

# Synthesizing Knowledge: A Cluster Analysis Approach Using Event Covering

DAVID K. Y. CHIU AND ANDREW K. C. WONG, MEMBER, IEEE

**Abstract**—An event-covering method [1] for synthesizing knowledge gathered from empirical observations is presented. Based on the detection of statistically significant events, knowledge is synthesized through the use of a special clustering algorithm. This algorithm, employing a probabilistic information measure and a subsidiary distance, is capable of clustering ordered and unordered discrete-valued data that are subject to noise perturbation. It consists of two phases: cluster initiation and cluster refinement. During cluster initiation, an analysis of the nearest-neighbor distance distribution is performed to select a criterion for merging samples into clusters. During cluster refinement, the samples are regrouped using the event-covering method, which selects subsets of statistically relevant events. For performance evaluation, we tested the algorithm using both simulated data and a set of radiological data collected from normal subjects and spina bifida patients.

## I. INTRODUCTION

**K**NOWLEDGE acquisition is a difficult, yet important process in the construction of knowledge-based systems. In most of the existing knowledge acquisition schemes, knowledge is either put into the system by experts or acquired through inductive learning or automatic deduction [2], [3]. For example, the acquisition of knowledge by learning from examples has been successfully applied to the concept formation of toy blocks [4]. In this paper we propose an approach whereby knowledge can be synthesized by extracting from empirical observations the statistical or deterministic patterns inherent in the data. The newly developed method is capable of synthesizing a large amount of data into statistical interdependence patterns through an event-covering [1] and data-clustering technique.

Clustering and classification methods have been constantly used in exploring new ways for constructing automated data analysis systems [1], [2], [5]–[15]. These methods can be used to detect patterns for concept acquisition, partitioning, and classification. The rationale is that if new patterns can be detected with high reliability, valuable additional insights on the data may be acquired. We have used an event-covering method to detect statistically significant amino acid groups in cytochrome c biomolecules and have used the information to classify taxonomical patterns in living organisms [1]. We believe that our pro-

posed methodology can play an important role in extending some of the existing knowledge-based systems for decision-support applications.

Among the clustering algorithms, numerical taxonomy methods, such as the construction of the dendrogram [9], cluster data according to a distance or similarity measure. These methods normally apply to data of the continuous type and usually do not perform very well on discrete-valued data due to the lack of a sensitive similarity measure. The conceptual clustering methods [6] group discrete-valued data using a set of predefined criteria. But problems arise when a set of relevant and reliable criteria is not easily available, and when the given set of observed data is very large. Other significant attempts include the decision-directed clustering algorithm [11] and DECA [12], both of which use a probability measure and the subsidiary Hamming distance. These methods are designed to cluster both ordered and unordered discrete-valued data. DECA uses Hamming distance in initiating clusters and a second-order probability estimate to group samples into clusters based on the Bayes' decision rule. It requires no assumed parameters. However, it too has some drawbacks. According to DECA, data that are closer to the mode in the probability-distance (P-D) space are grouped to form an initial cluster. Since the distances from the mode of the samples are all projected onto one axis in the P-D space, it may fail to separate the overlapping clusters in that space.

Knowledge involves the ability to select relevant action or information for a particular goal [16]. Thus, to a certain extent, feature extraction in pattern recognition can be considered as a form of knowledge acquisition. For discrete-valued data, even though relevant features can be identified using techniques in [5], [7], [8], [17], only limited effort has been made in developing techniques capable of discriminating between relevant and irrelevant events of a single variable or a joint variable. Though feature variables are selected for clustering or classification purposes, irrelevant events associated with these variables can still affect the effectiveness of an analysis. In this paper we will use an event-covering method to select useful statistical information (at the event representation level) for cluster analysis on discrete-valued data. By event-covering, we mean covering or selecting statistically significant events that are a subset of outcomes of the variable, disregarding whether or not the variable as a whole is statistically significant. This new approach eliminates the effect of "noise" at the event

Manuscript received April 23, 1985; revised November 1, 1985. This work was supported by the Natural Sciences and Engineering Research Council, Canada, under grant A4716.

The authors are with the PAMI Laboratory, Department of Systems Design Engineering, University of Waterloo, Waterloo, ON., Canada.

IEEE Log Number 8407312.

representation level. And when subsequent pattern analysis is performed, only this subset of relevant events is used.

With the event-covering method, we introduce a new cluster analysis algorithm that utilizes only significant statistical information in the data. It consists of a cluster initiation process based on an analysis of the nearest-neighbor distribution, followed by a regrouping process using the event-covering method. This algorithm can be extended to cluster incomplete multivariate data with mixed discrete and continuous values [18]. The extended work will be reported in a separate paper.

In the next section, we discuss how our approach can be used to acquire knowledge. Since the detection of statistical interdependency is a form of useful knowledge, inherent clusters extracted based on these relationships can be viewed as a synthesis of statistical knowledge. Section III introduces our cluster initiation method based on the distance measure. Section IV presents our cluster regrouping algorithm and the event-covering method. The event-covering method is a statistical method that can detect significant interdependent relationships at the event-representation level. This is used to select relevant events for clustering. In Section V, experiments using simulated data are used to evaluate the method in detail. Then in Section VI, we apply this method to real-life data and show that important medical knowledge can be extracted using our method.

