Constrained clusters of gene expression profiles with pathological features

Jun Sese¹, Yukinori Kurokawa³, Morito Monden⁴, Kikuya Kato³ and Shinichi Morishita¹,²

¹Undergraduate Program for Bioinformatics and Systems Biology and ²Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Bunkyo, Tokyo, Japan, ³Department of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara, Japan and ⁴Department of Surgery and Clinical Oncology, Graduate School of Medicine, Osaka University, Suita, Osaka, Japan

Received on December 28, 2003; revised on April 22, 2004; accepted on June 18, 2004
Advance Access publication June 24, 2004

ABSTRACT
Motivation: Gene expression profiles should be useful in distinguishing variations in disease, since they reflect accurately the status of cells. The primary clustering of gene expression reveals the genotypes that are responsible for the proximity of members within each cluster, while further clustering elucidates the pathological features of the individual members of each cluster. However, since the first clustering process and the second classification step, in which the features are associated with clusters, are performed independently, the initial set of clusters may omit genes that are associated with pathologically meaningful features. Therefore, it is important to devise a way of identifying gene expression clusters that are associated with pathological features.

Results: We present the novel technique of ‘itemset constrained clustering’ (IC-Clustering), which computes the optimal cluster that maximizes the interclass variance of gene expression between groups, which are divided according to the restriction that only divisions that can be expressed using common features are allowed. This constraint automatically labels each cluster with a set of pathological features which characterize that cluster. When applied to liver cancer datasets, IC-Clustering revealed informative gene expression clusters, which could be annotated with various pathological features, such as ‘tumor’ and ‘man’, or ‘except tumor’ and ‘normal liver function’. In contrast, the k-means method overlooked these clusters.

Contact: sesejun@gi.k.u-tokyo.ac.jp
Supplementary information: Our dataset is available on the following web page: http://love2.aist-nara.ac.jp/laboratory/data_download.html.

1 INTRODUCTION
Gene expression profiles that are generated using microarrays have many applications in biology, pharmacology and medicine. One of the most important of these applications is the diagnosis of disease, and various data-mining methods using gene expression profiles have been proposed for classifying disease (Golub et al., 1999; Tamayo et al., 1999; Alizadeh et al., 2000; Chung et al., 2002). In some cases, the classification may be uninformative particularly when each type of disorder includes similar pathological variants.

1.1 Gene expression profiling for disease treatments
Diseases are frequently classified using pathological information. Table 1 shows examples of pathological features that involve the liver. In this table, HBV+ means positive for hepatitis B, which is one of the most common liver diseases. Hepatitis C infection (HCV+) and cirrhosis are other well-known liver diseases. The pathological features also include same properties, such as normal liver, tumor cell and man. Variants may be distinguished by clustering the gene expression profiles, since cells that manifest the same disease show similar gene expression profiles. However, clustering provides little information on the optimal treatment choices, since clustering does not consider pathological data. Hence, there is an urgent need for a new method that combines the clustering of gene expression profiles with pathological information.

In this paper, most of the pathological information is expressed in Boolean, e.g. the characters are either positive or negative, and associations are made between the conjunctions of pathological features and clusters of gene expression profiles, thereby allowing the identification of variants. A striking example for the applicability of IC-Clustering is the discovery of the rule that the conjunction of pathological features ‘tumor’ and ‘man’ forms a specific cluster of...
Table 1. Pathological features involving the liver

<table>
<thead>
<tr>
<th>Normal liver</th>
<th>Abnormal liver</th>
<th>Tumor cell</th>
<th>Except tumor cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBV+</td>
<td>HBV−</td>
<td>HCV+</td>
<td>HCV−</td>
</tr>
<tr>
<td>Over 65 years old</td>
<td>Not over 65 years old</td>
<td>Man</td>
<td>Woman</td>
</tr>
</tbody>
</table>

