Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data

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

Abstract—The difficulties in analyzing and clustering (synthesizing) multivariate data of the mixed type (discrete and continuous) are largely due to: 1) nonuniform scaling in different coordinates, 2) the lack of order in nominal data, and 3) the lack of a suitable similarity measure. This paper presents a new approach which bypasses these difficulties and can acquire statistical knowledge from incomplete mixed-mode data. The proposed method adopts an event-covering approach which covers a subset of statistically relevant outcomes in the outcome space of variable-pairs. And once the covered event patterns are acquired, subsequent analysis tasks such as probabilistic inference, cluster analysis, and detection of event patterns for each cluster based on the incomplete probability scheme can be performed. There are four phases in our method: 1) the discretization of the continuous components based on a maximum entropy criterion so that the data can be treated as n-tuples of discrete-valued features; 2) the estimation of the missing values using our newly developed inference procedure; 3) the initial formation of clusters by analyzing the nearest-neighbor distance on subsets of selected samples; and 4) the reclassification of the n-tuples into more reliable clusters based on the detected interdependence relationships. For performance evaluation, experiments have been conducted using both simulated and real life data.

Index Terms—Cluster analysis, event-covering, incomplete probability scheme, mixed-mode data, probabilistic inference, statistical knowledge.

I. Introduction

A NEW challenge to computer-based pattern recognition is to detect probabilistic patterns from a database which is usually characterized by heterogenous features of different types, including the mixed discrete and continuous type [1], [2]. This challenge arises from the need in the decision-making process when management control and strategic planning are involved [3]. Such process usually requires unstructured and semistructured decision-making using information from a database. Unlike structured decision-making, which often has well defined objectives and is usually supported by the database schema and query language, unstructured and semistructured decision-making may have to select relevant information that often the decision-makers may not be previously aware of. Hence, extracted knowledge in the form of statistical patterns (based on statistical and cluster analysis) will be very useful in rendering the information [illegible] for this kind of decision-making.

However, while pattern analysis techniques [illegible] cluster analysis on multivariate continuous data [illegible] established [4] and methods to analyze discrete-[illegible] data have been proposed [5]–[7], the problem of [illegible] clustering patterns in multivariate data of the mixed [illegible] remains mostly unsolved. The problem posed by the [illegible]istence of continuous and discrete-valued features [illegible] vious. Methods based on similarity measures are [illegible]ally difficult, if not impossible, to apply to such data [illegible] Alternative methodologies based on probabilistic [illegible]ing [8] require an extremely large amount of data. [illegible] discrete-valued variables are transformed into binary [illegible]ued variables [9], this transformation will drastically [illegible]crease the number of variables in the analysis. [illegible] formation that certain outcomes are from the same [illegible] will be lost in the subsequent analysis. Furthermore, [illegible] the data contain considerable noise which are [illegible] for the analysis, and when the parametric form of [illegible] probability distribution on the data is unknown, the [illegible] becomes even more difficult. Despite these difficulties, [illegible] practical method to detect clustering and statistical [illegible]terns in such data will be very useful and desirable.

What we propose here is a practical approach [illegible] cumvent the difficulties. In our method, a mixed [illegible] probability model is approximated by a discrete [illegible] first discretize the continuous components using a [illegible]mum loss of information criterion. Treating a mixed [illegible] feature n-tuple as a discrete-valued one, we propose a [illegible] statistical approach for synthesis of knowledge [illegible] cluster analysis. The advantage of this method, as we [illegible] see later, is that it requires neither scale normalization [illegible] ordering of the discrete features. Hence, it bypasses [illegible] serious concerns in pattern recognition, namely, 1) [illegible] problem of nonuniform scaling in different feature [illegible]dinates, and 2) the lack of order in nominal data.

By synthesis of the data into statistical knowledge [illegible] refer to the following processes: 1) synthesize and [illegible] from the data inherent patterns which indicate [illegible] interdependency (between certain variables and [illegible] set of their outcomes); 2) group the given data [illegible] ent clusters based on these detected interdependency; [illegible] 3) interpret the underlying patterns for each cluster [illegible]tified. The method of synthesis is based on [illegible] developed event-covering approach [5], [6] [illegible] covering, we mean covering or selecting [illegible]

Manuscript received May 2, 1985; revised June 1, 1987. Recommended for acceptance by J. Kittler. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

A. K. C. Wong is with the Department of Systems Design Engineering, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada.

D. K. Y. Chiu is with the Department of Computing and Information Science, University of Guelph, Guelph, Ont. Canada.

IEEE Log Number 8716844.

atistically interdependent events in the outcome space variable-pairs, disregarding whether or not the variables (considering the complete outcome set) are statistically interdependent. From the detected statistical interdependence patterns of the data, a probabilistic decision le is used to group the data into clusters. Finally, again ing event-covering, we can detect the interdependence atterns between the feature events and the detected ubgroups.

Since the proposed method is based on a general pattern nalysis technique on a set of sample observations, it can applied to a broad spectrum of problem domains where mple self-learning and automatic information selection pability is desirable. Then, it can play an important role extending some of the existing decision-support and owledge-based systems. It can be used to provide new owledge of a problem domain, or to verify important erdependence relationships provided by human experts. Information thus obtained can also be used as additional knowledge to logical information in deductive atabases [10], or to data partitioning in distributed databases [1], [2], [11].

For performance evaluation, the proposed method is pplied to cluster incomplete data (or data with missing utcomes). The method has the following phases:

1) discretization of the continuous components based the maximum entropy criterion;

2) estimation of the missing values in the data set using developed inference method;

3) initiation of clusters by analyzing the distance and nearest-neighbor characteristics of selected samples;

4) reclassification of the samples into more reliable usters based on the statistical interdependence pattern of samples.

II. Mixed-Mode Data and Discretization of the Continuous Components

A. Data Representation and Definitions

Before describing in detail the new approach in handling mixed-mode data by discretizing the continuous mponents, here a few related definitions are introduced. The representation of data is similar to that in the relational model of database where the data are represented tuples.

Definition 1: Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_p, \cdots, x_q, \cdots, x_n)$ be an n-tuple ($1 \leq p \leq q \leq n$) such that the values of $x_1, x_2, \cdots, x_p$ are continuous, $x_{p+1}, x_{p+2}, \cdots, x_q$ are discrete ordinal and $x_{q+1}, x_{q+2}, \cdots, x_n$ are discrete nominal. Then $\boldsymbol{x}$ is called a *mixed-mode* n-tuple and the corresponding random n-tuple is represented as $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ where X_k, $1 \leq k \leq n$, is a continuous or discrete valued variable.

