Pattern Discovery: A Data Driven Approach to Decision Support

Andrew K. C. Wong, *Senior Member, IEEE*, and Yang Wang, *Member, IEEE*

***Abstract*—Decision support nowadays is more and more targeted to large scale complicated systems and domains. The success of a decision support system relies mainly on its capability of processing large amounts of data and efficiently extracting useful knowledge from the data, especially knowledge which is previously unknown to the decision makers. With a large scale system, traditional knowledge acquisition models become inefficient and/or more biased, due to the subjectivity of the experts or the pre-assumptions of certain ideas or algorithmic procedures. Today, with the rapid development of computer technologies, the capability of collecting data has been greatly advanced. Data becomes the most valuable resource for an organization. This paper presents a fundamental framework toward intelligent decision support by analyzing a large amount of mixed-mode data (data with a mixture of continuous and categorical values) in order to bridge the subjectivity and objectivity of a decision support process. By considering significant associations of artifacts (events) inherent in the data as patterns, we define patterns as statistically significant associations among feature values represented by joint events or hypercells in the feature space. We then present an algorithm which automatically discovers statistically significant hypercells (patterns) based on: 1) a residual analysis, which tests the significance of the deviation when the occurrence of a hypercell differs from its expectation, and 2) an optimization formulation to enable recursive discovery. By discovering patterns from data sets based on such an objective measure, the nature of the problem domain will be revealed. The patterns can then be applied to solve specific problems as being interpreted or inferred with. Classifiers can be developed, probabilistic density functions can be estimated, association rules can be generated and pattern based data query can be formulated. It presents a data driven system and a process from data to patterns and from patterns to applicable rules/models for decision support.**

***Index Terms*—Classification, clustering, data mining, decision support, event association, pattern discovery, residual analysis.**

I. Introduction

TODAY, intelligent decision support based on artificial intelligence, knowledge-based reasoning, data mining, data fusion, decision analysis, and optimization is more and more targeted to large scale complicated systems and domains [1]–[4]. The success of an intelligent decision support system relies significantly on its capability of processing large amounts of data and extracting useful knowledge from it, especially when patterns and knowledge are previously unknown to the decision makers. With the rapid development of computer technologies, the capability of collecting data has been greatly advanced. Now, more than ever, data is better organized and can be easily accessed for retrieval and analysis; a valuable resource for an organization. However, extracting information and knowledge contained in the data is still a very difficult problem, especially when the system is limited by legacy data storage format and by a lack of expert knowledge about the data, the visualization and the inefficiency of data mining tools and by there being too much customization. A major concern in the scientific and business world is to avoid DRIP (data rich information poor) [5] embarrassment. "How to get relevant information or to discover useful knowledge from a horrendous mass of data?"

Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Data mining is a step of knowledge discovery process consisting of particular algorithms to produce patterns. However, when data mining is applied in decision support, often information extracted from the data could be biased by the prior perception of a problem. A user may use the traditional data mining tools [3], [6] to obtain supporting evidences to confirm, or refute his/her preconceived ideas, yet he/she cannot be assured that there are no other comparable, or even better solutions to the problem and/or something important is missing in the search. For a complex problem with large size data, traditional knowledge acquisition and data mining models become obviously inefficient. They are likely to be biased due to the subjectivity dominated by the experts or the pre-assumptions of certain ideas and algorithmic procedures. Usually a long iterative trial and error process is used and that could be tedious, frustrating, and confusing.

The rationale of the current approaches is to provide a systematic way to prune the search space so as to acquire decision or classificatory rules inherent in the data. In most of the existing systems, accessory processes, such as pre-processing, data cleansing, filtering, attribute reduction are included [1] in order to remove noises, to bring out more relevant information from the data and to reduce the search space. These approaches allow users to mine from data what they would like to know or verify, using queries, or rules acquired from the training data specific to the pre-set classification, or query goals. They have to depend on prior knowledge, selected decision parameters and/or preconceived classificatory framework. Nevertheless, they are unable to discover new patterns and knowledge so as to provide a more objective and broader support base for

Manuscript received July 9, 2002; revised October 7, 2002 and January 23, 2003. This work was supported in part by Pattern Discovery Software Systems Ltd. This paper was recommended by Guest Editors K. W. Hipel and N. P. L. Cassaigne.

A. K. C. Wong is with the PAMI Lab, Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: akcwong@pami.uwaterloo.ca).

Y. Wang is with the Pattern Discovery Software Systems, Ltd., Waterloo, ON N2L 5V4, Canada (e-mail: yang@patterndiscovery.com).

Digital Object Identifier 10.1109/TSMCC.2003.809869

decision-making. Such processes could be very biased and usually involve long iterative search and examination-re-examination process. For large databases with unanticipated variations, this would be slow to produce useful results. Furthermore, three other related issues are also of concern to the decision-makers. They are: the flexibility and versatility of the pattern discovery process; the transparency of the supporting evidences, and; the processing speed.

For flexibility and versatility, most decision makers do not want just a single "optimal" solution. They would like to know other comparable alternatives and their relative merits. Sometimes they would even like to shift their decision goal or parameters a little so as to have a broader assessment base of the situation. Hence, too rigid a decision framework, or too narrow a set of decision rules or parameters, may not serve them comfortably.

For transparency, a decision maker would not be too comfortable relying on a black box for a final answer, but would rather, like to know in greater details what the underlying supporting evidences and their statistical significance could be. From the deterministic, or statistical patterns, that come to his awareness, he may like to draw inferences not only from the revealed patterns, but also from other external supporting evidences, so that he might come up with a comprehensive explanation or interpretation of the problem. In our view, it is only when the discovered patterns can be interpreted, explained, rationalized, confirmed, and made explicit, can they become verifiable knowledge. Otherwise, they are simply statistical patterns with limited value to a decision maker.

The speed of the pattern and rule extraction process is often crucial to a decision making process. This is true, not only because of the imminent response often required for a quick decision, but also that interactive processes are often needed in the incremental information and knowledge extraction process for a comprehensive decision. In view of this, it is indeed difficult for the decision makers to have an interactive exploratory process if they have to wait for hours, or even minutes, for an intermediate solution. In many situations, based on what they learn or discover from the explicit patterns displayed on the screen, they could make a judicial decision or they may like to look further into the data to discover more supporting evidences.

In summary, if machine intelligence could be used comfortably by the decision makers, it must be able to

1) discover multiple patterns from a database without relying on prior knowledge, so as to provide comprehensive, exhaustive, and unbiased supporting evidences;
2) cope with multiple and flexible decision scenarios and objectives;
3) provide explicit discovered patterns and rules associated to their problem of concern for interpretation, so as to enhance their understanding of the problem; and
4) render a high-speed interactive mode for information and knowledge extraction from the data.