Table 2. Example table

<table>
<thead>
<tr>
<th>Feature items (Pathological information)</th>
<th>Objective attributes (Gene expression patterns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i1</td>
<td>a1</td>
</tr>
<tr>
<td>i2</td>
<td>a2</td>
</tr>
<tr>
<td>i3</td>
<td>a1</td>
</tr>
<tr>
<td>i4</td>
<td>a2</td>
</tr>
<tr>
<td>i5</td>
<td>a1</td>
</tr>
<tr>
<td>i6</td>
<td>a2</td>
</tr>
<tr>
<td>i7</td>
<td>a1</td>
</tr>
<tr>
<td>i8</td>
<td>a2</td>
</tr>
</tbody>
</table>

gene expression patterns, which could not be detected using traditional k-means algorithms.

1.2 Motivation behind IC-Clustering

In this section, we present the major differences between the traditional clustering-classification approach and IC-Clustering method.

Table 2 contains eight samples, which are represented by tuples \( t_1, \ldots, t_8 \). Each tuple contains feature items \( i_1, \ldots, i_8 \), which represent pathological information, and objective attributes \( a_1 \) and \( a_2 \), which represent the gene expression levels. Figure 1A shows the objective vectors \((a_1, a_2)\) of the tuples, which are represented by white circles. The tuples are located at regular intervals on the same square. For example, tuple \( t_2 \) is located at \((1,2)\).

Let us form clusters by dividing the tuples into two groups, \( S \) and \( \bar{S} \), the objective vectors of which maximize some measure. Various values, such as the diameter or the connectivity of a group, can be used as the measure. Here, we use the interclass variance extended to the multiple dimension, which is often used in statistics to evaluate clusters. For simplicity, we designate this as the multiple dimensional version interclass variance in this paper. Let \( c(S) \) denote the centroid of the objective vector of \( S \), i.e., \( \frac{1}{|S|} \sum_{i \in S} \vec{x} / |S| \). The interclass variance is defined using the formula:

\[
|S| d(c(S), c(S \cup \bar{S}))^2 + |\bar{S}| d(c(\bar{S}), c(S \cup \bar{S}))^2,
\]

where \( d \) is the distance between the two vectors used as the arguments. The solutions that maximize interclass variance are indicated by the four dotted straight lines \((S1), (S2), (S3)\) and \((S4)\) in Figure 1B. For example, line \((S1)\) divides the tuples into cluster \( S = \{t_1, t_2, t_3, t_8\} \) and cluster \( \bar{S} = \{t_4, t_5, t_6, t_7\} \). These four segmentations are equally optimal in terms of interclass variance. Although the number of data points in this example was small, addition of numerous tuples to the square can yield many equally best divisions.

To understand why the tuples in each cluster are close, we employ traditional clustering-classification approaches such as conceptual clustering (Michalski and Stepp, 1983; Fisher, 1987; Gennari et al., 1989), which attempts to find classifiers that exactly classify the clusters using, e.g. additional feature items \( i_1, \ldots, i_8 \). In Table 2, ‘1’ denotes the presence of an item in each tuple, while ‘0’ denotes its absence. For instance, tuple \( t_1 \) includes items \( i_1, i_2 \) and \( i_4 \), and tuple \( t_5 \) contains \( i_4 \). Figure 1C shows the best classifier consisting of a single item for each cluster. No single item concludes with clusters divided by any of \((S1)–(S4)\). For example, although the classifier for \((S1) \) would be \( i_2, i_4 \) cannot conclude with cluster \((S1) \). The classifier has only 87.5% accuracy for \((S1) \). This example indicates that the simple process of putting similar tuples of objective vectors together does not produce clustering that is sufficiently informative for each cluster to be associated with special feature items.