Definition 2: Let the interval $[a, b]$ be the range space of the continuous random variable X_j in $\boldsymbol{X}$. A *partition* on it is defined as a set of L_j intervals $\{[z_0, z_1], [z_1, z_2], \cdots, [z_{Lj-1}, z_{Lj}]\}$, where $z_0 = a$, $z_{Lj} = b$, and, $z_{i-1} < z_i$ for $i = 1, 2, \cdots, L_j$.

Definition 3: In association with the partition, the *boundary set* is defined to be the set of ordered end-points $z_0, z_1, \cdots, z_{Lj}$ which delimit the L_j intervals. $\{a_{jr} | r = 1, 2, \cdots, L_j\}$ then denotes a set of *quanta* such that $z_{r-1} < a_{jr} < z_r$.

Definition 4: A *finite probability scheme* ψ on the partition is defined to be the set of probability values $\{p_i\}$ such that

$$p_i = \int_{z_{i-1}}^{z_i} f(X_j)\, dX_j = F(z_i) - F(z_{i-1})$$

$$\text{for } i = 1, 2, \cdots, L_j$$

where f and F are the probability density function and the cumulative probability function of X_j in $[a, b]$ respectively, and z_{i-1} and z_i are two consecutive elements in the ordered boundary set.

With these definitions in mind, *discretization* is referred to as the process that produces from the range of the continuous random variable X_j the partition of L_j intervals. Thus, there is associated with the intervals a boundary set and a quanta set. From the probability function and the partition, a finite probability scheme is obtained.

B. Maximum Marginal Entropy Discretization and Partitioning

It is clear that the number of ways to discretize the outcome of a continuous variable is infinite. A common procedure is to divide the range into equal length intervals. When the outcomes are not evenly distributed, a large amount of information may be lost after discretization using this method (see Section II-C). To minimize the information loss, we adopt the following partitioning method based on maximum entropy [12].

Formally, let Ψ be the set of all finite probability schemes that can be derived by all the discretization processes for a fixed probability function. The maximum entropy discretization problem is to find a $\psi^* \in \Psi$ such that:

$$H(\psi^*) \geq H(\psi) \qquad \forall \psi \in \Psi$$

where H is the Shannon's entropy function. This method will ensure maximum entropy with minimum loss of information after discretization.

Since high dimension discretization is highly combinatoric, an approximation using marginal entropy is proposed [13]. In practice, we are generally given an ensemble of samples with their probability distribution unknown. The discretization problem thus becomes a partition problem of the observed values for a variable X_j (where some of the outcomes may be repeated). The intervals on the range of a variable X_j are chosen so as to maximize the marginal entropy calculated on the finite probability scheme. Since the algorithm is still combinatoric in nature, to furnish a computationally efficient algorithm, local improvement technique is introduced [14].

When selecting the number of intervals L_j for a continuous variable X_j, it is obvious that in general the more intervals there are the less information will be lost. However, the reliability of the probability estimation based on L_j interval partitioning is affected by the sample size. Hence a rule of thumb [15] is adopted for determining the upper bound of L_j. Since second-order statistics are required in the probability estimation, the sample size for reliable estimation should be greater than A times L_j^2 for all X_j, where A can be taken as 3 for liberal estimation. Subject to this upper bound, the values of L_j in practice will depend on the size of available memory and computational resources.

The partitioning algorithm, using maximum marginal entropy, can be divided into two phases: 1) initial detection of interval boundaries; and 2) improvement on the interval boundaries. The first phase is devised to find intervals such that the sample frequency at each interval is as even as possible. The second phase is introduced to improve the interval boundaries iteratively by increasing the entropy value through local perturbation. It iterates until no improvement can be made. In practice, if the observed continuous data take distinct values, then iteration is not necessary.

Even though the maximum entropy discretization may not produce a unique solution for some data set, the heuristic algorithm [14] we adopted can arbitrarily select a unique set of maximum entropy intervals when more than one set of such intervals exists [16]. The partitioning algorithm can in principle be applied to ordinal-valued variables so as to reduce the number of distinct outcomes in the analysis. However, when the sample size and computational resources are sufficient, there is no need for such application. Despite the algorithm's heuristic nature, it is simple, computationally acceptable, and gives good results.

C. Comparison of the Maximum Entropy and the Equal-Width Discretization Approach

To evaluate the proposed maximum entropy discretization approach on discrete probability distribution estimation, we compare it with that based on equal-width interval discretization. Given an ensemble of sample observations with unknown probability density function, the number of observations falling into each interval is a maximum likelihood estimation of the probability density function [17]. This estimation is known as maximum likelihood estimation irrespective of how the intervals are chosen, given that the number of intervals is fixed. To illustrate the difference between these two approaches, we perform the following experiments.

1) Maximum Entropy Discretization Experiment 1: Consider a variable X and the following values are observed in 30 samples which are sorted in increasing order as:

Sample	1	2	3	4	5	6	7	8	9
X value	0.1	0.9	1.5	2.0	2.8	3.2	3.3	3.5	3.7
Sample	11	12	13	14	15	16	17	18	19
X value	4.0	4.5	4.9	5.5	6.0	7.3	8.5	8.8	9.1
Sample	21	22	23	24	25	26	27	28	29
X value	9.5	9.5	9.7	9.7	10.0	10.3	10.5	11.1	11.8

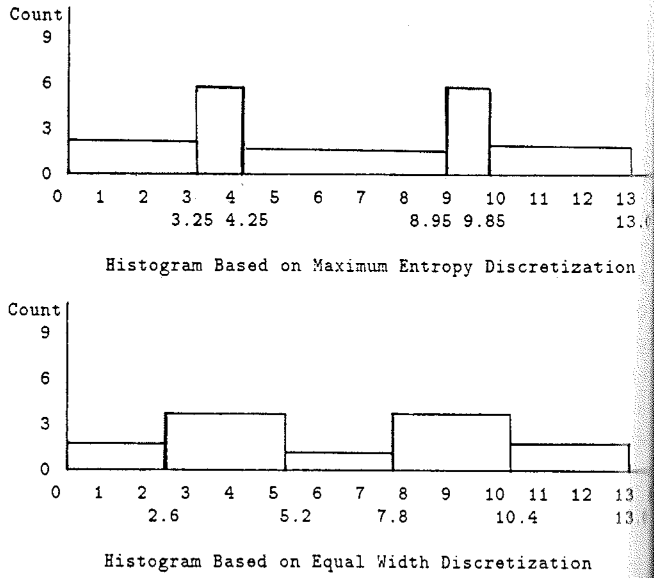

Fig. 1. Comparing histograms based on maximum entropy and equal w discretization.