To respond to these needs, a pattern discovery approach has been advanced [7]–[9]. Primarily data-driven, this approach is able to discover in an unbiased and exhaustive manner, statistically significant event or data associations (known as high order patterns) automatically, and to generate from them decision rules, classificatory modules for categorization, classification, prediction, and forecasting. A novel and unique feature of this system is that it is able to discover multiple explicit patterns of high order very fast, and rank them according to their statistical significance for interpretation, comparison, and assessment so that greater understanding of the data can be achieved and thereby better decisions can be made. Within this theoretical framework, a software system[1] has been implemented. Surrounding such core technology, many new modules are developed in support of decision making. This includes attribute clustering, entity clustering [10], class-dependent discretization of continuous data [11], classification [9], and forecasting [12], [13]. In this paper, we emphasize on the new theoretical development and demonstrate the performance of the proposed framework, especially that in real-world problems. Because the pattern discovery approach presented in this paper evolves essentially and mostly from our work over the last twenty five years, a brief historical background, rationale, and description of the system is given in next section.

II. Brief Description of Pattern Discovery

To discover patterns from data started in the seventies, the first author attempted to search for a quantitative basis of information measures and statistical patterns from bio-molecular data [14], [15], English text [16], digital images [17], [18], and biomedical and clinical data [19], [20]. With the belief that information in biomolecule sequences is coded for biomolecule structures and functions, an endeavor was made to obtain quantitative information measures and statistical patterns in biomolecules. Biomolecule data was chosen, so as to provide empirical evidences that information can indeed be measured, and statistical patterns be found in biomolecule code sequence to reveal the underlying biochemical and taxonomic characteristics of an ensemble of molecules such as cytochrome c from different spices. Later, along this line of thought, quantitative measures of information were found in English texts [16] and images [17], [18] based on how the data deviated from equal-probability and independence mode. This formed the early basis of our pattern discovery approach. Later, based on this concept, pattern recognition methods for discrete valued and continuous data were developed with applications in biological signal analysis [19], [21], and clinical diagnosis and prognosis [19], [20], [22], [23].

In the late seventies and early eighties, we realized that if the dimension of a relational database was large, the definition of patterns in the classical pattern recognition framework might not be too meaningful, for in the database, attributes might form interdependent groups and these groups could be independent of each other. Hence the concepts of classification and clustering based on pattern vectors broke down. Soon database partitioning was proposed [24], [25] which attempted first to cluster the attributes into interdependent groups according to the interdependence measure from information theory before clustering the entities of each group into subgroups. In that sense, each subgroup reflected local patterns based on the notion of attribute

[1]*discover*e*, developed by Pattern Discovery Software Systems Ltd.

interdependency. Later, various pattern recognition methodologies were developed [26]–[28]. They are still based on the interdependency of features as random variables or restriction of random variables. In the nineties, we shifted our pattern recognition paradigm from the variable level to event level based on event associations although only second order event associations was addressed in the early days with APACS [29].

To overcome the limitation of APACS, which is based on first order relationship, a higher order pattern discovery is designed [7], [9] for discrete data sets. In our method, patterns inherent in data are defined as statistically significant associations of two or more primary events of different attributes. To detect patterns from data with noise, *adjusted residual* analysis in statistics is engaged. It guarantees (with a fixed confidence level) that the patterns detected are significant. The discovered high order patterns can then be used to support decision making tasks such as classification and clustering. At the same time, high order pattern discovery with continuous data was also being advanced. In [8], events in continuous space are defined as Borel sets and the pattern discovery process is formulated as an optimization problem which recursively partitions the sample space for the best set of significant events. Tasks such as classification and probability density estimation can be easily performed when the patterns have been discovered. Significant results on both artificial and real-world data have been obtained. These automatic pattern discovery methodologies become ideal tools to support various types of decision making tasks.

In this paper, we present the new development along the path and the extension of pattern discovery theories and methodologies in a new framework as mixed-mode data analysis for decision support. It includes: the discovery of significant patterns that may interest the decision maker and, interpreting and inferencing with patterns for decision support. We give the details of various operations related to a decision making process.

III. Data, Events and Patterns

The purpose of pattern discovery, be it referred to as conceptual clustering [30], [31], or rule induction [32], is to find the relations among the attributes and/or among their values. We seek to find events whose occurrences are significantly different from those based on a "default" model. Thus, the co-occurrence of events detected may reflect the nature of the data set and provide useful information for future inferences and reasoning. We call these relationships patterns. Formal definitions will be given later.

Consider, our problem domain is described by N attributes (variables, or features), each of which can assume values from its own domain, be it real or a finite alphabet. Let $\mathbf{X} = \{X_1, \ldots, X_N\}$ represent this attribute set. Then each attribute, $X_i, 1 \leq i \leq N$, can be seen as a random variable taking on values from its domain d_i. For a continuous attribute, $d_i \subset \Re$ and for a categorical (discrete) attribute, $d_i = \{\alpha_i^1, \ldots, \alpha_i^{m_i}\}$, where m_i is the cardinality of the alphabet.

A. Generalized Event

In pattern discovery, we are concerned, in general, with high dimensional data. To draw an equivalence with probability theory, we simply consider an N-dimensional data point, either an observation or a measurement, as the outcome of an imaginary experiment on our problem domain. Denote an experimental outcome as ω. The set of all possible outcomes is the sample space, Ω, and may be finite or uncountably infinite.

When the sample space is finite, an event is defined as a subset of Ω. This subset may consist of a single event in which it is called an elementary event. If the subset is empty, it is called a null event. For events in a finite sample space, probabilities are assigned to elementary events. The axioms of probability are then easily satisfied. In pattern discovery, this definition of events applies when the data of interests is discrete. The discrete nature of the data may be manifested in two ways, a finite number of numerical values or a finite number of categorical labels. In the latter case, the data is symbolic. In either cases, the values are composed of the alphabet of the variable and an event consists of either one (primary event) or more (compound event) of the values in the alphabets [7].

For continuous data, the sample space is generally taken to be the N-dimensional Euclidean space, $\Re^N$. Clearly, this sample space consists of an uncountable number of outcomes. The use of elementary events becomes intractable, especially when defining probabilities. Instead, events are defined to be subsets of Ω which form a Borel σ-field, $B(\Re^N)$ [8]. A Borel σ-field is the collection of all rectangles in $\Re^N$, where a rectangle A is defined as a subset of $\Re^N$ if it has the form

$$A = I_1 \times \cdots \times I_N = \{\mathbf{x} : x_i \in I_i, 1 \leq i \leq N\} \qquad (1)$$

where $I_i = (a_i, b_i]$ and $-\infty < a_i < b_i < \infty$. Equivalently, the Borel σ-field is the smallest σ-field containing all rectangles. The sets in the Borel fields are the Borel sets, which include rectangles and countable unions and intersections of rectangles. An example can be seen in Fig. 1.

There are two advantages of defining events in this way. First, there is a nice geometric perspective. In Fig. 1 we see that events are just hyper-rectangles, or countable unions, or intersections of hyper-rectangles. Secondly, a probability measure can now be assigned to events without violating the axioms of probability.

In real world applications, however, we will have a sample space containing both discrete and continuous variables. That is to say, some of the attributes or variables take values from Euclidean space and others from finite alphabets. We then extend the concept of Borel field to mixed-mode space $\mathfrak{M}^N$. The extended Borel σ-field becomes the hyper σ-field, which is the collection of all hypercells in $\mathfrak{M}^N$. The following definitions then follow.