## II. SYNTHESIS OF STATISTICAL KNOWLEDGE

With our approach, knowledge is acquired by means of systematic observations rather than through deductions from a body of theory. However, through observations alone, some relevant knowledge is not immediately apparent since many useful interdependent relationships in the data are unknown. For example, in the diagnosis of a disease by observing the symptoms of a single patient, we may not fully comprehend the significance of some diagnostic features if previous knowledge of the relationships between the symptoms and the disease is unavailable. In this paper we propose a method to detect the interdependence relationships on the event level and use this information as "primitive" knowledge for further analysis.

Since this knowledge of event interdependence usually cannot be acquired by a single observation but can be synthesized through repeated observations, we extract this information based on a given set of observed samples. When the observed events are tabulated as according to their frequency of occurrence, such as in the construction of a contingency table, the frequency will reflect their inherent interdependent relationship. However, some of these associations indicated by the frequency of occurrence are just random associations, and do not necessarily indicate any of their true relationships. Hence, associations for which the frequency of occurrence significantly deviates from the random situations are most important. Using statistical tests on ordered and unordered discrete-valued

data, we can ascertain this independent relationship for further analysis.

When the statistical interdependent relationships are determined, we can gain further knowledge by observing the clustering patterns that take into consideration these acquired interdependent relationships. When the clustering subgroups are formed, the significance of the events within the subgroups can be analyzed. Patterns of the clustering subgroups are thus knowledge-synthesized. Further, the data have a structure described by their event interdependent relationships.

With these in mind, we can now introduce our cluster analysis algorithm and the event-covering method. But first, we shall describe how a cluster can be initiated using analysis on the distance between observations.

## III. CLUSTER INITIATION

### A. Relationship Between Nearest-Neighbor Distance Distribution and Cluster Characteristics

The cluster initiation process involves the analysis of the nearest-neighbor distance distribution on a subset of samples, the selection of which is based on a mean probability criterion. To describe the cluster initiation process, a few definitions and notations must first be introduced.

Let  $X = (X_1, X_2, \dots, X_n)$  be a random  $n$ -tuple of related variables and  $x = (x_1, x_2, \dots, x_n)$  be its realization. Then a sample can be represented as  $x$ . We call  $x$  a pattern  $n$ -tuple since it is analogous to the pattern vector in the Euclidean space.<sup>1</sup> Let  $S$  be an ensemble of observed samples represented as  $n$ -tuples.

**Definition 1:** The nearest-neighbor distance of a sample  $x_i$  with respect to a set of samples  $S$  is defined as

$$D(x_i, S) = \min_{\substack{x_j \in S \\ x_i \neq x_j}} d(x_i, x_j)$$

where  $d(x_i, x_j)$  is a distance measure.

**Definition 2:** Let  $C$  be a set of samples forming a single cluster. We define the maximum within-cluster nearest-neighbor distance as

$$D_c^* = \max_{x_i \in C} D(x_i, C)$$

$D_c^*$  reflects an interesting characteristic of the cluster configuration; that is, the smaller the  $D_c^*$ , the denser the cluster. If the clusters in  $S$  are unknown, we do not know the value of  $D_c^*$ . However, we can estimate  $D_c^*$  with the following analysis (Fig. 1). The estimation will depend on our conception of a cluster, which is as follows.

- If all the clusters  $C_i$  in an ensemble  $S$  have the same degree of denseness, then  $D_c^*$  is the same for all  $C_i$  in  $S$  and also the same as the maximum of all the  $D(x, S)$  values (Cases 1 and 2 in Fig. 1).

<sup>1</sup>For data represented as a string, a tree, or a graph, the  $n$ -tuple representation can also be used if the data can be mapped into a particular ordering scheme [10], [14], [15].

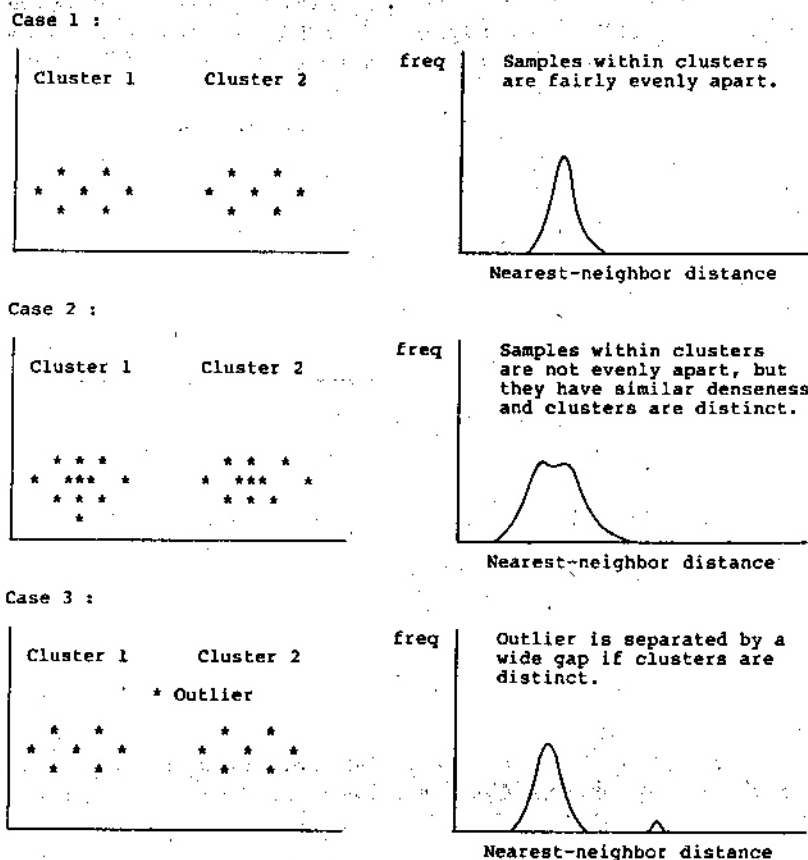


Fig. 1. Relationship between nearest-neighbor distance distribution and cluster characteristics.