To settle this problem, our constrained clustering method computes the optimal clusters, with the restriction that only splits which are expressible by a common itemset are allowed. We call such clusters itemset constrained clusters (IC-Clusters). For example, \{\( t_1, t_2, t_3 \)\} is an IC-Cluster annotated with \( i_4 \), since the split equals the set of tuples that contain \( i_4 \) \((S5)\) in Figure 1D. That is, classifier \( i_4 \) has 100% accuracy for cluster \{\( t_1, t_2, t_3 \)\}. \{\( t_1, t_2, t_7, t_8 \)\} is also an IC-Cluster associated with the special classifier \{\( t_1, t_2 \)\} \((S4)\) in Figure 1D. In these IC-Clusters, the optimal cluster maximizes interclass variance; the set of tuples whose interclass variance is larger is a better cluster. For example, \{\( t_1, t_2, t_7, t_8 \)\} split by \( S4 \) is a better IC-Cluster than \{\( t_1, t_2, t_3 \)\} split by \( S5 \). Moreover, the division by \( S5 \) is likely to be overlooked by traditional clustering algorithms such as two-clustering or k-means.

This example reveals that computing the optimal clusters is a non-trivial problem due to two major considerations. First, the cluster that maximizes the index, such as interclass variance, is not always associated with special features. In our example, although clusters segmented by \((S1), (S2)\) and \((S3)\) have the optimal index, these clusters are not associated with special features. Second, the number of combinations of items increases rapidly with the number of items.

1.3 Related work

A clustering-classification approach that refines clustering or classification may improve accuracy. Clustering studies have paid a great deal of attention to the choice of
Classified clustering with pathological features

Fig. 1. Motivating example of clustering constrained with itemset. (A) observed vectors, (B) optimal clusters, (C) best classifications and (D) best 2 IC-Clusters.

a measure that is tailored to the specificity of the given data. For example, measures of sub-clustering for gene expression profiles (Cheng and Church, 2000), projection-based measures for image analysis (Yang et al., 2002) and model-based measures (Tantrum et al., 2002) have been proposed. However, in these approaches, feature items are not supposed to be used to output directly constrained clusters.

To divide the data, in this paper, we use itemsets as classifiers. However, one might attempt to consider decision trees, such as CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993), to optimize the quality of the data division. But, this extension is computationally expensive, since searching for the optimal itemset is intractable. Therefore, in this paper, we focus on the itemset optimization problem, and present a practical branch-and-bound algorithm.

This paper is organized as follows. In Section 2, we formalize the IC-Clustering problem tailored it to finding variants of diseases. To resolve the abovementioned problem, we introduce the property of interclass variance in Section 3. Then, we present an effective IC-Clustering algorithm in Section 4. In Section 5, we test the IC-Clustering using a real liver cancer dataset, and compare it with the traditional clustering approaches. The results show that the algorithm works efficiently with real datasets. Finally, Section 6 contains our concluding remarks.

2 IC-CLUSTERING

To finding variants of diseases motivated us to consider the following IC-Clustering problem:

IC-Clustering. Suppose that we classify a tuple by checking to see whether it includes a feature itemset (e.g. \{i_1, i_2\}). Compute the optimal constraint that maximizes interclass variance with its corresponding cluster, or list the most significant N solutions.

In the running example, the optimal classifier is the itemset \{i_1, i_2\} and its corresponding cluster is \{t_1, t_2, t_7, t_8\}. Furthermore, when N = 10, the IC-Clustering problem demands the extraction of 10 optimal IC-Clusters.

Unfortunately, it is difficult to compute an optimal itemset that maximizes the interclass variance, because the problem is NP-hard if we treat the maximum number of items in an itemset as a variable. The NP-hardness has been proved by Morishita and Sese (2000) using Proposition 1. An effective method is required to compute the optimal itemset in practice. To compute the IC-Clustering problem, we present the properties of interclass variance in the next section.
3 INTERCLASS VARIANCE

3.1 Basic definitions

In this section, we first introduce the index; interclass variance.

DEFINITION 1. Here, we consider tuples (or records) that are associated with Boolean attributes and numerical attributes. The value of an attribute associated with a tuple \( t \) is denoted by \( t_i[x] \). Boolean attributes are called items, and are especially denoted by \( i_k \). When \( t_i[x] = 1 \) for \( k = 1, \ldots, n \), \( t \) is called to contain the itemset \( \{i_1, i_2, \ldots, i_n\} \). Numerical attributes are called objective attributes. We denote a set of tuples by \( D \).