Definition 1: Consider the sample space $\mathfrak{M}^N$. Let

$$I_i = \begin{cases} (a_i, b_i] : -\infty < a_i < b_i < \infty, & \text{if } X_i \in \Re \\ \alpha_i : \alpha_i \in d_i, & \text{if } X_i \text{ is discrete.} \end{cases}$$

A *hypercell* H of $\mathfrak{M}^N$ is called a hypercell if it has the form

$$H = I_1 \times \cdots \times I_N = \{\mathbf{x} : x_i \in I_i, 1 \leq i \leq N\}.$$

For dimensions defined on the $\mathfrak{M}^N$, this can be visualized as a hyper rectangle like E_1 in Fig. 1. For a dimension (called categorical dimension) defined on a finite set of discrete values, we

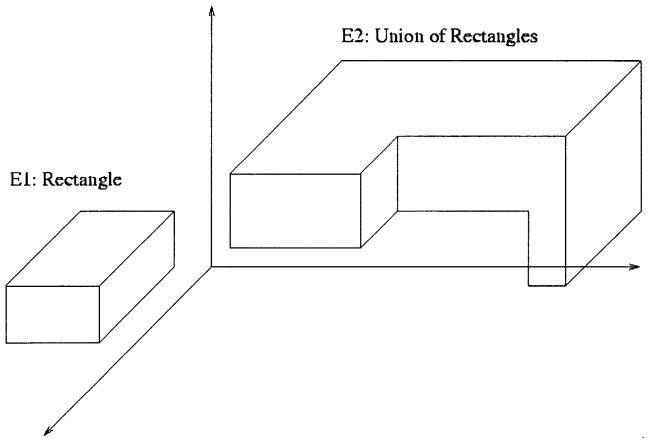


Fig. 1. Example of events in $\Re^3$.

could imagine an I_i as a finite set of discrete values (or categorical labels) assigned to subsets of a point set that make up the event space. Each label can be considered as a categorical label of a particular category. In that sense, it is not only the hyper rectangles that define events but also various subsets of categorical labels for various types of categories in the categorical dimensions. An event can then be visualized as a point that falls within a hyper rectangle and also bears values of various categorical dimensions. This is formally defined in the next paragraph.

The σ-field generated by the collection of all hypercells in $\mathfrak{M}^N$ is the hyper σ-field of $\mathfrak{M}^N$. As in the continuous case, the hyper σ-field is the smallest σ-field containing all hypercells. The sets in the field are called hyper sets. Hyper sets include hypercells and countable unions and intersections of hypercells.

Definition 2: An *event* in $\mathfrak{M}^N$ is a hyper set.

Apart from its location in space, a few other quantities of an event are also defined.

Definition 3: The *volume* of an event is the hypervolume of the N-dimensional subspace occupied by the event.

Suppose we have a data set $D = \{\omega\}$ from a sample space $\mathfrak{M}^N$.

Definition 4: The *observed frequency*, o_E, of an event E in the sample space Ω is the number of data points that fall within the volume of E. If we denote $\{\mathbf{x}\} \in D$ as the finite set of points falling inside the volume of E, then $o_E = |\{\mathbf{x}\}|$, where $|\cdot|$ denotes cardinality.

Definition 5: The *probability* of an event E is intuitively estimated by the proportion of data points contained in the event

$$\hat{P}_E = \frac{o_E}{|D|}. \tag{2}$$

B. Pattern

Within the classical pattern recognition framework (including artificial neural network), patterns are usually referred to as pattern vectors in the N-dimensional feature space. When the problem is small and can be contained, it may be possible to isolate a set of features for the purpose of a specific analysis. A database usually records information not confined to a particular interest or a specific problem domain. Some of the features may not be related to each other. In view of this, as observed in [24], [25], classification or clustering of the data may not be meaningful. Although, for solving a specific problem, selecting an optimal subset of features from a large data set has been proposed, yet in reality, such an approach faces many challenges. On one hand, the diversity of the real world situation often manifests that the subset of optimal features for one group may not be the same for the other and within the feature groups some of the feature values could be irrelevant to the problem or simply noise. On the other hand, when dimensionality is high, selecting optimal subset with no prior knowledge is very difficult. In our view, what actually makes up the patterns for a certain class (or group) are the significant event associations inherent in the data of that class or group. It is this group of event associations which make up the pattern vectors pertaining to a class. Statistical event associations discovered in a data set are the most fundamental set of information in the database or the data set. Hence, we consider patterns as associations. Here below we will give the formal definitions.

Definition 6: Let Ω be the sample space and $g(\cdot)$ be a test statistic corresponding to a specified discovery criterion c. Let θ_c^α be the critical value of the statistical test at a significant level of α. A *pattern* is an event E that satisfies the condition

$$g(E) \geq \theta_c^\alpha. \tag{3}$$

The test statistic $g(\cdot)$ measures the degree to which an event satisfies the objective of the discovery. If the test statistic is 2-tailed, three types of events can be identified, according to the value of $g(E)$.

1) $g(E) \geq \theta_c^\alpha$—The event E, as defined above, is a positive pattern, or a positive significant event.
2) $-\theta_c^\alpha < g(E) < \theta_c^\alpha$—The event E is an insignificant event.
3) $g(E) \leq -\theta_c^\alpha$—The event E is a negative significant event, which are contrary to the discovery objective.

With a 1-tailed test statistic, only insignificant and positive significant events are applicable.

Any event E of dimension $N > 1$ can be interpreted as the joint occurrence of lower dimensional events. This leads to the following simple definition [7]:

Definition 7: An *event association* is a significant joint occurrence of low-dimensional events. In particular, any N-dimensional event ($N > 1$) can be considered an event association, composed of N one-dimensional events.

Fig. 2 demonstrates an event as a hypercell in a three–dimensional (3-D) space. The event can also be viewed as the joint occurrence of three 1-D intervals (for continuous variable) or values (for discrete variable), I_x, I_y, and I_z.

To illustrate a hypercell of mixed-mode data, let us consider a four-dimensional mixed-mode space $\mathfrak{M}^4$, with the first three dimensions as those given in Fig. 2 and the fourth dimension assuming categorical values, say $\alpha = (\alpha_1, \alpha_2, \alpha_3)$. We could consider an event E as an association, say, of I_1, I_2, I_3, α_2.

We appreciate that terms such as "pattern," "significant event" and "event association" share the same meaning, but with variation in interpretation. While the word "pattern" has intuitive appeal, its statistical basis is intimately implied by

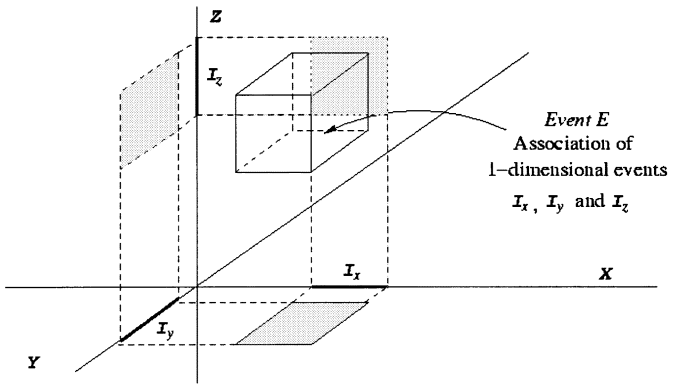


Fig. 2. Three-dimensional event association.

the term "significant event." On the pragmatic front, "event association" offers a geometric perspective, which will be useful in interpreting discovery results.