The probability distribution of X can be estimated fr the histogram constructed based on these values. Let arbitrarily select the number of intervals to be 5 and the range for the outcomes of X be [0.0, 13.0]. The ma imum entropy method then assigns 6 samples to each the five intervals whereas the equal width interval meth assigns the interval width to be 2.6. The histograms probability estimation are plotted in Fig. 1. Compari the two methods, we observe that the maximum entro method is more precise as an estimation than the equ width interval method. It is expected that the precisi would increase with the increase of discretization int vals, given that the sample size is large enough.

2) Maximum Entropy Discretization Experiment 2 supervised classification task based on Bayes' decision used in the second experiment to show the effectiven of maximum entropy discretization for class discrimin tion. Three classes of two-dimensional data of the form $= (x_1, x_2)$ and with different means are stochastic generated. Data from the first class are generated ba on the random combinations of two bivariate normal d tributions, whereas data from the second and third cla are generated based on a single bivariate normal dis bution. The variance matrices are then varied to pro 48 simulated data sets for each of the 7 sets of correla coefficients (Table I). The hold-out method of 10 pe is used to evaluate the classification result. The 7 sets correlation coefficients and the average misclassifica rate are also tabulated in Table I. The result shows the maximum entropy discretization approach is cons ently superior to the equal-width discretization appr

TABLE I
EVALUATE PROBABILITY DENSITY ESTIMATION IN CONTINUOUS-VALUED DATA

Correlation Coefficients				Ave. Misclassification for 48 Simulation Runs	
Class 1		Class 2	Class 3	Max.Entropy	Equal-Width
0	0	0	0	13.33%	13.68%
1/2	1/2	-1/2	-1/2	11.87%	12.84%
-1/2	1/2	0	-1/2	12.43%	12.98%
-0.7	0.7	-0.1	-0.7	12.50%	13.26%
1/4	1/4	0	0	13.33%	13.88%
1/4	1/4	1/4	1/4	13.40%	16.80%
1/4	-1/4	0	0	12.64%	14.72%

III. CLUSTER ANALYSIS ON INCOMPLETE MIXED-MODE DATA

After discretization, we can apply the cluster analysis algorithm on incomplete mixed-mode data. This method does not require the specification of an *a priori* distribution on the data. Furthermore, when certain values in an n-tuple are statistically irrelevant for classification, they will be disregarded by our proposed scheme.

A. *Event-Covering Based on Statistical Interdependency*

First, we present an event-covering method [5], [6] to detect statistically irrelevant outcomes from the mixed-mode data. Using our maximum entropy partitioning method, we obtain a set of intervals on the range of each continuous-valued variable. Then the continuous values observed in the ensemble can be replaced by the corresponding quantum values. Thus we can treat the continuous-valued components as discrete-valued ones. Consequently, the whole ensemble of incomplete mixed-mode data can assume a discrete-valued representation.

The following two procedures are used to estimate the interdependence relationships in the data for the purpose of imputation and cluster analysis. It should be noted that these procedures are applicable to any variable-pair in the mixed-mode n-tuple despite their variable type.

1) Value-to-Variable Interdependency: To estimate an unknown outcome of a particular component in the n-tuple, observed values from the other components can be used. The information for such process can be derived from the statistical interdependency between the observed and the unknown value. Conversely, if an observed value is unrelated to the unknown outcome, it should not be used in the estimation process. To extract this information, the following method is proposed.

In indicating the statistical interdependency between an observed value, say a_{ks}, and the outcome of another variable, say X_j, analysis based on the contingency table is proposed. For a variable-pair (X_k, X_j) in $\mathbf{X}$, a contingency table is constructed based on the discretized values.[1] Let a_{ks} and a_{jt} be the discrete values (or the discretized quantum values in the continuous case) of X_k and X_j, respectively. Let obs (a_{ks}, a_{jt}) represent the observed frequency of the joint outcomes of (a_{ks}, a_{jt}). Furthermore, let exp (a_{ks}, a_{jt}) represent the expected frequency of (a_{ks}, a_{jt}) calculated from the marginal frequencies of X_k and X_j, or estimated based on some expert's judgment. We can estimate exp (a_{ks}, a_{jt}) as

$$\frac{\text{obs}\,(a_{ks}) \times \text{obs}\,(a_{jt})}{M(X_k, X_j)}$$

where $M(X_k, X_j)$ is the size of the sample set such that both the outcome of X_k and X_j are observed; and obs (a_{ks}) and obs (a_{jt}) are the marginal frequencies of $X_k = a_{ks}$ and $X_j = a_{jt}$ in the sample set, respectively. The following expression obtained from the contingency table,

$$D(a_{ks}, X_j) = \sum_{t=1}^{L_j} \frac{\left(\text{obs}\,(a_{ks}, a_{jt}) - \text{exp}\,(a_{ks}, a_{jt})\right)^2}{\text{exp}\,(a_{ks}, a_{jt})},$$

can be used for testing statistical interdependence between a_{ks} and the outcomes of X_j at a given significance level. Notice that $D(a_{ks}, X_j)$ is the summation of L_j terms and each term corresponds to the joint outcome of each distinct value of X_j and a_{ks}. It indicates the deviation of the observed frequency from the expected frequency on this subset of joint outcomes.

The following example shows the selection of certain cells corresponding to a_{ks} in the contingency table of (X_k, X_j). Each cell indicates the observed and expected frequency from an ensemble of totally 400 complete samples. Assuming that there are four distinct outcomes for X_j (denoted as $a_{j1}, a_{j2}, a_{j3}, a_{j4}$), we describe the subcontingency table of $X_k = a_{ks}$ as follows:

	Outcomes of X_j				Marginal frequency
	a_{j1}	a_{j2}	a_{j3}	a_{j4}	
Observed frequency	8	8	40	24	80
Expected frequency	16	32	16	16	
Marginal frequency	80	160	80	80	

The value of $D(a_{ks}, X_j)$ is then calculated from the subcontingency table as: $D(a_{ks}, X_j) = (8 - 16)^2/16 + (8 - 32)^2/32 + (40 - 16)^2/16 + (24 - 16)^2/16 = 62$.