If the clusters in  $S$  have different degrees of denseness, then when all  $D(x, S)$  values are projected onto a real axis, distinct groups will result. An isolated sample  $x$ , which does not belong to any cluster (i.e. an "outlier"), will have a relatively large  $D(x, S)$  value. Thus one way to characterize the denseness of all distinct clusters is by the maximum value of  $D(x, S)$  for all  $x$  in  $S$  after the large values associated with isolated samples are removed (Case 3 in Fig. 1). We represent this value as  $D^*$ .

Using a mean probability criterion to select a similar subset of samples, the isolated samples can be easily detected by observing the wide gaps in the nearest-neighbor distance space. The probability distribution from which the criterion is derived for the samples can be estimated using a second-order probability estimation [19], [20]. An estimation of  $P(x)$ , known as the *dependence tree product approximation* [19], can be expressed as

$$\hat{P}(x) = \prod_{j=1}^n P(x_{m_j} | x_{m_{k(j)}}), \quad 0 < k(j) < j$$

where (1) the index set  $\{m_1, m_2, \dots, m_n\}$  is a permutation of the integer set  $\{1, 2, \dots, n\}$ , (2) the ordered pairs  $\{x_{m_j}, x_{m_{k(j)}}\}$  are so chosen that they represent the set of branches of a spanning tree defined on  $X$  with their summed mutual information maximized, and (3)  $P(x_{m_1} | x_{m_0}) = P(x_{m_1})$ . The probability defined above is

known to be the best second-order approximation of the high-order probability distribution [19]. Then corresponding to each  $x$  in the ensemble, a probability  $P(x)$  can be estimated.

In general, it is more likely for samples of relatively high probability to form clusters. By introducing the mean probability as below, samples can be divided into two subsets: those above the mean and those below. Samples above the mean will be considered first for cluster initiation.

**Definition 3:** Let  $S = \{x\}$ . The *mean probability* is defined as

$$\mu_s = \sum_{x \in S} P(x) / |S|$$

where  $|S|$  is the number of samples in  $S$ .

#### B. Cluster Initiation Algorithm

When distance is considered for cluster initiation, we can use the following criteria in assigning a sample  $x$  to a cluster.

- 1) If there exists more than one cluster, say  $\{C_k | k = 1, 2, \dots\}$ , such that  $D(x, C_k) \leq D^*$  for all  $k$ , then all these clusters together with  $x$  can be merged.
- 2) If there exists exactly one cluster  $C_k$ , such that  $D(x, C_k) \leq D^*$ , then  $x$  can be grouped into  $C_k$ .
- 3) If  $D(x, C_k) > D^*$  for all clusters  $C_k$ , then  $x$  may not belong to any cluster.

We use the mean probability to control this merging process at each iteration in the algorithm outlined below:

- 1) Calculate  $P(x)$  for all  $x$  in  $S$ .
- 2) Set  $K := 0$ ;  $t := 0$ .
- 3) Let  $C_0$  be a dummy subgroup representing samples of unknown cluster. Initially  $C_0$  is empty.
- 4) If  $|S| > T$  then  $P' := \mu$ , else  $P' := 0$ . ( $T$  is a size threshold indicating the smallest size of a cluster.<sup>2</sup>)
- 5) List all  $x \in S$  in a table  $L$ , if  $P(x) > P'$ .
- 6) Calculate  $D(x, L)$  for all  $x$  in  $L$ .
- 7)  $D^* := \max_{x \in L} D(x, L)$  and assume that  $x$  is not isolated.<sup>3</sup>
- 8) For all  $x$  in  $L$  do the following.
  - a) Find  $x$  such that  $P(x)$  is the highest.
  - b) If  $D(x, C_k) \leq D^*$  for more than one cluster, say  $C_k$ ,  $i = 1, 2, \dots$ , then do
    - i) if one of the cluster, say  $C_{k_i}$ , is found at a previous iteration, i.e.  $k_i < K$ , then  $C_0 := C_0 \cup \{x\}$ ;
    - ii) else all the clusters,  $C_{k_i}$ ,  $i = 1, 2, \dots$ , are merged.
  - c) If  $D(x, C_k) \leq D^*$  for exactly one cluster  $C_k$ , then  $C_k := \{x\} \cup C_k$ .
  - d) If  $D(x, C_k) > D^*$  for all clusters  $C_k$ ,  $k = 1, 2, \dots, t$ , then  $t := t + 1$  and  $C_t := \{x\}$ .
  - e) Remove  $x$  from  $L$  and  $S$ .
- 9)  $K := t$ .
- 10) Go to 4 until  $S = 0$ .
- 11) For  $k = 1$  to  $t$  do the following.
  - If  $|C_k| < T$ , then  $C_0 := C_0 \cup C_k$ .

#### IV. CLUSTER REFINEMENT

The cluster regrouping process uses a decision rule based on statistical significant events obtained through event-covering method. In this section, we describe how this can be achieved.