IV. Pattern Discovery

With the definitions of event and pattern, pattern discovery becomes a process of searching significant events in the sample space. Theoretically, candidate events may lie anywhere in the sample space. In practice, we restrict the search to a compact subspace of the sample space as demarcated by the available samples.

Definition 8: Suppose we have a data set with sample space Ω. *Pattern Discovery* is the search for significant events (hypercells) in a compact subspace of the sample space Ω demarcated by the available data set D.

This definition covers a wide range of data driven, statistics based decision support approaches. For example, trees (decision trees, dependence trees, CART, etc.) and neural networks can be viewed as inherently event discovery mechanisms.

Recall that decision trees partition the sample space into subregions each time a node splits. Suppose the sample space is mixed-mode. At a given node i, if i is continuous, the range of a variable X undergoes a binary split, say, $X < a$ and $X \geq a$, where a is some scalar value. If i is discrete, the variable X also undergoes a binary split, say, $X = a$ and $X \neq a$. Whatever the type of the variable, the subregion corresponding to node i is then partitioned into two smaller regions. The continuation of this process successively demarcates regions of space. Hence, the subspace, delineated by tree leaves, represents a subspace with maximally homogeneous class composition. Whereas, the subspace is simply a hypercell.

In neural networks, discrete variables are often coded before feeding into the networks in a way they can be treated as continuous variables. The relationships of both neural classifiers and neural function approximators with event pattern discovery have been stated in [33].

A. Pattern Discovery as Residual Analysis

With the definition of patterns, our pattern discovery paradigm discovers associations at event level, that is a localized approach. In statistics, residual analysis provides valuable local information about the organization of a domain characterized by a sample set.

We propose to use residual as the statistic in testing the significance of the pattern candidates. The problems of unstandardized, wildly varying quantities, and arbitrary thresholds are overcome by the use of residual. Furthermore, the residual is easily interpreted in terms of the degree of satisfaction of the discovery.

Suppose we have an event E, the following terms are defined:

Definition 9: The *residual* of an event E against a pre-assumed model is defined as the difference between the actual occurrence of the event and its expected occurrence. That is

$$\delta_E = o_E - e_E \tag{4}$$

where e_E is the expected occurrence under the pre-assumed model estimated by the given sample set.

In practice, the residual is first standardized before any analysis is conducted.

Definition 10: [34] The *standardized residual* of event E is defined as the ratio of its residual and the square root of its expectation

$$\zeta_E = \frac{\delta_E}{\sqrt{e_E}}. \tag{5}$$

Definition 11: [34], [35] The *adjusted residual* of event E is defined as

$$\gamma_E = \frac{\delta_E}{\sqrt{\nu_E}} \tag{6}$$

where ν_E is the variance of δ_E.

If the pre-assumed model is log-linear, the standardized residual and the adjusted residual are both asymptotically normal distributed [34]–[36]. The adjusted residual is a better estimate than standardized residual. Hence, we use γ_E as $g(\cdot)$ as the test statistic for significant hypercells.

For a log-linear default model which is hierarchical and decomposable, we can obtain a close form maximum likelihood estimate of the variance of the residual. See [37] for the general format and [36] for the special cases in mutual independent models. The calculation of expected occurrence of an event is tightly related to the chosen default model. In our system, there are two special default models.

1) uniform distribution;
2) mutual independence.

In the uniform distribution case, the data is uniformly distributed throughout the volume of the bounded subspace, $S \subset \Omega$, under consideration. Hence, within an event E, of volume, v_E, the expected number of observations would be

$$e_E = M \cdot \frac{v_E}{V} \tag{7}$$

where V is the volume of S, and M is the total number of observations.

With the default uniform distribution model, the data is tested against randomness, therefore, a pattern is a hypercell with significantly larger number of data points. We call it *concentration-driven discovery*.

In the independent model, all variables are mutually independent. The joint probability is just the product of the marginal probabilities. Under this assumption, for an event E, the expected number of observations would be

$$e_E = M \cdot \prod_{i=1}^{N} \hat{P}_i \tag{8}$$

where $\hat{P}_i$ is the marginal probability, which can be estimated as

$$\hat{P}_i = \frac{n_i}{M} \tag{9}$$

where n_i is the number of data points falls into the interval I_i (if variable i is continuous) or are equal to α_i (if variable i is discrete) when all data points are projected to the i-th axiom.

With the default independent model, the data is tested against independence, therefore, a significant event marks a subspace where the different dimensions of the data exhibit strong relationships or interactions. Regions populated with data but void of dependencies are detected as insignificant events. Negatively significant events implies that the space is sparser than expected, which may support some type of dependency. We call it *dependency-driven discovery*.

B. Pattern Discovery as Optimization

In practice, to automatically discover patterns in a given sample set, we need to reformat the discovery into a mathematical problem, so that we can take advantage of standard computing scheme. Now that we have established pattern discovery as a process of searching for events which maximize a discovery criterion (the test statistic or adjusted residual), it is natural to formulate discovery as an optimization problem.

If we have an arbitrary hypercell E, we can represent E by a pair $\{C, L\}$, where C represents one of the corners of E, and L represents the lengths of the edges. We have,

$$C = \{c_i | 1 \le i \le N\}$$
$$L = \{l_i | 1 \le i \le N\}$$

where c_i is the i-th coordinate of the reference point (corner) and l_i is the length of the edge of the i-th axiom. We further define

$$l_i = \begin{cases} b_i - a_i, & if\ x_i\ is \text{ continuous} \\ 1, & if\ x_i\ is \text{ discrete} \end{cases} \tag{10}$$

The pattern discovery problem is to

$$\begin{aligned} &\text{maximize} : O(E) \\ &\text{subject to} : \begin{vmatrix} 0 < l_i \le b_i & 0 \le i \le N \\ a_i < c_i \le b_i & 0 \le i \le N \\ \prod_{i=1}^{N} l_i \ge V_{\min} \end{vmatrix} \end{aligned} \tag{11}$$

where $V_{\min}$ is the minimum volume under consideration. See [8], [33], [36] for details how to determine this value.

The objective function $O(E)$ is the adjusted residual:

$$O(E) = \begin{cases} \gamma_c(E), & \text{concentration discovery} \\ \gamma_d(E), & \text{dependency discovery} \end{cases} \tag{12}$$

With the objectives and constraints, optimization methodologies, such as, generic algorithm and stimulated annealing can be applied recursively in the sample space to discover the significant hypercells as patterns.

V. Inferencing With Patterns for Decision Support

A. Building Classifiers

Classification is an important task for decision support. A classifier is a model of the problem domain that is able to guess the "best" membership of an unseen object according to the rules learned from previous experience. The "class" attribute is normally categorical.

The classification problem can be formalized as determining the value of a missing discrete attribute, X_i, given a set of observations in sample space Ω containing X_i. Let $(\underline{\mathbf{X}}, Y)$ be jointly distributed random variables with q-dimensional vector $\underline{\mathbf{X}}$ denoting a feature (observation) vector in space $\Phi(\Phi \subset \Omega)$ and Y denoting the attribute whose value is to be determined. The missing-value problem is to find a decision rule $\rho(\cdot)$ that maps Φ into the domain of Y such that certain properties of the original data set are preserved. The feature vector $\underline{\mathbf{x}}$ is called a new *observation* or a new *object*, and Y its class label, or predicting attribute. It is assumed *a priori* that attribute Y is discrete. In traditional classification, or supervised learning, Y is a special pre-defined attribute called "class" while in unsupervised learning or for flexible prediction, Y can be any attribute describing the problem domain.