$D(a_{ks}, X_j)$ possesses an asymptotic chi-square property with $(L_j - 1)$ degree of freedom. To select a subset of interdependent events, a function h_k^j which maps the value-variable pair into a binary decision state is defined as:

$$h_k^j(a_{ks}, X_j) = \begin{cases} 1 & \text{if } D(a_{ks}, X_j) > \chi^2_{L_j - 1} \\ 0 & \text{otherwise.} \end{cases}$$

where $\chi^2_{L_j - 1}$ is the tabulated chi-square value. The function indicates whether or not the event is statistically interdependent with the variable based on the significance of the chi-square test. The subset of outcome events of X_k having statistical interdependency with X_j is defined as:

$$E_k^j = \left\{ a_{ks} \,\middle|\, h_k^j(a_{ks}, X_j) = 1 \right\}.$$

E_k^j is called the *covered event subset* of X_k with respect to X_j. The subset E_j^k of variable X_j (with respect to X_k) can

[1] For simplicity, we use the same notation for the variables in the random sample, even though a continuous-valued variable here will assume discretized outcomes rather than continuous ones.

be identified similarly. E_j^k represents the subset of the hypothesized values which are interdependent with the outcomes of X_k. It should be noted that $E_j^k \times E_k^j$ then represents an event subspace of the complete outcome space of the variable-pair selected by this covering process. Statistical information can be analyzed based on the incomplete probability scheme [18] defined on the event subspace spanned by $E_j^k \times E_k^j$ rather than on the complete set of outcomes.

2) Interdependency between Restricted Variables: Based on $E_j^k \times E_k^j$, the interdependency between the two restricted variables defined on $E_j^k \times E_k^j$ can be calculated. Let the restricted variables be represented as X_k^j and X_j^k, defining on E_k^j and E_j^k, respectively. An information measure called *interdependence redundancy* [7] defined on the incomplete probability schemes of the subsets is calculated as:

$$R(X_k^j, X_j^k) = I(X_k^j, X_j^k)/H(X_k^j, X_j^k)$$

where $I(X_k^j, X_j^k)$ and $H(X_k^j, X_j^k)$ are the expected mutual information and the Shannon's entropy defined on X_k^j and X_j^k, respectively. The value of $R(X_k^j, X_j^k)$ will indicate the degree of statistical interdependency between the two restricted variables. We have chosen the interdependency redundancy measure because it is normalized and bounded by 0 and 1. Note that if either $|E_k^j| = 1$ or $|E_j^k| = 1$ then $R(X_k^j, X_j^k) = 0$ since there is only redundancy information for each of the variables rather than interdependency information between the variables. If the redundancy information is also desirable in this situation, we can adopt a two-phased approach [5] which makes inferences based on the interdependency information in the first phase (our proposed method) and then when a rejection occur, make inferences based on the redundancy information. $R(X_k^j, X_j^k)$ has an asymptotic chi-square property [7]:

$$2\, R(X_k^j, X_j^k)\, M\,(X_k^j,\ X_j^k)\, H(X_k^j, X_j^k) \sim \chi^2_{df}$$

where df is the corresponding degree of freedom having the value $(|E_k^j - 1)(|E_j^k - 1)$ and $M(X_k^j, X_j^k)$ is the number of observations in the incomplete scheme of (X_k^j, X_j^k). The chi-square test is then used to detect statistical interdependency between the two restricted variables at a given significance level.

B. Probabilistic Imputation of Missing Values in Mixed-mode Data

Before performing cluster analysis, the missing values in the data set are estimated from the other observed values which are selected based on the detected statistical interdependency. Since a missing value can occur in any of the variables in the tuple, statistical interdependency is calculated between all the variable-pairs. Using the two statistical tests described in the previous sections, only values which are statistically significant for the estimation process are selected. Let the unknown value in an n-tuple $\boldsymbol{x}$ be x_j and its hypothesized value be a_{jr}. The conditions for a value $x_k = a_{ks}$ $(k \neq j)$ in $\boldsymbol{x}$ to be selected for estimation are:

1) the value x_k is an observed value;
2) $R(X_k^j, X_j^k)$ is statistically significant;
3) $a_{ks} \in E_k^j\ a_{jr} \in E_j^k$.

An information measure called the *normalized surprisal* (NS) is used in the decision rule for estimating the missing values. NS corresponds to the weighted information of a hypothesized value a_{jr}, and is conditioned on the selected values [denoted here as $\boldsymbol{x}'(a_{jr})$]. $\boldsymbol{x}'(a_{jr}) = \{x_1', x_2', \cdots, x_m'\}$ represents a sub-n-tuple of $\boldsymbol{x}$ where m $(m < n)$ is the number of values selected. $\mathrm{NS}(x_j = a_{jr}|\boldsymbol{x}'(a_{jr}))$ is defined as follows:

$$\mathrm{NS}(x_j = a_{jr}|\boldsymbol{x}'(a_{jr})) = \frac{I(x_j = a_{jr}|\boldsymbol{x}'(a_{jr}))}{m\left(\sum_{k=1}^{m} R(X_k^j, X_j^k)\right)}$$

where

1) $I(x_j = a_{jr}|\boldsymbol{x}'(a_{jr})) = \Sigma_{k=1}^m \{R(X_k^j, X_j^k)\, I(a_{jr}|x_k')\}$
2) $I(a_{jr}|x_k')$ is the conditional information defined on the incomplete probability scheme on $E_k^j \times E_j^k$ where

$$I(a_{jr}|x_k') = -\log \frac{P(a_{jr}|x_k')}{\sum_{a_{ju}\in E_j^k} P(a_{ju}|x_k')}$$

and $\Sigma_{a_{ju}\in E_j^k} P(a_{ju}|x_k') > T$ such that T is chosen as a size threshold for reliable probability estimation [15].[2]

NS is normalized by the total weights and the number of selected events after weighting each conditional information by $R(X_k^j, X_j^k)$, its measure of interdependence redundancy. In [5], we have discussed more thoroughly the intuitive properties of NS which are as follows:

1) larger the weights, more reliable the estimation;
2) larger the conditional information, more reliable the estimation.