##### A. Event-Covering and Covered-Subset

The event-covering method can be conceptualized as a mapping which maps events onto a binary decision state which indicates whether or not they are relevant for clustering. Let  $C = \{a_{c1}, a_{c2}, \dots, a_{cq}\}$  be the set of labels for all possible clusters to which  $x$  can be assigned. Initially,  $C$  is

<sup>2</sup>Since second-order statistics are required in the probability estimation, the minimum sample size for a reliable estimation can be assumed to be

$$T = A \times \max_{j=1, \dots, n} L_j^2,$$

where the constant  $A$  may be taken as three for liberal estimation and  $L_j$  is the number of possible events for variable  $X_j$  in  $X$ . Since the sample size is small for our experiments, we choose a smaller value for  $T$  based on some initial trials of the experiments. We also perform experiment using simulated data to determine the sensitivity of the choice of this value.

<sup>3</sup>To avoid including distance calculation of outlier, we use a simple method which assigns  $D^*$  the maximum value of all nearest-neighbor distances in  $L$  provided there is a sample in  $L$  having a nearest-neighbor distance value of  $D^* - 1$  (with the distance values rounded to the nearest integer value).

the set of cluster labels found after the initiation process. Since each  $x$  in  $S$  is a realization of  $X = (X_1, X_2, \dots, X_n)$  and also associates with a value in  $C$ ,  $C$  can be considered as an additional variable associated with  $X$ . The information of significant events associated with the cluster configuration is obtained by analyzing the frequency of events observed in the ensemble through the use of a contingency table. For  $X_k$  in  $X$ , we can form a contingency table between  $X_k$  and  $C$ . Let  $a_{ks}$  and  $a_{cj}$  be possible outcomes of  $X_k$  and  $C$ , respectively, and let  $\text{obs}(a_{ks})$  and  $\text{obs}(a_{cj})$  be the respective marginal frequencies of their observed occurrences. The expected relative frequency of  $(a_{ks}, a_{cj})$  is expressed as

$$\exp(a_{ks}, a_{cj}) = \frac{\text{obs}(a_{ks}) \times \text{obs}(a_{cj})}{|S|}$$

Let  $\text{obs}(a_{ks}, a_{cj})$  represent the actual observed frequency of  $(a_{ks}, a_{cj})$  in  $S$ . The expression

$$D = \sum_{j=1}^q \frac{(\text{obs}(a_{ks}, a_{cj}) - \exp(a_{ks}, a_{cj}))^2}{\exp(a_{ks}, a_{cj})}$$

summing over the outcomes of  $C$  in the contingency table, possesses an asymptotic chi-square property with  $(q - 1)$  degrees-of-freedom.  $D$  can then be used in a criterion for testing the statistical dependency between  $a_{ks}$  and  $C$  at a presumed significant level as described below. For this purpose, we define a mapping

$$h_k^c(a_{ks}, C) = \begin{cases} 1, & \text{if } D > \chi^2(q - 1) \\ 0, & \text{otherwise} \end{cases}$$

where  $\chi^2(q - 1)$  is the tabulated chi-square value. The subset of selected events of  $X_k$ , which has statistical interdependency with  $C$ , is defined as

$$E_k^c = \{a_{ks} | h_k^c(a_{ks}, C) = 1\}.$$

We call  $E_k^c$  the covered event subset of  $X_k$  with respect to  $C$ . Likewise, the covered event subset  $E_c^k$  of  $C$  with respect to  $X_k$  can be defined.

##### B. Selection of Significant Variables using Event-Covering

After finding the covered event subsets of  $E_k^c$  and  $E_c^k$  between a variable-pair  $(C, X_k)$ , information measures can be used to detect the statistical pattern of these subsets. These information measures are based on an incomplete probability scheme [21] defined over the subset of significant events in the outcome space of the variables. Let  $X_k^c$  and  $C^k$  represent the restricted variables of the covered event subsets  $E_k^c$  and  $E_c^k$  respectively. An interdependence redundancy measure [17] between  $X_k^c$  and  $C^k$  can be defined as

$$R(X_k^c, C^k) = \frac{I(X_k^c, C^k)}{H(X_k^c, C^k)}$$

where  $I(X_k^c, C^k)$  is the expected mutual information and  $H(X_k^c, C^k)$  is the Shannon's entropy defined respectively



on  $X_k^c$  and  $C^k$ . Mathematically, they are expressed as

$$I(X_k^c, C^k) = \sum_{a_{cu} \in E_c^k} \sum_{a_{ks} \in E_k^c} P(a_{cu}, a_{ks}) \log \frac{P(a_{cu}, a_{ks})}{P(a_{cu})P(a_{ks})}$$

and

$$H(X_k^c, C^k) = - \sum_{a_{cu} \in E_c^k} \sum_{a_{ks} \in E_k^c} P(a_{cu}, a_{ks}) \log P(a_{cu}, a_{ks}).$$

The interdependence redundancy measure has a chi-square distribution [17]:

$$I(X_k^c, C^k) \sim \frac{\chi_{df}^2}{2|S|H(X_k^c, C^k)}$$

where  $df$  is the corresponding degree of freedom having the value  $(|E_c^k| - 1)(|E_k^c| - 1)$ . A chi-square test is then used to select interdependent variables in  $X$  at a presumed significant level.

### C. Second-Order Event Selection using Event-Covering

For a data set with low-noise level, analysis based on the marginal probability distribution of the first-order events (events of a single variable) may be adequate. However, for data with higher noise level, the second-order probability distribution, defined on the joint events corresponding to a variable pair, may be needed. We call these joint events of a variable-pair the *second-order events*. The second-order events are of particular importance because 1) reliable probability estimates can be obtained in an ensemble of a reasonable size and 2) random noise which may affect the outcome of one variable is less likely to simultaneously affect the joint outcome of two variables. Thus during the clustering process, it is desirable that only second-order events are included.