In Table 2, let \( D = \{t_1, t_2, \ldots, t_k\} \) and \( i_1, i_2, i_3 \) be items, and \( a_1 \) and \( a_2 \) be objective attributes. Then, \( t_1 \) contains itemset \( \{i_1, i_2, i_3\} \), and \( t_4[a_2] = 1 \).

We divide \( D \) into two groups using itemset \( I_i \) and \( I_o \). \( I_i \) is a set of tuples including itemset \( I \), and \( I_o = D - I_i \). In the running example, when \( I = \{i_1\} \), \( I_i = \{t_1, t_2, t_4, t_7, t_8\} \) and \( I_o = \{t_3, t_5, t_6\} \).

DEFINITION 2. Let \( n \) be \( |D| \) (the number of given tuples) and \( x(I) \) be \( |D_I| \) (the number of tuples including itemset \( I \)). Let \( s_I = \sum_{t \in D_I} t_i[x] \) and \( y_I = \sum_{t \in D_I} t_i'[x] \). We define the interclass variance of itemset \( I \) as:

\[
x(I) \sum_{i=1}^{m} \left[ \frac{y_I[i_i] - s_I}{y_I} - \frac{s_I}{n} \right]^2 + \left[ n - x(I) \right] \sum_{i=1}^{m} \left[ \frac{s_I - y_I[i_i]}{n - x(I)} - \frac{s_I}{n} \right]^2.
\]

In the running example, \( n = 8, s_I = s_2 = 16, x(\{i_1\}) = 5, y(\{i_1\}) = 9 \) and \( y(\{i_1\}) = 12 \). Therefore, the interclass variance of itemset \( \{i_1\} \) is

\[
5 \left\{ \left( \frac{9}{5} - \frac{16}{8} \right)^2 + \left( \frac{12}{5} - \frac{16}{8} \right)^2 \right\} + 3 \left\{ \left( \frac{7}{3} - \frac{16}{8} \right)^2 + \left( \frac{4}{3} - \frac{16}{8} \right)^2 \right\} = 2.67.
\]

As \( s_I \) and \( n \) are independent of the choice of itemset \( I \) according to the definition of interclass variance, the values of \( x(I) \) and \( y(I) \) uniquely determine interclass variance. Therefore, we will refer to interclass variance as \( \text{var}(x, y_1, \ldots, y_m) \).

DEFINITION 3.

\[
\text{var}(x, y_1, \ldots, y_m)
\]

\[
= x \sum_{i=1}^{m} \left( \frac{y_I[i_i] - s_I}{x} \right)^2 + (n - x) \sum_{i=1}^{m} \left( \frac{s_I - y_I[i_i]}{n - x} \right)^2.
\]

In the running example, let \( I = \{i_1\} \). We can calculate \( \text{var}(x(I), y(I), y_2(I)) = 2.67 \).

When \( m = 1 \), this measure equals the interclass variance, a well-known statistical measure. Therefore, this index is a multi-dimensional generalization of the interclass variance.

From the definition of interclass variance, we can prove the convexity of \( \text{var}(x, y_1, \ldots, y_m) \). The convexity is useful for conducting an effective search for significant itemsets.

DEFINITION 4. A function \( f(x, y_1, \ldots, y_m) \) is convex if for any \( (x, y_1, \ldots, y_m) \) and \((x', y'_1, \ldots, y'_m)\) in the domain of \( f \), and for any \( 0 \leq \lambda \leq 1 \),

\[
\lambda f(x, y_1, \ldots, y_m) + (1 - \lambda) f(x', y'_1, \ldots, y'_m)
\]

\[
\geq f[\lambda(x, y_1, \ldots, y_m) + (1 - \lambda)(x', y'_1, \ldots, y'_m)].
\]

PROPOSITION 1. \( \text{var}(x, y_1, \ldots, y_m) \) \( (0 \leq x \leq n) \) is a convex function.