In rendering a meaningful calculation, x_k is selected only if a reasonable sample size is available, or:

$$\sum_{a_{ju}\in E_j^k} P(a_{ju}|x_k') > T.$$

The following decision rule based on the information measure NS is designed. Given $T_j = \{a_{jr}|r = 1, \cdots, L_j\}$ as the set of all possible values that can be assigned to an unknown x_j, then

$$x_j = a_{jt} \quad \text{if} \quad \mathrm{NS}(x_j = a_{jt}|\boldsymbol{x}'(a_{jt})) = \min_{a_{jr}\in T_j} \mathrm{NS}(x_j = a_{jr}|\boldsymbol{x}'(a_{jr})).$$

[2]Since the second-order statistics are required in the probability estimation, the minimum sample size for a reliable estimation is recommended to be:

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

where the constant A may be taken as 3 for liberal estimation and L_j is the number of possible events for variable X_j in $\boldsymbol{X}$. A size threshold [illegible] used in the cluster initiation phase, however, we do not find the [illegible] the value T to be sensitive in affecting the result in our experiments. [illegible] the sample size or the cluster size is small, T can be chosen [illegible] based on some initial trial of the experiments, and small clusters [illegible] be detected while large clusters are not affected.

f x' is an empty set for all hypothesized values or if there re more than one hypothesized value which yields the inimum NS values, then the estimation cannot be made, nd the estimated value is still unknown. These samples which are incomplete because of unknown estimation are aken out initially for cluster initiation.

The computational complexity of the inference method s relatively low. The number of chi-square test applications is $(L_k + L_j + 1)$ for a variable-pair (X_k, X_j) where L_k and L_j represent the number of distinct events for X_k nd X_j, respectively. The tests determine the statistical ignificance of the events for X_k and X_j with respect to heir interdependency. For data represented as an n-tuple, here are $nC_2(=n(n-1)/2)$ different variable-pairs, and he total number of statistical test applications is

$$\sum_{j=1}^{n} \sum_{k=1, k\neq j}^{n} (L_k + L_j + 1)$$

r

$$O\left[n^2(\max_k L_k)\right], \qquad k = 1, 2, \cdots, n.$$

ncluding the calculation of the probability estimates, the omplexity of the event-covering process is

$$O\left[Mn^2(\max_k L_k)\right], \qquad k = 1, 2, \cdots, n$$

here M is the number of samples for probability estimation. The NS calculation is also linearly proportional the number of selected events in the estimation.

Unbiased Probability Estimator

When estimating the probability based on an ensemble samples, zero probability may be encountered if the robability estimation is based on direct frequency count. order to have a better probability estimate for these ses, an unbiased probability estimate proposed by [19], 0] is adopted.

Consider a pair of restricted variables (X_k^j, X_j^k) with the omplete probability scheme involving events in E_k^j and the unbiased marginal distribution of X_j^k is defined as

$$P(x_j^k = a_{jr}) = \{M(a_{jr}) + |E_j^k|\}/\{M(X_k^j, X_j^k) + |E_j^k|^2\}$$

here $M(a_{jr})$ and $M(X_k^j, X_j^k)$ are respectively the frequency of occurrence of a_{jr} and the sample size for the omplete scheme of X_j^k. Similarly, the unbiased joint ribution of X_j^k and X_k^j is defined as

$$P(x_j^k = a_{jr}, x_k^j = a_{ks}) = \{M(a_{jr}, a_{ks}) + 1\}/\{M(X_k^j, X_j^k) + |E_j^k| \times |E_k^j|\}$$

here $M(a_{jr}, a_{ks})$ is the number of occurrence of the joint ome (a_{jr}, a_{ks}) in the incomplete scheme of the ensem- Hence the conditional information $I(a_{jr} \mid a_{ks})$ is calculated as

$$I(a_{jr} \mid a_{ks}) = -\log \frac{P(a_{jr}, a_{ks})/P(a_{ks})}{\sum_{a_{jt}\in E_j^k} P(a_{jt}, a_{ks})/P(a_{ks})} = -\log \frac{M(a_{jr}, a_{ks}) + 1}{\sum_{a_{jt}\in E_j^k} \{M(a_{jt}, a_{ks}) + 1\}}.$$

D. Cluster Analysis on the Data

After the missing values are imputed, cluster analysis can be performed. First, clusters are initiated based on the nearest-neighbor characteristics of the ensemble. Then clusters are regrouped based on the statistical interdependency detected from the data.

1) Cluster Initiation: The cluster initiation process involves three phases: 1) selecting samples which are not yet clustered and are more likely to form clusters first; 2) finding a data-dependent nearest-neighbor criterion which reflects the cluster characteristic; and 3) merging reliable samples to form clusters based on this criterion. These three phases of the process are applied iteratively until all the samples are considered.

The first phase of cluster initiation estimates the probability for a sample to occur and then selects a subset of sample with higher probability estimation. The probability of a sample is estimated by a second-order product approximation on the discretized data [21].[3] Further, the process involves the calculation of a mean probability [6].[4] With the probability estimates of each sample calculated and the mean probability on a given set of samples defined, a subensemble of the unclustered samples with relatively higher probability estimate is selected by the following criterion. A sample is selected if its probability is greater than the mean probability of the remaining unclustered samples. We represent these selected samples as S'.

The second phase involves the calculation of nearest-

[3]An estimation of $P(x)$, known as the dependence tree product approximation [21] can be expressed as:

$$P(x) = \prod_{j=1}^{n} P(x_{m_j} \mid x_{m_{k(j)}}), \qquad 0 \le k(j) < j$$

where 1) the indexes $\{m_1, m_2, \cdots, m_n\}$ are a permutation of the integer set $\{1, 2, \cdots, n\}$ and k is a function of j, 2) the ordered pairs $(x_{m_j}, x_{m_{k(j)}})$ are identified from the branches of a spanning tree (defined on X) where the branch weights are the expected mutual information between the variable nodes; and the ordered pairs are chosen such that the summed expected mutual information of all the branches is maximized, and 3) $P(x_{m_1} \mid x_{m_0}) = P(x_{m_1})$. The probability defined above is known to be the best second-order approximation of the high-order probability distribution [21].