When selecting joint events for clustering purposes, those reflecting interdependency usually contain more information. In other words, their observed frequency should deviate significantly from the expected marginal relative frequency derived from its first-order event. Thus the second-order event corresponding to  $(X_k, X_i)$  must be in  $E_k^i \times E_i^k$ , if they contain additional information as compared to the marginal events. Hence, we accept only these second-order events for further testing while the others are disregarded. Since only a subset of second-order events is now involved, the number of events for analysis during the regrouping phase is substantially reduced.

Now, a new variable corresponding to a variable-pair  $(X_k, X_i)$  in  $X$  can be used to associate with the second-order events in the outcome space of  $E_k^i \times E_i^k$ . For samples represented as  $X = (X_1, X_2, \dots, X_n)$ , we can construct a new representation  $X_e = (X_1, X_2, \dots, X_N)$ .  $X_e$  consists of all the variables in  $X$  as well as those representing all the possible combination of the variable-pairs. Thus,  $N$  has the value  $n + n(n - 1)/2$ . We call  $X_e$  the extended tuple of  $X$ . We can then extend the selection of significant events and variables for clustering as described before to  $X_e$ .

### D. Cluster Regrouping Algorithm

Since not all the components in a sample are statistically relevant for clustering purposes, components (first- and second-order events) of a sample  $x$  are chosen based on the subsets of events selected in the event-covering process. The component of a sample is selected if it has significant interdependency with the hypothesized cluster label. Let  $x'(a_{cj}) = \{x'_1, x'_2, \dots, x'_m\}$  ( $m > 0$ ) be the set of selected components of  $x_e$  in estimating the cluster label as  $a_{cj}$ . The event  $x_k$  in the set  $x'(a_{cj})$  is chosen if the following conditions are satisfied.

- 1) The value of  $x_k$  is not a second-order event that is disregarded.
- 2) The value of  $x_k$  is in  $E_k^c$  and  $a_{cj}$  is in  $E_c^k$ .
- 3)  $R(X_k^c, C^k)$  is significant.

The cluster regrouping process uses an information measure to regroup data iteratively. In [1] we have proposed an information measure called *normalized surprisal* (NS) to indicate significant joint information. Using this measure, the information conditioned by an observed event  $x_k$  is weighted according to  $R(X_k^c, C^k)$ , their measure of interdependency with the cluster label variable. Therefore, the higher the interdependency of a conditioning event, the more relevant the event is. NS measures the joint information of a hypothesized value based on the selected set of significant components. It is defined as

$$NS(a_{cj}|x'(a_{cj})) = \frac{I(a_{cj}|x'(a_{cj}))}{m \left( \sum_{k=1}^m R(X_k^c, C^k) \right)}$$

where  $I(a_{cj}|x'(a_{cj}))$  is the summation of the weighted conditional information defined on the incomplete probability distribution scheme [1] as

$$\begin{aligned} I(a_{cj}|x'(a_{cj})) &= \sum_{k=1}^m R(X_k^c, C^k) I(a_{cj}|x_k) \\ &= \sum_{k=1}^m R(X_k^c, C^k) \left( -\log \frac{P(a_{cj}|x_k)}{\sum_{a_{cu} \in E_c^k} P(a_{cu}|x_k)} \right), \\ &\quad \sum_{a_{cu} \in E_c^k} P(a_{cu}|x_k) > 0. \end{aligned}$$

In rendering a meaningful calculation in the above incomplete probability scheme formulation,  $x_k$  is selected if

$$\sum_{a_{cu} \in E_c^k} P(a_{cu}|x_k) > T$$

where  $T \geq 0$  is a size threshold for meaningful estimation. NS can be used in a decision rule as given below in the regrouping process. Let  $C = \{a_{c1}, a_{c2}, \dots, a_{cq}\}$  be the set of possible cluster labels. We assign  $a_{cj}$  to  $x_e$  if

$$NS(a_{cj}|x'(a_{cj})) = \min_{a_{cu} \in C} NS(a_{cu}|x'(a_{cu})).$$

If no component is selected with respect to all hypothesized cluster labels, or if there are more than one label associated with the same minimum NS, then the sample is assigned a dummy label, indicating that the estimated cluster label is still uncertain. Also, zero probability may be encountered in the probability estimation, an unbiased probability estimate based on [22] is adopted.

In the regrouping algorithm, the cluster label for each sample is estimated iteratively until a stable set of label assignments is attained. The cluster regrouping algorithm is outlined as follows.

- 1) Construct  $x_e$  from  $x$  in the ensemble.
- 2) Identify  $\{E_k^e\}$ ,  $\{E_c^e\}$  and compute the finite probability schemes based on the current cluster labels  $C$ .
- 3) Set number\_of\_change = 0.
- 4) For each  $x_e$  in the ensemble do the following.
  - a) If estimation is uncertain, then assign the dummy label  $a_{ej}$ .
  - b) Otherwise assign  $x_e$  to cluster  $a_{ej}$  if
 
$$NS(a_{ej}|x'(a_{ej})) = \min_{a_{cu} \in C} NS(a_{cu}|x'(a_{cu})).$$
  - c) if  $a_{ej} \neq$  previous\_cluster\_label then do the following.
    - i) Set number\_of\_change = number\_of\_change + 1.
    - ii) Update cluster label for  $x_e$ .
- 5) If number\_of\_change > 0 then go to 2; else stop.

