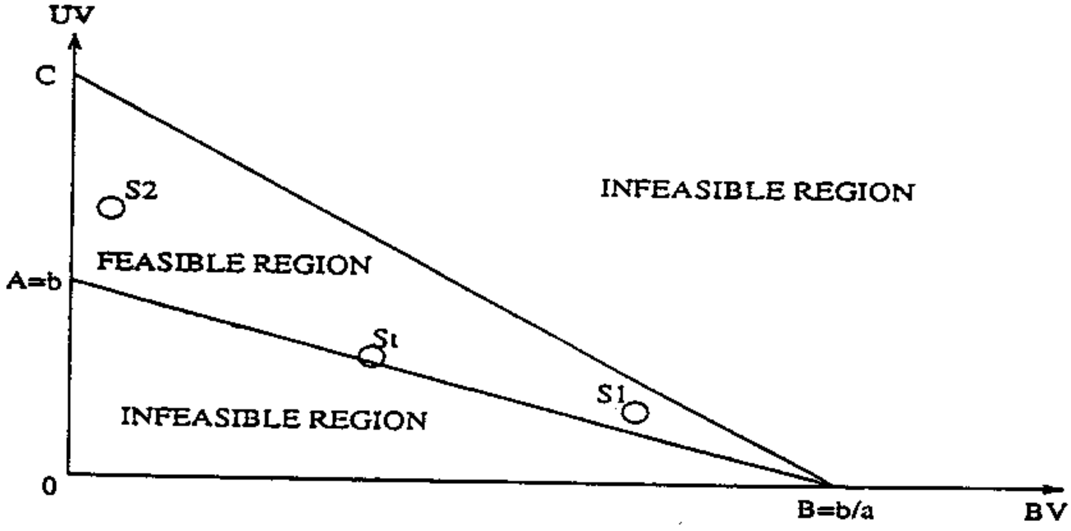

Figure 2. Solution structure of LGP

line.

For the purpose of evaluating the performance of a given set, we consider the following two trivial solutions S_1 and S_2.

1. S_1: first, sort the bale characteristic values in increasing order; then assign the first n bales to laydown 1, the next n bales to laydown 2, and so on.
2. S_2: first, sort the bale characteristic values in increasing order; then assign the first $n/2$ bales and the last $n/2$ bales to laydown 1; and repeat this procedure with the remaining N-n bales, and so on.

These two solutions are shown in Figure 2. Indeed, S_1 exemplifies a case where UV is made small at the expense of BV, whereas S_2 shows a case where BV is made small at the expense of UV. We also depict a target solution S_t which is the lower bound of all feasible solutions. Based on two extreme solutions S_1 and S_2, it is clear that the optimal solution should be somewhere between these two solutions. In problem Q1, we try to minimize UV with a given upper limit for BV, which tackles the problem in one direction. However, there should be a way to attack the problem from both directions by minimizing a weighted sum of UV and BV. Therefore, to make an optimal solution as close to S_t as possible, we want to minimize both UV and BV simultaneously by using the control parameters k_1 and k_2. Thus, the objective function can be written as

$$Z_{LGP} = k_1 UV + k_2 BV. \tag{19}$$

This objective function is identical to the function (7.22) - (7.23) in Lee (1995) with $k_1 = 2N(N-1)\alpha_1$ *and* $k_2 = 2g(g-1)n^2 \alpha_2$; and we use (19) as an objective function in our experiments.

3.3. Experimental Results Using SA Algorithm

Recall the SA procedure explained in Section 3.1. At first, we describe a solution-generation mechanism. Let G = [G_{ij}] be a $|\Omega| \times |\Omega|$ stochastic matrix -- i.e. all elements of G are nonnegative and for every solution (state) $i \in \Omega$, $\sum_j G_{ij} = 1$, where G_{ij} is the probability of selecting j as the destination state when moving from the current state i. We assume that G_{ij} is independent of T_k in such a way that

$$G_{ij} = \begin{cases} 1/|\Omega_i|, & \text{if } j \in \Omega_i \\ 0, & \text{otherwise} \end{cases},$$

where Ω_i is the set of all states j to which we can move starting from state i. In order to select a neighboring state j, we choose randomly two bales from N samples and swap these two bales.

For an annealing schedule in LGP, the temperature T_k is updated by the geometric law

$$T_{k+1} = \alpha T_k \qquad \text{with } 0 < \alpha < 1.$$

Initial temperature T_0=125 is used. At each temperature T_k, the number of replications required to reach an equilibrium (or stationary) state is called n_{equil}. Through sensitivity analysis, we select the parameter values as of $n_{equil} > 150$ and α=0.90. We also choose the values of control parameters in (19) as of k_1=1.0 and k_2=g.