PROOF. (Omitted).

3.2 Upper bound

To calculate the set of significant itemsets, it is useful to estimate the upper bound of the interclass variance of itemset \( J \) because this information allows us to restrict the search space of the itemsets. For example, let us compute the interclass variance and the upper bound of \( \{i_1\} \), and then calculate the upper bound of itemset \( \{i_1, i_2\} \). At the time if the upper bound of \( \{i_1, i_2\} \) is less than the interclass variance of \( \{i_1\} \), \( \{i_1, i_2\} \) and its supersets (e.g. \( \{i_1, i_2, i_3\} \)) can be pruned because all the interclass variances of \( \{i_1, i_2\} \) and its supersets are lower than interclass variance of \( \{i_1\} \).

To describe the estimation, we map each itemset \( J \) to a tuple \( [x(J), y_1(J), \ldots, y_m(J)] \), which we call the shape stamp point of \( J \). The stamp point enables us to describe an upper bound estimate of the index of \( J \) with a hyperpolyhedron, which encloses all the stamp points of \( J \) for a given itemset \( I \).

DEFINITION 5. Let \( y_{i,k}(I) \) be the multi-set \( \{\sum_{t \in D_I} a_{ti} \} = |D'| = k, D' \subseteq D_I \). Let \( S_k(I) \) be

\[
\{[k, z_1, \ldots, z_m] | z_i = \max y_{i,k}(I) \}
\]

or

\[
\{[k, z_1, \ldots, z_m] | z_i = \min y_{i,k}(I) \}
\]

where \( m \) is the number of objective attributes.

For example, in Table 2, let \( I = \{i_1\} \). Then, \( y_{1,1}(I) = \{1, 1, 2, 2, 3\}, y_{1,2}(I) = \{2, 3, 3, 3, 4, 4, 4, 5, 5\} \). Furthermore, \( y_{2,2}(I) = \{3, 4, 4, 4, 5, 5, 5, 5, 6, 6\} \). Therefore, \( S_2(I) = \{2, 2, 3\}, \{2, 2, 6\}, \{2, 5, 3\}, \{2, 5, 6\} \).

Each element in \( S_k(I) \) is a vertex of the wrapping hyperpolyhedron on \( x = k \). Figure 2 illustrates the hyperpolyhedron enclosing all the stamp points. Each circle represents the stamp point of a superset of \( \{i_1\} \) on the hyper-plane \( k = 2 \). On this hyper-plane, superset \( J \) of itemset \( \{i_1\} \) contains four itemsets, such that \( x(J) = 2 \). These are \( \{i_1, i_3\}, \{i_1, i_4\}, \{i_1, i_5\} \) and \( \{i_1, i_2, i_4\} \). Each rhombus indicates
the vertex of a (hyper-)rectangle of $S_2(I)$. We can confirm that the rectangle surrounds all four stamp points. Note that if $x(J) \leq x(I)$ for any itemset $J \supseteq I$, we can prove the following lemma.

**Lemma 1.** For any itemset $J \supseteq I$,

$$\text{var}(x(J), y_1(J), \ldots, y_m(J)) \leq \max_{0 \leq k \leq x(I)} \{\text{var}(\tilde{x}) \mid \tilde{x} \in S_k(I)\}.$$

**Proof.** It is known that any convex function is maximized at one of the vertices on the boundary of a convex hyper-polyhedron (Horst and Tuy, 1993). From Proposition 1, interclass variance is a convex function. Owing to its convexity, it is sufficient to prove that the hyper-polyhedron of $\bigcup_{x(I) = k} S_k(I)$ encloses all the stamp points of itemsets $J \supseteq I$.