## V. EXPERIMENTS USING SIMULATED DATA

For comparing and evaluating the algorithm, four sets of simulated data taken from [12] are used. They are labeled  $E1$ ,  $E2$ ,  $E3$ , and  $E4$  and have the form  $X = (X_1, X_2, \dots, X_6)$ . The possible discrete values taken up by the variables are from the set  $\{A, B, C\}$ . Thus there are nine possible second-order events. To test the algorithm's sensitivity to cluster size, the sample sizes of the subgroups in  $E2$  and  $E4$  are made uneven and all the subgroups in  $E3$  are made specially small. The sample sizes of the subgroups in the four data sets are tabulated in Table I.

Each of the data sets are stochastically generated tuples consisting of three subgroups. The subgroups in  $E1$  and  $E3$  are generated using the same probability distribution scheme; whereas the subgroups in  $E2$  and  $E4$  are generated using different probability distribution scheme. As an illustration, let us look at the data set  $E1$ .  $E1$  is generated according to the following probability distribution scheme:

$$P(X) = P(X_1)P(X_2|X_1)P(X_3|X_1) \\ \cdot P(X_4|X_1)P(X_5|X_1)P(X_6|X_1).$$

When generating the value for the variables in the ensemble, first, 40 A's, 40 B's and 40 C's are generated for  $X_1$ . Then we predetermine a fixed value for the joint probability  $P(X_1, X_j)$ ,  $j = 2, 3, \dots, 6$  and generate the outcome for the rest of the variables.

The experiment is to detect the subgroups and their membership. We use the unnormalized Hamming distance.

TABLE I  
CLUSTER SIZE FOR THE SIMULATED DATA SETS

	Cluster 1	Cluster 2	Cluster 3	Total
$E1$	40	40	40	120
$E2$	90	70	40	200
$E3$	16	16	16	48
$E4$	60	40	20	120

TABLE II  
CLUSTERING RESULT FOR THE SIMULATED DATA SETS<sup>1</sup>

Data sets	Proposed Method		
	Initiation	Regrouping	DECA
$E1$	100/1/19	120/0/0	112/6/0
$E2$	149/3/48	200/0/0	184/16/0
$E3$	35/0/13	48/0/0	46/2/0
$E4$	82/4/34	120/0/0	92/28/0

<sup>1</sup>The entries indicate the number of correct/incorrect/unknown cluster label.

TABLE III  
CLUSTERING RESULT USING DATA SETS OF DIFFERENT SIZE<sup>1</sup>

Data Set Size	Initiation	Regrouping
200	149/3/48	200/0/0
175	130/3/42	175/0/0
150	111/2/37	150/0/0
125	94/1/30	125/0/0
100	78/2/20	100/0/0
75	59/4/12	74/1/0 <sup>2</sup>

<sup>1</sup>The entries indicate the number of correct/incorrect/unknown cluster label.

<sup>2</sup>Cluster regrouping is also applied to small clusters detected at the initiation phase.

in the cluster initiation phase. Thus, the distance has integer value ranging from zero through six. In the regrouping phase, a 95-percent significance level is used in all the chi-square tests. After several iterations, the algorithm terminates with the cluster result.

The clustering results in both phases together with the final result obtained from DECA [12] are tabulated in Table II. It is noted that even in the cluster initiation phase, the cluster members with high correct rate are detected, though some samples still have unknown cluster label. The sizable group of unknown cluster labels may be due to the insensitivity of the Hamming distance. The final result shows a significant improvement over the previous studies [11]–[12] and gives a 100-percent correct rate in all four data sets. The superiority of the information measure using event-covering is thus obvious. The use of probability and distance criterion for clustering discrete-valued data, however, is better than the result obtained using the distance measure alone.

In the above experiment, we choose the value 10 for  $T$  in the cluster initiation algorithm. The choice of this value has two major effects on the execution of the algorithm: 1) the determination of the smallest cluster size (step 11) and 2) the termination of the cluster initiation process (step 4). Although a rule of thumb has been proposed in footnote 2 to select the value of  $T$ , a smaller value can be used to










FEATURES	A Normal	B	C	D
$X_1$ SHAPE				
$X_2$ TRABECULATION	no			
$X_3$ DIVERT SACCULATION	no	divert reflux	sac in bladder wall	outside bladder wall
$X_4$ OUTLET	normal	intermit- tent filling	persist- ent filling	
$X_5$ THICKNESS OF BLADDER WALL	(normal) 1-3 mm	3-6 mm	> 6 mm	
$X_6$ RIGHT REFLUX	no	1-2	3	4
$X_7$ LEFT REFLUX	no	1-2	3	4

Fig. 2. Features observed on the X-ray plates of human bladder diagnosed by a radiologist.

obtain clusters of smaller sizes if the clusters are distinct. To evaluate the performance sensitivity of the algorithm over a range of  $T$  values, we apply the algorithm on all four data sets using different  $T$  values. First, we choose  $T$  to be five. We find that there is no change in the result for all four data sets. Then we choose  $T$  to be 15. We find that for  $E1$  and  $E2$ , which have clusters of size greater than 15, the result remains the same as before. However, for  $E3$  and  $E4$ , which originally have small clusters (less than 15) and the cluster label of some of the samples remains unknown, this criterion will cause the small clusters to be rejected. If we apply the cluster regrouping to small clusters as well, we achieve the same final result as in Table II. In brief, from our experiment we find that the choice of this value is not sensitive in altering the clustering result.

Next, we evaluate the effect of sample size on the statistical stability of the cluster result. Our cluster regrouping consists of the event-covering process and the decision rule which combines the information from the significant events. By the use of the event-covering method and the selection of significant events, better results can be achieved than when the complete set of outcomes is considered without statistical screening. The event-covering method is basically a statistical method and subject to two types of statistical errors: type 1 error, which rejects an event even though it is relevant for clustering and type 2 error, which accepts an event even though it is irrelevant. Needless to say, the more samples we have, the more reliable our method is. Since our method combines the information

from multiple events, the effect of the statistical errors committed can be minimized.