[4]Let a set of selected samples be denoted as S. The *mean probability* for S is defined as

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

where $|S|$ is the sample size.

neighbor distance[5] for all the samples in S'. Let $D(x, S')$ be the nearest-neighbor distance value of x considering all the samples in S'. Among these distances, let D^* be the maximum value.[6] Using the clustering procedure reported in [6], samples can be merged into a cluster basing on the analysis of the nearest-neighbor distance. The cluster initiation is outlined as follows:

1) Calculate $P(x)$ for all x in the ensemble.
2) Set $K := 0$; $t := 0$;
3) Let C_0 be a dummy subgroup representing samples of unknown cluster. Initially C_0 is empty. Initialize the first cluster C_1 containing the sample x such that $P(x)$ is highest.
4) If the number of unclustered samples $> T$ then P' is assigned the mean probability of unclustered samples else P' is assigned 0; (T is a size threshold indicating the smallest size of a cluster.)
5) List all the unclustered samples in a table S' if $P(x) > P'$;
6) Calculate $D(x, S')$ for all x in S'.
7) $D^* := \max_{x \in S'} D(x, S')$ (see footnote 6).
8) For all x in S' do:
   - Get x in S' such that $P(x)$ is highest.
   - If $D(x, C_{k_i}) \leq D^*$ for more than 1 cluster (say C_{k_i} for $i = 1, 2, \cdots, t, t > 1$), then:
     —If $k_i < K$ for some i then $C_0 := C_0 \cup \{x\}$;
     —else $C_{k_1} := \{x\} \cup C_{k_i}$ for all i;
   - If $D(x, C_k) \leq D^*$ for exactly 1 cluster C_k, then $C_k := \{x\} \cup C_k$;
   - If $D(x, C_k) > D^*$ for all clusters C_k, $k = 1, 2, \cdots, t$, then: $t := t + 1$; $C_t := \{x\}$;
   - Remove x from S':
9) $K := t$;
10) Goto 4 until all samples are considered;
11) If $|C_k| < T$, the size threshold, then $C_0 := C_0 \cup C_k$ for all k.

Computationally, the proposed cluster initiation procedure is reasonably fast. It requires the calculation of nearest-neighbor distance between sample-pairs in a subensemble only. The probability estimate is calculated only once for each sample. Also it can apply to any distance measure and it allows uncertain samples to be temporarily assigned as belonging to unknown cluster.

[5]We use the Hamming distance on the complete discretized n-tuples. Let x and y be two n-tuples; then the Hamming distance, $d(x, y)$, is defined as

$$d(x, y) = \sum_{k=1}^{n} \delta_k$$

where

$$\delta_k = \begin{cases} 0 & x_k = y_k \\ 1 & \text{otherwise.} \end{cases}$$

[6]Since outlier has a large nearest-neighbor distance and will affect the value of D^* which is the maximum of such distances, we use a heuristic method to choose D^* as the maximum value of all nearest-neighbor distance in S' provided there is a sample in S' having a nearest-neighbor distance value of $D^* - 1$ (with the distance values rounded to the nearest integer value). In another word, this method screens out the outliers in affecting the value of D^*.

2) *Cluster Regrouping:* After finding the initial clusters along with their membership, the cluster membership (or the cluster label) of each sample x can be considered as an additional value of x. Let the cluster label variable be C and its current set of detected outcomes be $\{c_1, c_2, \cdots, c_g\}$ where g is the current number of clusters detected. The regrouping process is thus essentially an inference process for estimating the cluster label of a sample. During this process, the values which are statistically interdependent with the cluster label (now treated as a variable) are selected. Joint outcomes (second or higher order outcomes) which are found to be interactive in a sample x can be considered as additional observed features if computational resources and storage space are available [6]. Then the decision rule based on the minimum NS value (see Section III-B) can be applied to estimate the cluster label of a sample. The process of estimation iterates until stable clusters are found. The cluster regrouping algorithm is outlined as follows:

1) Compute the finite probability schemes based on the current cluster labels.
2) Identify the events in the covered event subspace for all variables with respect to the cluster label variable;
3) Set number_of_change := 0;
4) For each x in the ensemble do;
   - If estimation is uncertain because more than one cluster label satisfies the minimum criterion or because no value in the n-tuple has been selected, then assign the label as missing.
   - Otherwise assign x to cluster label c_j if:

$$\mathrm{NS}(c_j | x'(c_j)) = \min_{c_u \in C} \mathrm{NS}(c_u | x'(c_u));$$

   - if $c_j \neq$ previous_cluster_label then:
     —number_of_change := number_of_change + 1;
     —update cluster label for x;
5) If number_of_change > 0 then goto 1 else stop.

Because there is no distance measure defined for mixed-mode data, the cluster analysis is performed based on the statistical properties rather than distance measure in the final phase of the algorithm with all the variables treated as nominal variables, including the ordinal variables. However, since the cluster initiation is based on the nearest-neighbor characteristics, the final clusters consist of both distance and statistical information of the data ensemble.

When the clusters are found, interdependency between the class and the event values is a piece of synthesized knowledge which is extracted from the ensemble of data as a whole, and could not be acquired from individual sample in isolation.

IV. Experiment Using Simulated Data

In evaluating this approach to mixed-mode data analysis, an experiment using simulated data is performed. To generate a set of simulated data, four clusters are created based on four n-tuples (Fig. 2). The data are represented as $x = (x_1, x_2, \cdots, x_7)$. These n-tuples are repeated

Class	X_1	X_2	X_3	X_4	X_5	X_6	X_7	frequency
	Contin.			Ord.		Nominal		
1	(1,	6,	3,	6,	1,	F,	A)	200
2	(6,	3,	1,	6,	6,	C,	C)	150
3	(3,	1,	6,	1,	3,	A,	F)	75
4	(6,	3,	6,	3,	1,	A,	C)	75
Total								500

Fig. 2. Original n-tuples for generating the simulated data involving four classes.