**Observation 1.** (Morishita and Sese, 2000) Let us evaluate itemsets using an index satisfying convexity. Let $N$ be the user-specified number of IC-Clustering rules. Let $L$ be a list of the best $N$ itemsets, and $\tau(L)$ be the $N$-th best value in $L$. For any itemset $J \supseteq I$, since $u(J) \leq u(I)$, $J$ can be pruned when $u(I) < \tau(L)$.

This observation enables us to design the algorithm ‘IC-Clustering’. To describe the IC-Clustering, we define the following notation.

**Definition 6.**

$$u(I) = \max_{0 \leq k \leq x(I)} \{|\text{var}(\tilde{x})| \mid \tilde{x} \in S_k(I)\}.$$

**4 IC-CLUSTERING ALGORITHM**

The estimation of an upper bound enables the design of an algorithm to solve the IC-Clustering problem as a result of the following pruning observation.

Figure 3 illustrates the wrapping strategy used in Lemma 1. Indeed, according to this lemma, we can estimate an upper bound of the interclass variance of any itemset $J \supseteq I$.
IC-Clustering
1 (Q1, L) := ICC-init;
2 // L: list of the best N rules
3 B1 := Q1; k := 1;
4 repeat until Qk = φ
5 for each B ∈ B1, Q ∈ Qk
6 st. tail(Q) < head(B) and u(Q) ≥ τ(L)
7 // Search only productive Q
8 // τ(L) : Nth best value in L
9 if u(B) < τ(L)
10 // Dispose of unproductive 1-itemsets
11 Remove B from B1;
12 next;
13 end
14 // Construct new candidate itemset and
15 // update Qk+1 and L if necessary
16 (Qk+1, L) := ICC-update(Q ∪ B, Qk+1, L);
17 end
18 k++;
19 Return L; // the best N rules

Fig. 4. The pseudo-code for IC-Clustering.

ICC-init
1 L := φ;
2 for each I ∈ {J | J is a 1-itemset s.t. x(J) ≤ S}
3 // Calculate upper-bound and var for each 1-itemset
4 (Q1, L) := ICC-update(I, Q1, L);
5 end
6 Return Q1 and L;

Fig. 5. The pseudo-code for ICC-init.

ICC-update (Itemset I, Set of Itemsets Q, List of the best N rules L)
1 Calculate u(I) according to Definition 6;
2 // Update Q and L if necessary (Observation 1)
3 if u(I) ≥ τ(L)
4 Put I into Q;
5 if var(x(I), y1(I), ..., ym(I)) ≥ τ(L) and x(I) ≥ C
6 // Update the best list L, except when
7 L := list of the best N rules in L ∪ {I};
8 end
9 end
10 Return Q and L;

Fig. 6. The pseudo-code for ICC-update.

Based on the set enumeration tree (Bayardo, 1998; Sese and Morishita, 2002), which is tailored to compute the best N rules using a statistical measure.

In the pseudocode, we use the user-specified values of C and S to extract informative clusters.

Definition 9. Let C and S be user-specified values. Let us compute the significant itemset I, such that x(I) ≥ C and x(I) ≤ S, for any i ∈ I.

C indicates the minimum cluster size, which has the effect of removing clusters that include few tuples. S allows us to avoid finding constraints that include ubiquitous items, which are contained within almost all tuples.

Example. Let us illustrate IC-Clustering using the data in Table 2. The procedure is illustrated in Figure 7. Let us compute the optimal IC-Cluster in terms of the interclass variance, namely N = 1, where C = 1 and S = 8.

First, we calculate 1-itemsets (the pseudo-code is shown in Fig. 5). According to the total order of items, let I be {i1}. Therefore, var[x(i1)], y1(i1), y2(i1)] and u(i1) are calculated, and L is updated. L = {i1} and τ(L) = 2.67 [(1) in Fig. 7]. We then put {i1} into Q1. Next, we generate {i2}. Although var[x(i2)], y1(i2), y2(i2)] < τ(L), this itemset cannot be pruned because u(i2) ≥ τ(L). Therefore, we add {i2} to Q1. We then estimate the upper bound of the superset of {i1}. Since u(i1) = 2.29 and this value is less than τ(L), Observation 1 (line 4 in Fig. 6) allows us to prune {i1} and its superset without calculating var[x(i3), y1(i3), y2(i3)] [(3) in Fig. 7].