However, an experiment is set up to evaluate the stability of our method using data sets of different size. We choose the data set  $E2$  for this experiment since it has uneven number of samples in the subgroups. In our experiment, we attempt to show that the reduction of sample size, to a certain degree, will not change the results significantly. From the ensemble of the original 200 samples, we obtain smaller subensembles by incrementally taken out 25 samples randomly each time. The resulting five data sets then consist of 175, 150, 125, 100, and 75 samples. The clustering result is tabulated in Table III. It is noted that change in the sample size of the ensemble does not deteriorate the overall performance of the method in both the cluster initiation and the regrouping phase (Table III). The gradual decrease in the samples of unknown cluster label during cluster initiation probably reflects the decrease in sample size, whereas the number of incorrect cluster label samples remains fairly stable. From this result, we have shown statistical stability on the performance of the algorithm.

## VI. EXPERIMENT USING CLINICAL DATA

To demonstrate the feasibility of the algorithm when applied to real-life data with fairly high noise level, a set of clinical data [23] is used. It consists of features diagnosed from X-ray plates of both normal subjects and spina bifida

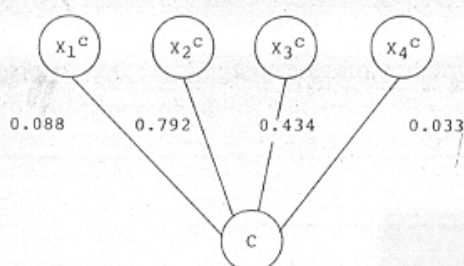


Fig. 3. Interdependence redundancy measures between the restricted variables and cluster label variable. Covered event subsets of the following variables:

$$E_1^c = \{A, C, D\}, \quad E_3^c = \{A, C, D\}$$

$$E_2^c = \{A, C, D\}, \quad E_4^c = \{A, C\}.$$

**X<sub>1</sub> : Shape**

A : highly probable to be normal  
B : not indicative  
C : more likely to be abnormal  
D : more likely to be abnormal

**X<sub>2</sub> : Trabeculation**

A : indicates normal subgroup  
B : not indicative  
C : indicates abnormal subgroup  
D : indicates abnormal subgroup

**X<sub>3</sub> : Divert sacculation**

A : highly probable to be normal  
B : not indicative  
C : indicates abnormal subgroup  
D : highly probable to be abnormal

**X<sub>4</sub> : Outlet**

A : more likely to be normal  
B : not indicative  
C : less likely to be normal

Fig. 4. Significance of the events (of  $X_1$  through  $X_4$ ) in indicating the subgroups. Events which are not in the covered event subsets are not indicative of the subgroups. Events are labelled as according to Fig. 2.

patients by a radiologist. Spina bifida is a family of spinal disorders that occur when a section of spine is left exposed by vertebral abnormalities. Seven features on X-ray plates of the bladder are diagnosed on each sample. They are 1) shape of bladder, 2) presence of trabeculation, 3) divert sacculation, 4) outlet (indicating filling), 5) thickness of bladder wall, 6) presence of right reflux (i.e. reverse urine flow to the right kidney), and 7) presence of left reflux. These features are described for each X-ray plate indicating the different degrees of severity or abnormality (Fig. 2). It happens that all the features selected have four possible different descriptions except for features 3 and 4, which have only three. The data is then represented as  $X = (X_1, X_2, \dots, X_7)$ .

When the clustering algorithm is applied, it is found that the noise level of this data set is fairly high. Since the features are rank-ordered values, Euclidean distance measure is used in the cluster initiation phase. After cluster initiation, interdependence redundancy is calculated between each variable in the extended tuples and the cluster label variable. The tests indicate a low interdependency level and that only the first four variables together with some of their second-order variables are significant. In

TABLE IV  
CLUSTERING RESULT OF THE CLINICAL DATA

Subgroups	Size	Unique characteristics found in the subgroups
Normal subgroup	295	1) No trabeculation.
Abnormal subgroup	139	1) Severe trabeculation. 2) Sacculation in the bladder wall. 3) Divert reflux sacculation with a bell shape bladder.
Unknown subgroup	18	4) Sacculation outside of the bladder wall with a bell shape bladder.
Total	452	

TABLE V  
ABNORMAL SUBGROUPS AND THEIR CHARACTERISTICS

Subgroups	Size	Unique Characteristics Found in the Subgroups
Abnormal subgroup1	116	1) Sacculation in the bladder wall.
Abnormal subgroup2	19	2) Sacculation out of the bladder wall. No divert sacculation.
Unknown subgroup	4	
Total	139	

order to improve the distance measure, we include only the first four variables and apply cluster initiation again. When the cluster regrouping algorithm is applied on the extended tuples of the four variables, a normal subgroup and an abnormal subgroup are found. The result remains the same when the clusters are regrouped based on the extended tuples of the seven variables, since the other variables are not found to be significant. The interdependency patterns are described for the first four significant restricted variables (Fig. 3). When the events are analyzed, a set of unique events (first order or second order) is found in each subgroup. The significance of the events detected by the event-covering and the clustering process is described in more detail in Figs. 4 and Table IV. The characteristic of spina bifida having trabeculate bladder is consistent with what has been reported in [24].