We argue that the patterns detected by dependency-driven discovery can be dynamically re-organized to be a classifier, solving the missing-value problem.

Based on the mutual information in information theory, the difference in the gain of information when Y takes on the value y_i and when it takes on some other values, given $\underline{\mathbf{x}}$, is a measure of evidence provided by $\underline{\mathbf{x}}$ in favor of y_i being a plausible value of Y as opposed to other values. This difference, denoted by $W(Y = y_i / Y \ne y_i | \underline{\mathbf{x}})$, is defined as the *weight of evidence*, which has the following form:

$$\begin{aligned} W\Big(Y = & y_i / Y \ne y_i | \underline{\mathbf{x}}\Big) \\ &= I(Y = y_i : \underline{\mathbf{x}}) - I(Y \ne y_i : \underline{\mathbf{x}}) && (13) \\ &= \log \frac{P(Y = y_i | \underline{\mathbf{x}})}{P(Y = y_i)} - \log \frac{P(Y \ne y_i | \underline{\mathbf{x}})}{P(Y \ne y_i)} && (14) \\ &= \log \frac{P(\underline{\mathbf{x}} | Y = y_i)}{P(\underline{\mathbf{x}} | Y \ne y_i)} && (15) \end{aligned}$$

where $I(\cdot)$ is the mutual information.

The weight of evidence is positive if $\underline{\mathbf{x}}$ provides positive evidence supporting Y taking on y_i, otherwise, it is negative, or zero. A negative weight of evidence implies that there is negative evidence provided by $\underline{\mathbf{x}}$ against Y taking on the value y_i. In other words, it is more likely for this attribute to take on another value. A zero weight of evidence suggests that $\underline{\mathbf{x}}$ is irrelevant to the prediction of Y.

To calculate the weight of evidence, we need to estimate the conditional probabilities or know the distribution, which are not available on hand. However, according to [7], [36], the weight

of evidence can be decomposed if significant event associations related to y_i and $\underline{\mathbf{x}}$ are known. That is

$$\begin{aligned} W\Big(Y =& y_i/y \neq y_i|\underline{\mathbf{x}}\Big) \\ &= \log \frac{P(\underline{\mathbf{x}}_1|Y=y_i)}{P(\underline{\mathbf{x}}_1|Y\neq y_i)} + \cdots + \frac{P(\underline{\mathbf{x}}_n|Y=y_i)}{P(\underline{\mathbf{x}}_n|Y\neq y_i)} \\ &= W(Y=y_i/Y\neq y_i|\underline{\mathbf{x}}_1) + \cdots \\ &\quad + W(Y=y_i/Y\neq y_i|\underline{\mathbf{x}}_n) \\ &= \sum_{k=1}^{n} W(Y=y_i/Y\neq y_i|\underline{\mathbf{x}}_k) \end{aligned} \tag{16}$$

where $\underline{\mathbf{x}}_k$ is a sub-event of $\underline{\mathbf{x}}$ and satisfies

$$\begin{aligned} &\underline{\mathbf{x}}_p \bigcap \underline{\mathbf{x}}_q = \phi, \quad p \neq q, 1 \leq p, q \leq n \text{ and} \\ &\bigcup_{p=1}^{n} \underline{\mathbf{x}}_p = \underline{\mathbf{x}}. \end{aligned}$$

Based on [29], [36], events which are not statistically significant can be considered irrelevant for the inference process. The underlining model used in the pattern discovery process assumes that the attributes are mutually independent. This implies that, if an event is not statistically significant, the primary events in it are randomly combined. The mutual information is approximately zero. When calculating the weight of evidence, these events can be eliminated. Thus, only the significant event associations discovered from the data set are used in the inference process. Then the calculation of weight of evidence is to find a proper set of disjoint significant event associations from $\underline{\mathbf{x}}$ and to sum each individual weight of evidence provided by each of them. That is to maximize

$$W(Y=y_i/Y\neq y_i|\underline{\mathbf{X}}) = \sum_{q=1}^{m} W\left(Y=y_i/Y\neq y_i|\underline{\mathbf{X}}_q\right) \tag{17}$$

with sub-compound events $\underline{\mathbf{x}}_1, \ldots, \underline{\mathbf{x}}_m$, such that $(\underline{\mathbf{x}}_q, Y=y_i)$ $(1 \leq q \leq m)$ is a significant event association and the intersection between $\underline{\mathbf{x}}_p$ and $\underline{\mathbf{x}}_q$ is empty if $p \neq q$. Using significant association patterns in the inference process makes it possible not to go back to the original data set.

The classification process based on weight of evidence can be summarized as follows. A set of primary events $\underline{\mathbf{x}}$ are observed. The value of an unobserved attribute Y is going to be determined with the significant event associations discovered from the training data set. Given a set of significant associations related to attribute Y, the weight of evidence for each possible value of Y provided by the observation is calculated. These weights are compared to find the most plausible value of Y. The value y_i can be considered as the most plausible value if the following conditions stand:

$$\begin{cases} W(Y=y_i/Y\neq y_i|\underline{\mathbf{x}}) > W(Y=y_j/Y\neq y_j|\underline{\mathbf{x}}) \\ W(Y=y_i/Y\neq y_i|\underline{\mathbf{x}}) > 0 \end{cases} \tag{18}$$

where $1 \leq j \leq d_Y$ and $j \neq i$. For more details, refer to [9], [36]

Unlike traditional classifier, using the patterns as a model, any missing values of a discrete variables can be classified.

B. Multivariate Probabilistic Density Estimation

The estimation of the probability density function (pdf) is a central problem in multivariate data analysis, as evidence by the large body of literature from a diversity of disciplines. The density function gives a probabilistic description of the data's organization. Such a description is useful for data interpretation, regression, classification and prediction. In this section, we demonstrate how to estimate the probability density function for both discrete and continuous variables with patterns by concentration-driven discovery.

1) Discrete pdf Estimation: To simplify notation, the event indicator function is defined.

Definition 12: The *indicator function* for a event, E_i, is defined as,

$$I_j(\mathbf{x}) = \begin{cases} 1, & if\ \mathbf{x} \in E_j \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

Employing the definitions for event volume and observed frequency, we can assign to an event, E_j, the following probability density estimate

$$\hat{p}_j = \frac{o_j}{M \cdot v_j} \tag{20}$$

where, as usual, M is the total number of sample points under consideration. This definition is along the line of the general nonparametric density estimate of Duda *et al.* [38]. Note that this probability density is just the probability estimate $\hat{P}_j$ of (2), divided by the event volume v_j. We see immediately that the normalization condition

$$\sum_j \hat{P}_j = \sum_j \hat{p}_j \cdot v_j = 1$$

is satisfied.