Similarly, we can calculate {i4} and {i5} and update L and Q1. Therefore, Q1 = B1 = {i1}, {i2}, {i4}.

Then, we construct Q2 from Q1 = B1 = {i1}, {i2}, {i4}. We choose Q = {i1} from Q1 and B = {i2} from B1 according to its total order, and generate {i1, i2}. Its interclass variance is 5.00 > τ(L). Therefore, we can update L [(6) in Fig. 7]. Next, we generate {i1, i4}. Since u(i4) < τ(L) = 5.00, it is no longer necessary to construct {i1, i4}. Furthermore, we can remove {i4} from B1 [line 11 in Fig. 4 and (7) in Fig. 7].

Finally, {i1, i2, i7, i8} is classified using {i1, i2} with an optimal IC-Cluster. Figure 8 illustrates the IC-Cluster.

5 EXPERIMENTAL RESULTS

This section presents the results of the experiment examining the efficiency of IC-Clustering using a liver cancer gene expression dataset.

This dataset includes expression levels of 3072 genes over 238 samples obtained by ATAC-PCR (Kato, 1997), which shows similar to those obtained by microarray analyses but with a higher degree of precision. The samples consist of three different types of cells: 120 samples from tumors of liver cancer patients, 86 samples from normal livers and 32 samples from normal livers.

Each sample has 16 pathological feature items shown in Table 1.

5.1 Preprocessing

Gene expression data generally have a lot of missing values, and our ATAC-PCR is not exception to this. Because most of
the missing values locate in several genes and several samples, we eliminate genes containing more than 30% missing values of all the samples. Furthermore, we remove samples containing more than 30% missing values of all the genes. We then fill in the remaining missing values with the average expression level of each gene over all the samples.

After this preprocessing, the dataset holds 1993 genes over 213 samples.

5.2 Results

In this subsection, we summarize our results obtained by application of IC-Clustering to liver cancer analysis.

We compute five IC-Clusters with the largest interclass variance from the dataset. In this test, $C$ and $S$ are 10 and 90% of the number of samples, respectively. In Java on a PC with a 1-GHz Pentium III processor and 512 MB of main memory on Windows XP, this task took 39 s to complete.

Table 3 shows the results of the IC-Clustering. Each row represents an IC-Cluster, and contains the rank of cluster, itemset constraint, cluster size divided by the constraint and interclass variance of the IC-Cluster. For instance, in the first line, the optimal IC-Cluster can be characterized by \{tumor\} containing 106 samples, and the interclass variance of the IC-Cluster is 3126.9.

Table 4 shows the results of the traditional clustering-classification approach. As the clustering, we use $k$-means clusterings ($k = 2, 3, 4$ and 5) with Euclidean distance because this type of clustering has often been used for associating cell features with gene expression patterns. Each row in this table represents a cluster and its optimal classifier expressed by a pathological feature itemset in terms of accuracy for the cluster. For example, first two rows indicate two clusters computed by two-means clustering. Upper row in the two indicates that \{tumor\} is the optimal classifier in terms of accuracy; the classifier has 89.2% accuracy and the cluster contains 108 samples.

The results of IC-Clustering shown in Table 3 indicate that the classifier \{tumor\} forms the most significant cluster in all of the combinations of features in terms of interclass variance. This result suggests that the differences in gene expression patterns between tumor cell and non-tumor cell in cancerous liver are the most significant consideration in any conjunction of pathological features.

A similar result was found using $k$-means algorithm (Table 4), in that the clusters that were computed using two-means are associated with \{tumor\}. In contrast to the clustering-classification approach, IC-Clustering guarantees that the division by \{tumor\} forms the best classifier. These results verify that the significant association found in clustering-classification approach is also found in IC-Clustering.