In this experiment, actual HVI data are used with N= 754, n=29, g=26. For simplicity, before its usage, these HVI data are standardized by subtracting the overall mean from each data item; and then the resulting difference is divided by the standard deviation of the entire data set. We compared our SA results to Robin's earlier work in the 3-dimensional case (using three characteristics, strength, micronaire, and length), when the target within-group variance (TWV) in Robin's model is twv = 1.2. Note that twv is represented by UV in our notation. In order to handle the 3-dimensional case, we use a weighted average of the three characteristics. Thus, the objective function (19) has been changed to

$$Z_{LG_P}(W) = \sum_{d=1}^{D} W_d [k_1 UV_d + k_2 BV_d]. \qquad (20)$$

where D = 3 and W_d = 1/3 for d =1,2,3.

Figures 3 and 4 represent the laydown means and within-group variances for the characteristic of micronaire, respectively. Figures 3 and 4 show that SA results are definitely superior to the Robin's results on both laydown means and within-group variances. In other words, SA results have more consistent laydown means and less variation in within-group variances of all laydowns than Robin's results do. We also plotted the initial grouping which is directly clustered from a given HVI data without any treatment. This depicts how much SA result is improved from initial grouping. Similar results are obtained for the other two characteristics, strength and length.

4. Conclusions and Further Research

We introduced a new class of clustering problem with a multicriteria objective function called laydown grouping problem (LGP); and we analyzed the structure of LGP and its complexity. A definition of a better solutionin LGP is proposed. We also conjectured that LGP is a NP-complete combinatorial optimization problem. The SA algorithm is applied to LGP using real HVI data. Based on our experiments, it has been shown that the SA algorithm outperforms the previous methods with respect to both the laydown means and the laydown within-group variances.

We also bring up several issues for future research. First, parametric analysis is needed mainly for the solution-generation mechanism and the annealing schedule to improve the performance of the SA algorithm in LGP. Secondly, some other newly developed techniques such as *neural networks* for combinatorial optimization problems can be applied to LGP, and the results obtained with these techniques should be compared with those of SA