To obtain a discrete probability density function, $\hat{p}$, valid for the entire sample space, since the events, $\{E_j\}$, do not overlap, we may write compactly

$$\hat{p}(\mathbf{x}) = \sum_j I_j(\mathbf{x}) \cdot \hat{p}_j \tag{21}$$

where $\mathbf{x} \in \mathfrak{M}^N$ and $I_j(\mathbf{x})$ is the indicator function previously defined. Note that for each $\mathbf{x}$, only one term in the summation will have $I_j(\mathbf{x}) \neq 0$ as the data point can only fall into one event.

2) Continuous pdf Estimation: The set of discovered events forms a discrete representation of the data. However, a continuous description of the data's organization is often preferred. Evidently, smoothing is an important method of data analysis. It turns out that the discrete representation of events can be easily relaxed to produce a smooth representation of the data's structure.[2]

The basic idea is to estimate a kernel for each event. The cohort of these kernels then provides a smooth approximation over the sample space. The most popular kernel is the multivariate

[2]For the nature of the problem, we default that the variables for continuous pdf estimation are all continuous.

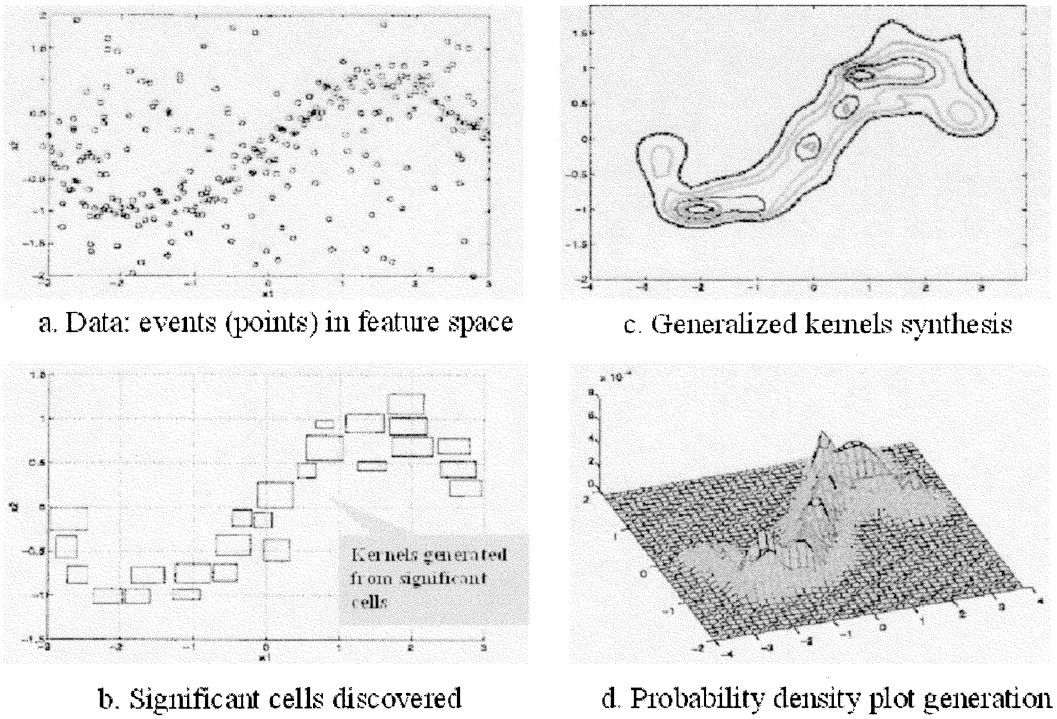


Fig. 3. From data to pdf. (a) Data events (points) in feature space. (b) Significant cells discovered. (c) Generalized kernels synthesis. (d) Probability density plot generation.

Gaussian kernel since it is continuous everywhere and it satisfies

$$\int_{-\infty}^{\infty} \psi(\mathbf{x}) d\mathbf{x} = 1 \tag{22}$$

where

$$\psi(\mathbf{x}) = \frac{1}{(2\pi)^{N/s}\Delta(\Sigma)^{1/2}} e^{-1/2(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)} \tag{23}$$

where μ is the mean, Σ is the covariance matrix, $\Delta(\Sigma)$ is the determinant of Σ and the prime denotes transpose.

To fit a kernel $\psi(\mathbf{x})$ to the event E, with observed frequency o_E, we simply compute the mean and covariance matrix for the data points contained in E.

$$\mu = \frac{1}{o_E} \sum_{i=1}^{o_E} \mathbf{x}_i \tag{24}$$

$$\Sigma = \sum_{i=1}^{o_E} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)' \tag{25}$$

The covariance matrix is always symmetric and positive semi-defined. In practice, the use of the kernel is restricted to cases when Σ is positive definite, so that the determinant is strictly positive. The smoothed events can be strategically combined to yield a continuous pdf estimate that satisfies probability axioms. Suppose discovery yields events $\{E_j : 0 \leq j \leq J\}$. The estimated discrete density is $\hat{p}_j$. Each events is fitted with a kernel, $\psi_j(\mathbf{x})$ as explained above. The combined pdf is estimated by

$$\hat{f}(\mathbf{x}) = \sum_{j=1}^{J} W_j \psi_j(\mathbf{x}) \tag{26}$$

where the kernel weight, W_j, is defined as

$$W_j = \frac{\hat{p}_j}{\sum_{j=1}^{J} \hat{p}_j} \tag{27}$$

We see that the normalization condition for densities is satisfied. An example of continuous density estimation can be illustrated by Fig. 3.

C. Interpretation of Patterns

Since events delineate a region of the sample space, production rules and/or association rules can be easily extracted. It is a straightforward process to transfer a significant hypercell (a pattern) into an association rule and measure the strength of the rule.

An association rule [39], [40] has the form of $X \Rightarrow Y$, where X and Y are sets of items (attribute-value pairs). Each association rule is measured by two parameters, the support and the confidence. The support of an association rule is basically the probability of $P(X,Y)$ and the confidence is the conditional probability $P(Y|X)$ estimated by frequency count from a given data set.

Suppose there is a pattern $E = \{I_i | 1 \leq i \leq N\}$, where I_i is either an interval $(a_i, b_i]$ or a discrete value α_i. We can easily transfer a pattern into an association rule by selecting a subset

of E as the right-hand side X and the rest as the left-hand Y and justify the strength of the rule by calculating the support and confidence. Without loss of generality, we assume there is only one item (attribute-pair) at the right-hand side. Then a rule has the form

$$\bigwedge_{i=1,\ i\neq j}^{r} I_i \Rightarrow I_j \text{ with quantitative measures, } 1 \leq j \leq N$$

In addition to support and confidence as in association rules, the weight of evidence can be used to measure the strength of such a rule. The weight of evidence in support of I_j, given $I_1, \ldots, I_{j-1}, I_{j+1}, \ldots, I_N$ can be calculated by (13). Recall that the weight of evidence is the gain of information when X_j takes on value of I_j and when it takes on other values, given $\{I_i | 1 \leq i \leq N, i \neq j\}$. It can be interpreted as a measure of similarity between I_j and rest of d_j, provided that $\{I_i | 1 \leq i \leq N, i \neq j\}$ happens. This similarity shows the strength of the association between the right-hand events and the left-hand conditions. The higher the value, the stronger the association.

When we transform a pattern into association rules, thresholds such as support and confidence can be used to filter resulting rules. Weight of evidence can also be calculated.