Next, we discuss other IC-Clusters. None of the rules ranked from 2 to 5 in Table 3 appeared in $k$-means classifiers. Of these rules, a notable one was ranked number 4. The rule means that gene expressions at liver cancer tumor is different between
Table 4. \(k\)-means results

<table>
<thead>
<tr>
<th>No. of clusters ((k))</th>
<th>Cluster no.</th>
<th>Best classifier</th>
<th>Cluster size</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>{tumor}</td>
<td>108</td>
<td>89.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>{except tumor}</td>
<td>105</td>
<td>89.2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>{tumor}</td>
<td>102</td>
<td>76.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>{tumor, not over 65 years old}</td>
<td>62</td>
<td>78.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>{HBV(\sim), not over 65 years old, cirrhosis}</td>
<td>49</td>
<td>72.3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>{tumor}</td>
<td>65</td>
<td>78.4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>{HBV(\sim), not over 65 years old, cirrhosis}</td>
<td>44</td>
<td>82.2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>{tumor}</td>
<td>33</td>
<td>81.7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>{tumor, HBV(\sim), man, normal liver function, no cirrhosis}</td>
<td>11</td>
<td>84.5</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>{tumor}</td>
<td>96</td>
<td>79.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>{tumor, not over 65 years old}</td>
<td>60</td>
<td>78.9</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>{abnormal liver, not over 65 years old, normal liver function}</td>
<td>26</td>
<td>80.3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>{tumor, HCV(\sim)}</td>
<td>25</td>
<td>82.7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>{tumor, HBV(\sim), man, normal liver function, no cirrhosis}</td>
<td>6</td>
<td>85.0</td>
</tr>
</tbody>
</table>

man and woman. This subcluster of tumors means that the treatment of liver cancer may have to be changed according to the sex of the patient. This rule is important in medical treatment, because finding such subclusters provides information that is tailored to each patient on an individual basis. This cluster is not found in \(k\)-means results, because the clustering and classification steps are performed independently of one another.

Let us investigate whether other clusters that are found in \(k\)-means can also be discovered by IC-Clustering. The classifier \{tumor, not over 65 years old\} appears in clusters of three-means and five-means clustering analyses. This classifier has no significant association with gene expression in terms of interclass variance, because it was ranked 17th in the IC-Clustering result.

We have so far made two major observations: one is that the cluster found by \(k\)-means could also be recognized by using IC-Cluster if we compute many IC-Clusters. The other is that traditional approach may disregard the cluster associated with pathological features such as \{tumor, man\}. IC-Clusters have members which have close gene expression levels to each other and are tightly associated with their constraints of cancer feature items. Therefore, IC-Clusters associated with pathological features provide information to remark which combination of pathological features make the clusters of gene expressions for improving patient care.

### 6 CONCLUDING REMARKS

It is desirable to identify variants of diseases that are associated with combinations of pathological features using gene expression profiles. To discover these disease subtypes, we introduced the novel paradigm of IC-Clustering which is a clustering method that allows only splits expressible by a common feature itemset, and computes the optimal itemset that maximizes the interclass variance of objective attributes or lists the most significant \(N\) solutions.

When applied to liver cancer datasets, our IC-Clustering technique uncovered informative clusters of gene expression that were annotated with pathological features. Using the constraints that are associated with particular cluster should facilitate the correct diagnosis and appropriate treatment of disease. In contrast, these clusters are overlooked by the \(k\)-means method.

Currently, we are planning two future studies. The first study will examine the different outcomes when the interclass variance is replaced with other measures, since our proofs used only the convexity of the interclass variance. The second study will seek to confirm the results of IC-Clustering using biological or medical methods.

The resolution of the IC-Clustering problem has applications in various fields, since this problem suggests that both clustering and classification analyses should be reconsidered.

### ACKNOWLEDGEMENT

This research has been partly supported by Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, and Culture Japan (Grant no. 12208003).

### REFERENCES