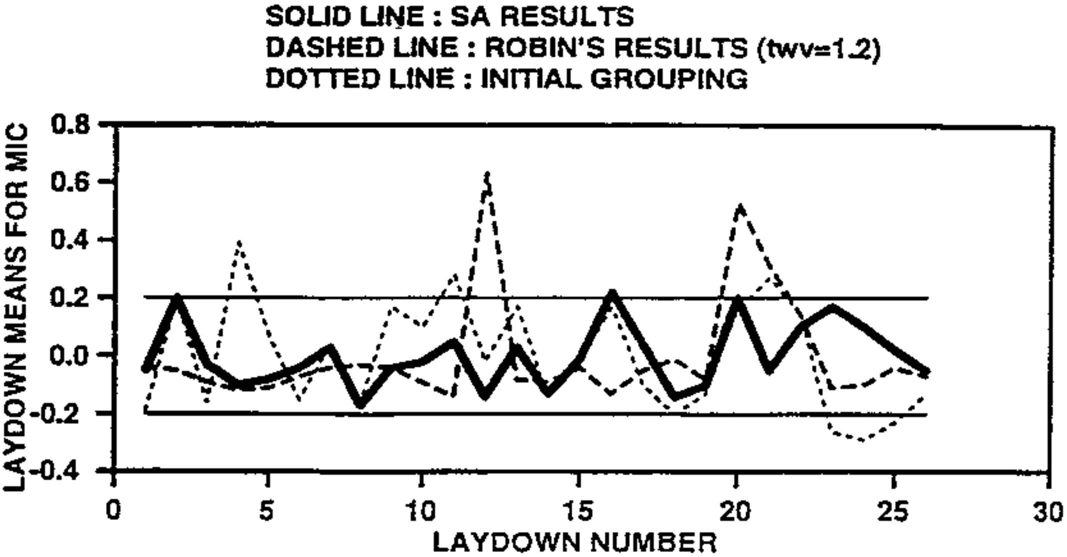

Figure 3. Comparison of laydown means for *Micronaire*

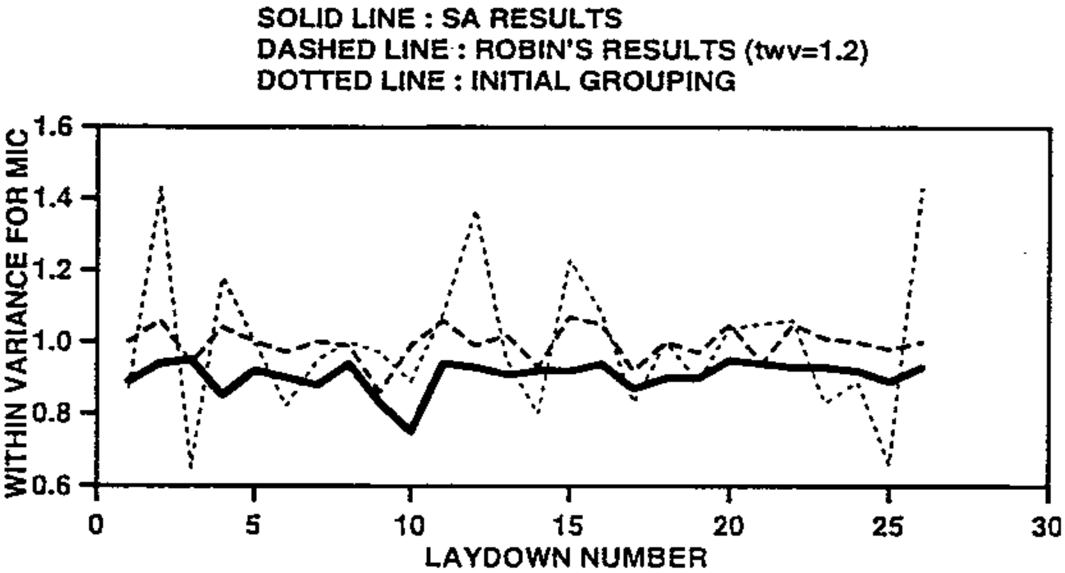

Figure 4. Comparison of within-group variances for *Micronaire*

algorithm. Thirdly, the SA procedure of LGP can be applied to other similar practical problems such as --

1. Military personnel and equipment assignment problems to keep all unit forces as nearly equal as possible in combat effectiveness.
2. Student and teacher assignments in the public school system to achieve more uniform quality of all schools.
3. Grape-blending problems to achieve uniform wine quality.

Consequently, LGP is a theoretically challenging and practically rewarding problem.

References

[1] Cerny, V. (1985). "Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm," *J. of Optimization Theory and its Application*, Vol. 45. pp. 41-51.

[2] Eglese, R. W. (1990). "Simulated Annealing: A Tool for Operations Research," *European J. of OR*, Vol. 46, pp. 271-281.

[3] Garey, M. R., and D. S. Johnson. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, San Francisco.

[4] Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. (1983). "Optimization by Simulated Annealing." *Science*, Vol. 220, pp. 671-680.

[5] Klein, R. W., and R. C. Dubes. (1989). "Experiments in Projection and Clustering by Simulated Annealing," *Pattern Recognition*, Vol. 22, No. 2, pp. 213-220.

[6] Lee, J. Y. (1995). "Faster Simulated Annealing Techniques for Stochastic Optimization Problems, with Application to Queueing Network Simulation," Ph.D. Thesis in the Department of Statistics and the Graduate Program in Operations Research in North Carolina State University, Raleigh, NC 27695-7913.

[7] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth. A. H. Teller, and E. Teller. (1953). "Equation of State

Calculations by Fast Computer Machines," *J. Chem. Phys.*, Vol. 21, pp. 1087-1092.

[8] Papadimitriou, C. H., and K. Steiglitz. (1982). *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall. Inc., Englewood Cliffs, New Jersey 07632.

[9] Park, C.-I., K.-H. Park, and M.-H. Kim. (1988). "Simulated Annealing Method for Clustering Problem in Large-Scale Systems," *IEEE Int. Symposium on Circuit and Systems*, Vol. 3, pp. 2347-2350.

[10] Robin, F. and M. W. Suh. (1993). "A Heuristic Algorithm for the Bale Assignment Problem," Master's Thesis in Operations Research Program in North Carolina State University, Raleigh, NC 27695-7913.

[11] Samuelson, P. A. (1968). "How Deviant Can You Be?" *J. of the American Statistical Association*, Vol. 63, pp. 1522-1525.

[12] Selim, S. Z., and K. Alsultan. (1991). "A Simulated Annealing Algorithm for the Clustering Problem," *Pattern Recognition*, Vol. 24, No. 10, pp. 1003-1008.

[13] Späth, H. (1980). *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. John Wiley & Sons.

96년 11월 최초 접수, 97년 3월 최종 수정