The major differences between pattern based association rule and the conventional association rules are

- A pattern is statistically significant. There is no guarantee that conventional association rules are statistically significant.
- A pattern based rule can be directly used in classification, as shown in the previous section. However, association rules are not designed for classification purpose [9], [41]. Two conventional association rules cannot be easily combined to integrate partial information available for a classification task. This is not a problem with pattern based association rules, as the weight of evidence provides such a mechanism to combine partial information. See [9] for more details.

D. Discovered Patterns as Queries for Class Data Retrieval

In the classical framework of pattern recognition, once a classifier is developed, it can be used to classify new set of data given. In today's data retrieval setting, the data could be widely distributed and it is very difficult to retrieve data at large for classification and analysis. Traditional pattern recognition or neural net approaches are based on the class probability model, discriminant models, distance/similarity measures from a subset class prototypes or class regions in the feature space. Almost all of these methods are comparison based. That is, if we want to classify a new candidate, we have to submit the candidate into the classifier and find out though comparison, to which class it belongs. If no prior knowledge is provided, we have to submit all candidates into the classifier for comparison before we could find out which of them belong to which class. This is a horrendous task if the database is huge.

With the pattern discovery approach, a new data retrieval and classification process can be realized. First, if one is given a training set of data, the system is able to generate all of the significant patterns or rules for a class and rank them according to their statistical strength. We can turn each significant pattern into a query and use it to retrieve all candidates matching the query from a huge database.

For instance, from the Wisconsin Cancer Data [42], [43], altogether 52 patterns are discovered with confidence level at 90%. The discovered patterns are ranked according to their residue value.

These patterns can be used to select a subset of data from very large data sources. An obvious choice for a query technology is the well-known structured query language (SQL). The following figure shows an internal application to induce these subsets of data based on certain patterns. The Wisconsin Breast Cancer dataset has buried within it a number of patterns. One such pattern relates the following attributes to class $= 4$ (cancerous):

$$\begin{aligned}&\text{Clump Thickness} = 5\\ &\text{and Normal Nuclei} = 10 \text{ and Mitosis} = 1\end{aligned}$$

The internal application, shown below, uses a variant of SQL query that is generated interactively by selecting attribute value pairs. The following query is generated:

The query:

Select* As Subset1
From BreastCancerWisconsin
Where ("Clump Thickness" = "5") And ("Normal Nucleoli" = "10") And ("Mitosis" = "1")

to retrieve all the cancerous patients in the database with the above patterns is shown the lower left window. Patients with all other patterns can also be retrieved.

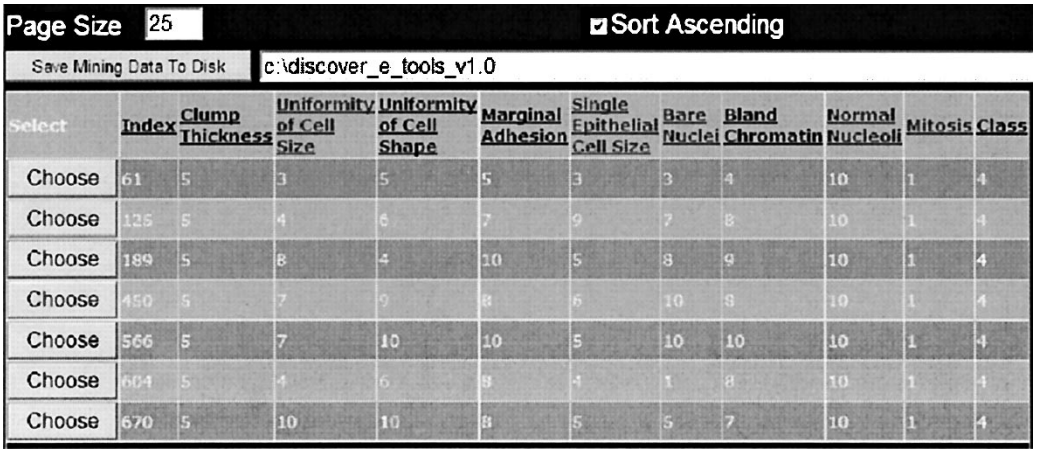
Fig. 4 shows the resulting data set.

In that case, we have shifted the class candidate retrieval from a comparison base into a query search. Furthermore, for all those candidates, we know also based on which pattern they are retrieved and classified. In the advent of information network of data and world wide web, this would have significant impact on knowledge retrieval and sharing.

VI. Summary and Discussion

In this paper, we have presented the new development in the framework of data driven decision support based on pattern discovery. It has been evolved for over twenty years and matured and expanded rapidly in the last few years [7]–[10]. In this paper, we have presented

1) the motivation, historical background and the rationale of our approach;
2) a novel unified framework to define and represent mixed-mode data, the most general and common data encountered in today's database;
3) the theoretical basis of pattern discovery based on statistical residual and optimization principles;
4) a novel and unified framework to represent probability models for discrete, continuous and mixed-mode data in the form of pdf;



Page Size 25 Sort Ascending

Save Mining Data To Disk c:\discover_e_tools_v1.0

Select	Index	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitosis	Class
Choose	61	5	3	5	5	3	3	4	10	1	4
Choose	125	5	4	6	7	9	7	8	10	1	4
Choose	189	5	8	4	10	5	8	9	10	1	4
Choose	450	5	7	9	8	6	10	8	10	1	4
Choose	566	5	7	10	10	5	10	10	10	1	4
Choose	604	5	4	6	8	4	1	8	10	1	4
Choose	670	5	10	10	8	5	5	7	10	1	4

Fig. 4. Result of the query.

5) an inferencing system using the discovered patterns and weight of evidence for classification and prediction of missing values under a general and flexible framework;
6) a new way of data retrieval by converting the discovered patterns of various orders (ranked according to their statistical significance if preferred) for each class into queries for class data retrieval in a distributive database setting with unlimited size—a very important paradigm shift from a comparison based approach of class pattern recognition and retrieval to automatic and comprehensive query-based search; and
7) supporting validation evidences of the efficacy of the proposed framework and its new development in solving large scaled problem with online interactive capability.

From data to patterns and from patterns to rules/models, this pattern based approach tries to bridge the gap between data and knowledge. We argue that the real value of data (or information) resides in the internal event associations occur in the domain. Discovering these patterns without biased assumptions is the first step in utilizing the collected data for decision support. In our framework, patterns in mix-mode data are discovered by residual analysis and an optimization process.

The patterns alone, however, only reflect local or point information, which has limited power in decision making. To unleash the power of the patterns, we have to develop inference mechanisms that bring out the applicable knowledge. In the paper, we present several such mechanisms including classification model building, discrete and continuous probability density function estimation, association rule generating and a new pattern based class information retrieval method.

Further knowledge exploration via further mining is our current research focus. In further mining, the discovered patterns or rules pertaining to whichever classes, categories, decisions or any instances within the data domain are at our finger tip and can be synthesized for decision and presentation used as queries for class data retrieval even outside of our original data domain.

In summary, with this approach, all patterns discovered (aware or unaware by the users) are made explicit to enhance understanding and interpretation. The system is flexible and versatile in handling any kinds of data with noise and provides the ways for a decision maker to have a broader, less biased assessment of the problem domain. Our special algorithms also make it one of the fastest association detection and inference systems in the world. The system has been used successfully in a number of complicated projects such as oil sands extraction process optimization [44].