Next, the samples of the subgroups are clustered separately based on the first 4 variables. After cluster initiation, nine abnormal subgroups and eight normal subgroups are identified. The abnormal subgroups have sample sizes ranging from four to 19 and the normal subgroups naturally have a dominant one. When the abnormal subgroups are regrouped based on this small sample result, only variable 3 is found to be significant in estimating any further groupings. The regrouping algorithm merges some of the initial clusters and identifies two abnormal subgroups and one normal subgroup. They are described in Table V. These subgroups identified seem to provide a meaningful grouping and may be useful for medical diagnosis.



## VII. CONCLUSION

It is generally accepted in philosophy that knowledge is a true belief (or a fact, with justification [25]). Our method uses statistical tests as evidence to detect "primitive" knowledge in the form of event association. We then select the relevant events for clustering to reflect the inherent data interdependence relationship. Once the clustering subgroups are found, subsequent analysis on the class pattern can be performed. This synthesized knowledge cannot be obtained by observing individual events in isolation, but it must be based on a set of observations. We have tested the algorithm using simulated data, and both the cluster initiation and the regrouping process show superior results. When we tested the algorithm in a real-life pattern recognition problem using clinical data, we became convinced that the patterns indicated by the clusters and the covered event subsets may be very useful for diagnostic study.

## REFERENCES

- [1] A. K. C. Wong and D. K. Chiu, "A probabilistic inference system," in *Proc. 1984 IEEE Conf. Pattern Recognition*, pp. 303-306.
- [2] P. R. Cohen and E. A. Feigenbaum, *The Handbook of Artificial Intelligence*. New York: William Kaufmann, Inc., vol. 3, 1982.
- [3] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds., *Machine Learning an Artificial Intelligence Approach*. Tioga, 1983.
- [4] T. G. Dietterich and R. S. Michalski, "A comparative review of selected methods for learning from examples," *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell and T. M. Mitchell, Eds. Tioga, 1983.
- [5] D. Wang and A. K. C. Wong, "Classification of discrete data with feature space transformation," *IEEE Trans. on Automatic Contr.*, vol. AC-24, no. 3, 434-437, June 1979.
- [6] R. Michalski and R. E. Stepp, "Automated construction of classifications: conceptual clustering versus numerical taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 4, July 1983.
- [7] A. K. C. Wong and H. C. Shen, "Data base partitioning for data analysis," in *Proc. 1979 Intern. Conf. Cybern. Society*, pp. 514-518.
- [8] H. C. Shen, M. S. Kamel, and A. K. C. Wong, "Intelligent data base management systems," in *Proc. 1983 Intern. Conf. Syst. Man, Cybern.* pp. 1131-1135.
- [9] P. H. R. Sneath and P. R. Sokal, *Numerical Taxonomy*. San Francisco: Freeman, 1973.
- [10] A. K. C. Wong and L. Goldfarb, "Pattern recognition of relational structure," *Pattern Recognition Theory and Applications*, J. Kittler, K. S. Fu, and L. F. Pau, Ed. Dordrecht: D. Reidel, pp. 157-175, 1982.
- [11] A. K. C. Wong and T. S. Liu, "A decision-directed clustering algorithm for discrete data," *IEEE Trans. Comput.*, vol. C-26, pp. 75-82, 1977.
- [12] A. K. C. Wong and D. C. C. Wang, "DECA: a discrete-valued data clustering algorithm," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. PAMI-1, no. 4, pp. 342-349, Oct. 1979.
- [13] M. C. Yovits, G. T. Jacobi and G. D. Goldstein, Eds., *Self-Organizing Systems*. Washington: McGraw-Hill, 1962.
- [14] A. K. C. Wong and M. L. You, "Distance and entropy measure of random graph with application to structural pattern recognition," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. PAMI-7, no. 5, pp. 599-609.
- [15] A. K. C. Wong and T. P. Liu, "Random graph mappings and distribution," Institute of Computer, Science Research Report, University of Waterloo, ON, Canada, 1985.
- [16] A. Newell, "The knowledge level," *Artificial Intelligence*, vol. 18, pp. 87-127, 1982.
- [17] A. K. C. Wong and T. S. Liu, "Typicality, diversity, and feature pattern of an ensemble," *IEEE Trans. Computers*, vol. C-24, no. 2, 158-181, Feb. 1975.
- [18] A. K. C. Wong, D. K. Y. Chiu, and M. de Lascourain, "Inference and cluster analysis of mixed-mode data," unpublished.
- [19] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 462-467, 1968.
- [20] H. H. Ku and S. Kullback, "Approximating discrete probability distributions," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 368-372, 1970.
- [21] S. Guisan, *Information Theory with Applications*. New York: McGraw-Hill, 1977.
- [22] R. Christensen, "Entropy minimax, a non-Bayesian approach to probability estimation from empirical data," in *Proc. IEEE 1973 Int. Conf. Cybern. Society*, pp. 321-325.
- [23] J. Brule, K. Teylonni, and B. R. Smith, "Pattern analysis of unlabelled biomedical data," internal report, Dept. of Systems Design, University of Waterloo, ON, Canada, 1981.
- [24] D. C. Utz and D. M. Barrett, "Stasis involving the lower part of the urinary tract," in *Emmett's Clinical Urography, An Atlas and Textbook of Roentgenologic Diagnosis*, D. M. Witten, G. H. Myers and D. C. Utz (ed.), Vol. 2, 1977, W. B. Saunders Co.
- [25] P. Edwards, Ed., *Encyclopedia of Philosophy*. New York: MacMillan, 1967, pp. 345.