TABLE II
RESULT OF EXPERIMENT IN ESTIMATING MISSING VALUES

Variables Type	X_1 Continous	X_2	X_3	X_4 Ordinal	X_5	X_6 Nominal	X_7	Total
Incorrect	19	10	9	9	5	1	2	55
Reject	0	0	0	8	1	0	0	9
Correct	37	46	36	31	36	49	51	286
Total	56	56	45	48	42	50	53	350

number of times to produce an ensemble of 500 n-tuples. Note that the clusters are not determined by a single value but by the joint information of the n-tuple. To create mixed-mode data with noise perturbation, 40 percent of the values are randomly replaced by a value with equal probability from the set $\{B, D, E\}$ for nominal variables, and $\{2, 4, 5\}$ for continuous and ordinal variables. These replaced values have no information about the cluster. Then noise with normal distribution of zero mean and 0.5 standard deviation are imposed on all the values in the ensemble. The variables X_1, X_2, X_3 are designed as continuous, X_4, X_5 as ordinal discrete, and X_6, X_7 as nominal discrete. The generated values are added to continuous and ordinal discrete values. Thus, x_1, x_2, x_3 takes up the real value after the addition. For x_4, x_5, the value is rounded to the nearest integer value bounded by 1 and 6. For x_6 and x_7, if the Gaussian noise value generated is greater than 1, then the resulting value is randomly changed to any arbitrary possible outcome with equal probability. To create a set of incomplete n-tuples, 10 percent of all values is randomly taken out, so that there is a total of 350 missing values.

The purpose of the experiment is to cluster this set of incomplete n-tuples. First, the maximum entropy discretization on the continuous values is applied. Each continuous value will be represented by one of six discrete quantum values indicating six intervals (i.e., $L_j = 6$ for $j = 1, 2, 3$). Then the inference method is applied to estimate any missing value and perform cluster analysis on the data. For continuous variables, the original value is compared to see if it falls in the range of the estimated interval. The 95 percent confidence level is used in all the chi-square tests.

The result of the experiment in estimating the missing values is tabulated (Table II). The number of errors on the different types of variables is probably proportional to the amount of noise imposed. The error rate of the initial

TABLE III
RESULT OBTAINED IN CLUSTERING SIMULATED DATA (INITIAL CLUSTERS)

Class	Misclass.	Unknown	Correct	Total
1	1	26	173	200
2	3	28	119	150
3	1	15	59	75
4	2	48	48	75
Total	7	117	376	500

Note : There are 8 incomplete n-tuples in class 1 and 1 incomplete n-tuple in class 2 among the n-tuples of the unknown class.

TABLE IV
RESULT OBTAINED IN CLUSTERING SIMULATED DATA (FINAL CLUSTERS)

Class	Misclass.	Unknown	Correct	Total
1	2	0	198	200
2	3	0	147	150
3	4	0	71	75
4	10	0	65	75
Total	19	0	481	500

clustering result is very low even though the unknown rate is high (Table III). The unknown rate is the highest in class 4 because, compared to the other classes, the original n-tuple that represents it is the most similar to the others. The final result is given in Table IV. The overall result indicates that the method achieves high reliability for this set of data.

V. EXPERIMENT USING HYDROMETRIC NETWORK DATA

The next experiment involves hydrometric network data. In order to integrate the hydrologic, meteorologic, and physiographic aspect of hydrometric network in a quantitative analysis, a set of samples are collected over 131 catchment areas in British Columbia, Canada [22], for cluster analysis. Seven hydrometric features are chosen for each catchment area 1) mean annual runoff; 2) mean annual precipitation; 3) mean runoff coefficient; 4) relief and bathymetry; 5) ground water activities; 6) moisture index; and 7) forest coverage. They are expressed as $\boldsymbol{x} = (x_1, x_2, \cdots, x_7)$ (Fig. 3), where the first three features are of the continuous type and the others are of the discrete type (nominal as well as ordinal). Since the data are complete n-tuples, the discretization process on the continuous features can be applied immediately, and then the event-covering and cluster analysis are performed. The continuous variables are discretized into four intervals ($L_j = 4$ for $j = 1, 2, 3$). All the statistical tests are based on a confidence level of 95 percent. After cluster regrouping, the final result is shown in Fig. 4.

When examining the two clusters found, feature values characterizing the clusters are noted. Generally speaking, samples of cluster 1 are flat river basins such as flatland and plateau. They have relatively low annual runoff and low precipitation and with noticeable underground water

Var.	Data type	Possible values
X_1	continuous	3.0 to 65.0
X_2	continuous	11.0 to 71.0
X_3	continuous	0 to 1.0
X_4	nominal	{mountain, flatland, valley or plateau}
X_5	nominal	{underground water, underground water in downstream, no underground water}
X_6	ordinal	{sub-arid, semi-arid, sub-humid, humid}
X_7	ordinal	{poorly-covered, half-covered, fully-covered}

Fig. 3. A description of hydrometric network data.

Cluster 1: Size 50

General River Basin Characteristics
relatively low annual runoff
relatively low annual precipitation
relatively high runoff coefficient
majority of flat topography
noticeable underground water activities
low to medium moisture content
less forest coverage

Unique Feature Values
Mean annual runoff below 13.3
Mean annual precipation below 18.5
Mean runoff coefficient above 0.675
flatland and plateau
Underground water activity

Cluster 2: Size 81

General River Basin Characteristics
Relatively high annual runoff
Relatively high annual precipitation
Relatively low runoff coefficient
Mostly mountainous topography
Relatively scarcity of underground water activities
Relatively high moisture content
More forest coverage

Unique Feature Value :
none

Fig. 4. A description of clusters on the hydrometric network data.

Restricted Variables	$R(X_k{}^c, C)$	$E_k{}^c$
$X_1{}^c$	0.200	{3.00-6.3, 6.4-13.3, 13.4-24.9, 25.0-65.0}
$X_2{}^c$	0.138	{<18.5, 18.6-23.7, 23.8-36.1, > 36.2}
$X_3{}^c$	0.175	{< 0.29, 0.30-0.51, 0.52-0.67, > 0.68}
$X_4{}^c$	0.169	{mountain, flatland, plateau}
$X_5{}^c$	0.567	{underg. water, no underg. water}
$X_6{}^c$	0.160	{semi-arid, sub-humid, humid}
$X_7{}^c$	0.071	{poorly covered, half-covered }

Fig. 5. Measure of interdependent redundancy between cluster and the restricted variables.

activities, whereas cluster 2 are river basins having relatively high annual runoff and precipitation. They are mostly mountainous with relatively low underground water activities. The measures of interdependence redundancy between the restricted variables and the cluster label variable are described in Fig. 5. They indicate that the

Var.	events	statistical significance
X_1	< 13.30	indicate cluster 1
	> 13.30	more likely cluster 2
X_2	< 18.50	indicate cluster 1
	18.50-23.70	more likely cluster 1
	23.70-36.10	more likely cluster 2
	> 36.10	highly probable cluster 2
X_3	< 0.295	highly probable cluster 2
	0.295-0.515	not indicative
	0.515-0.675	highly probable cluster 1
	> 0.675	indicates cluster 1
X_4	mountain	more likely cluster 2
	flatland	indicates cluster 1
	valley	not indicative
	plateau	indicates cluster 1
X_5	underg.water	indicates cluster 1
	underg.water in downstream	not indicative
	no underground water	highly probable cluster 2
X_6	sub-arid	not indicative *
	semi-arid	highly probable cluster 1
	sub-humid	highly probable cluster 1
	humid	highly probable cluster 2
X_7	poorly-covered	highly probable cluster 1
	half-covered	more likely cluster 2
	fully-covered	not indicative *

* may be due to small sample size

Fig. 6. The significance of the events in indicating the subgroups.

ground water activities are the most important factor in determining the subgroups and the forest coverage is the least important. Fig. 6 shows the significance of the different events in indicating the subgroups.

VI. Conclusion

In order to acquire more information in tackling complicated tasks involving high-level skills, there is an increasing need to analyze complex multivariate data with variables from different sources and of different forms of description. This paper has proposed a feasible solution to detect clustering patterns in mixed-mode data in an integrated way. The method is mathematically and intuitively meaningful [13], [16]. Furthermore, it possesses algorithmic simplicity. When a reasonably large set of observations is analyzed by a general inference and cluster analysis algorithm using the event-covering approach, new knowledge is acquired indicating different forms of interdependent patterns: subset of interdependent events, interdependent patterns between the restricted variables involving only these events, and clustering patterns based on these acquired interdependence relationships. Once the clusters are formed, further class-value interdependent patterns can be extracted. Information thus obtained reflect synthesized knowledge inherent in the data as a whole. Despite the influence of statistically irrelevant events in the data, experiments using simulated incomplete data and real life hydrometric network data have produced very encouraging results.

References

[1] A. K. C. Wong and H. C. Shen, "Data base partitioning for data analysis," in *1979 Proc. Int. Conf. Cybernetics and Society*, Denver, CO, pp. 514-518.

[2] H. C. Shen, M. S. Kamel, and A. K. C. Wong, "Intelligent data base management systems," in *Proc. 1983 Int. Conf. Systems, Man, and Cybernetics.*

[3] R. J. Thierauf, *Decision Support Systems for Effective Planning and Control.* Englewood Cliffs, NJ: Prentice-Hall, 1982.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York, Wiley, 1973.

[5] A. K. C. Wong and D. K. Chiu, "An event-covering method for effective probabilistic inference," *Pattern Recognition*, vol. 20, no. 2, pp. 245-255, 1987.

[6] D. K. Chiu and A. K. C. Wong, "Synthesizing knowledge: A cluster analysis approach using event-covering," *IEEE Trans. Syst., Man, and Cybern.*, vol. SMC-16, pp. 251-259, Mar./Apr. 1986.

[7] A. K. C. Wong and T. S. Liu, "Typicality, diversity, and feature pattern of an ensemble," *IEEE Trans. Comput.*, vol. C-24, pp. 158-181, Feb. 1975.

[8] S. S. Wilks, *Mathematical Statistics.* New York: Wiley, 1962.

[9] D. J. Hand, "Statistical pattern recognition of binary variables," in *Pattern Recognition Theory and Applications*, J. Kittler, K. S. Fu, and L. F. Pau, Eds. Dordrect, The Netherlands: D. Reidel, 1982, pp. 19-33.

[10] H. Gallaire, J. Minker, and J. Nicolas, "Logic and databases: A deductive approach," *Comput. Surveys*, vol. 16, no. 2, pp. 153-185, June 1984.

[11] S. Navathe, S. Ceri, G. Wiederhold, and J. Dou, "Vertical partitioning algorithms for database design," *ACM Trans. Database Syst.*, vol. 9, no. 4, pp. 680-710, Dec. 1984.

[12] F. M. Reza, *An Introduction to Information Theory.* New York: McGraw-Hill, 1961.

[13] B. Forte, M. de Lascurain, A. K. C. Wong, "The best lower bound of the maximum entropy for discretized two dimensional probability distributions," *IEEE Trans. Inform. Theory*, to be published.

[14] R. S. Garfinkel and G. L. Nemhauser, *Integer Programming.* New York: Wiley, 1972.

[15] J. C. Stoffel, "A classifier design technique for discrete variable pattern recognition problems," *IEEE Trans. Comput.*, vol. C-23, pp. 428-441, 1974.

[16] C. T. Ng and A. K. C. Wong, "On nonuniqueness of discretization of two-dimensional probability distribution subject to maximization of Shannon's entropy," *IEEE Trans. Inform. Theory*, to be published.

[17] R. A. Tapia and J. R. Thompson, *Nonparametric Probability Density Estimation.* Baltimore, MD: The John Hopkins University Press, 1978.

[18] S. Guiasu, *Information Theory with Applications.* New York: McGraw-Hill, 1977.

[19] R. Christensen, "Entropy minimax, a non-Bayesian approach to probability estimation from empirical data," in *Proc. IEEE 1973 Int. Conf. Cybernetics and Society*, pp. 321-325.

[20] R. D. Smallwood, *A Decision Structure for Testing Machine.* Cambridge, MA: M.I.T. Press, 1962.

[21] 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.

[22] A. K. C. Wong, "Problem definition for the pattern analysis and display systems of hydrometric networks in British Columbia," Report to Environment Canada, Mar. 1982.

Andrew K. C. Wong (M'79) received the Ph.D. degree from Carnegie-Mellon University, Pittsburgh, PA, in 1968.

For several years, he taught at Carnegie-Mellon. He is currently a Professor of Systems Design Engineering and the Director of the PAMI Group, University of Waterloo, Waterloo, Ont., Canada. In 1984, he also assumed the responsibility of Research Director, in charge of the research portion of the Robotic Vision and Knowledge Base Project at the University. He has authored and coauthored chapters/sections in several engineering books and published many articles in scientific journals and conference proceedings.

Dr. Wong is an Associate Editor of the *Journal of Computers in Biology and Medicine.*

David K. Y. Chiu received the M.Sc. degree in computing and information science from Queen's University, Kingston, Ont., Canada, in 1979 and the Ph.D. degree in system design engineering from the University of Waterloo, Waterloo, Ont., in 1986.

Currently, he is on the faculty of the University of Guelph, Guelph, Ont., Department of Computing and Information Science. His research interests include pattern analysis and machine intelligence, knowledge based system and computer vision.