In conclusion, knowledge discovery provides an advantage in data driven decision support. If applied successfully, it will deliver many economic benefits in sectors such as health care, financial services and customer relationship management. The discovery of new knowledge from existing data will greatly advance the state-of-the-art of the current decision making/support technologies. It has the potential to change the way we live, work and interact. With the proposed system and other pattern discovery technologies, our ability to discover knowledge has already had an impact on these sectors by deeper understanding of the problem, more efficient planning and process controls. As these technologies continue to expand, new models and new modes of solution, to decision support are anticipated.

ACKNOWLEDGMENT

The authors would like to take the opportunity to thank Pattern Discovery Software Systems Ltd. for the support of this research and allowing them to use their data mining product *discover*e*.

REFERENCES

[1] J. Donato, J. Schryver, G. Hinkel, R. Schmoyer, N. Grady Jr., and M. Leuze, "Mining multi-dimensional data for decision support," in *Proc. High Performance Computing Networking*, Amsterdam, The Netherlands, Apr. 1998, pp. 433–441.
[2] E. Blanzieri, P. Giorgini, P. Massa, and S. Recla, "Data mining, decision support and meta-learning: Toward an implicit culture architecture for kdd," in *Proc. Workshop Positions, Developments Future Directions Connection IDDM-2001*, Freiburg, Germany, Sept. 2001, pp. 9–20.
[3] R. Kohavi, D. Sommerfield, and D. J., "Data mining using mlc++: A machine learning library in c++," in *Proc. Tools AI*, Ann Arbor, MI, 1996, pp. 234–245.
[4] K. Vicente, K. Christoffersen, and Pereklita, "Supporting operator problem solving through ecological interface design," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, pp. 529–545, Apr. 1995.
[5] J. M. Tien, "Toward a systematic evaludation of large-scale information system: A framework and appliication," in *Proc. Pacific Asia Conf. Information Systems*, Brisbane, Australia, Apr. 1997, pp. 495–506.

[6] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Mateo, CA: Morgan Kaufmann, 2002.
[7] A. K. C. Wong and Y. Wang, "High order pattern discovery from discrete-valued data," *IEEE Trans. Knowledge Data Eng.*, vol. 9, pp. 877–893, Nov./Dec. 1997.
[8] T. Chau and A. K. C. Wong, "Pattern discovery by residual analysis and recursive partitioning," *IEEE Trans. Knowledge Data Eng.*, vol. 11, pp. 833–852, Nov./Dec. 1999.
[9] Y. Wang and A. K. C. Wong, "From association to classification: Inference using weight of evidence," IEEE Trans. Knowledge Data Eng., to be published.
[10] A. K. C. Wong, D. K. Y. Chiu, and W. H. Huang, "A discrete-valued clustering algorithm with applications to biomolecular data," *J. Inf. Sci.*, vol. 139, no. 1–2, pp. 97–112, 2001.
[11] J. Y. Ching, A. K. C. Wong, and K. C. C. Chan, "Class-dependent discretization for inductive learning from continuous and mixed-mode data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 641–651, July 1995.
[12] D. K. Y. Chiu and A. K. C. Wong, "Synthesis of statistical knowledge from time-dependent data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 265–271, Mar. 1991.
[13] K. C. C. Chan, A. K. C. Wong, and D. K. Y. Chiu, "Learning sequential patterns for probabilistic inductive prediction," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, pp. 1532–1547, Oct. 1994.
[14] A. K. C. Wong and T. S. Liu, "Typicality, diversity and feature patterns of an ensemble," *IEEE Trans. Comput.*, vol. C-24, pp. 158–181, 1975.
[15] A. K. C. Wong, T. S. Liu, and C. C. Wang, "Statistical analysis of residue variability in cytochrome c," *J. Mol. Bio.*, vol. 102, pp. 287–295, 1976.
[16] A. K. C. Wong and D. Ghahraman, "A statistical analysis of interdependence in character sequences," *J. Inform. Sci.*, vol. 8, pp. 173–188, 1975.
[17] A. K. C. Wong and M. A. Vogel, "Resolution-dependent information measures for image analysis," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 49–61, 1977.
[18] H. Raafat and A. K. C. Wong, "A texture information-directed region growing algorithm for image segmentation and region classification," *Comput. Vis. Graph. Image Process.*, vol. 43, no. 1, pp. 1–21, 1988.
[19] A. K. C. Wong, A. H. Vagnucci, and T. S. Liu, "Pattern detection of multivariate hormonal systems," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 33–45, 1976.
[20] A. K. C. Wong and D. C. C. Wang, "DECA: A discrete-valued data clustering algorithm," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI–1, pp. 342–349, 1979.
[21] A. C. Sanderson and A. K. C. Wong, "Pattern trajectory analysis of nonstationary multivariate data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-10, pp. 384–393, 1980.
[22] A. K. C. Wong and T. S. Liu, "A decision-directed clustering algorithm for discrete date," *IEEE Trans. Comput.*, vol. C–26, pp. 75–82, 1977.
[23] A. K. C. Wong and K. C. Chan, "Automating the knowledge acquisition process in the construction of medical expert systems," *Artif. Intell. Med.*, vol. 2, pp. 267–292, 1990.
[24] A. K. C. Wong and H. C. Shen, "Data base partitioning for data analysis," in *Proc. Int. Conf. Cybernetics Society* , Denver, CO, 1979, pp. 514–518.
[25] H. Shen, M. Kamel, and A. K. C. Wong, "Intelligent data base management systems," in *Proc. Int. Conf. Systems, Man, Cybernetics*, Bombay, Delhi, India, 1983, pp. 1131–1134.
[26] D. K. Y. Chiu and A. K. C. Wong, "Synthesizing knowledge: A cluster analysis approach using event-covering," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC–16, pp. 251–259, 1986.
[27] A. K. C. Wong and D. K. Y. Chiu, "An event-covering method for effective probabilistic inference," *Pattern Recognit.*, vol. 20, no. 2, pp. 245–255, 1987.
[28] ——, "Synthesizing statistical knowledge form incomplete mixed-mode data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 796–805, 1987.
[29] K. C. C. Chan and A. K. C. Wong, "APACS: A systems for automated pattern analysis and classification," *Comput. Intell.*, vol. 6, no. 3, pp. 119–131, 1990.
[30] R. S. Michalski and P. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI–5, pp. 396–409, Apr. 1983.
[31] R. S. Michalski and R. Stepp, "Learning from observation: Conceptual clustering," in *Machine Learning: An Artificial Intelligence Approach*, J. G. Michalski, R. S. Carbonell, and T. M. Michell, Eds. San Mateo, CA: Morgan Kaufmann, 1983.
[32] P. Smyth and R. M. Goodman, "Information theoretic approach to rule induction from database," *IEEE Trans. Knowledge Data Eng.*, vol. 4, pp. 301–316, Aug. 1992.
[33] T. Chau, "Event level pattern discovery in multivariate continuous data," Ph.D. dissertation, Univ. Waterloo, Waterloo, ON, Canada, 1997.
[34] S. J. Haberman, "The analysis of residuals in cross-classified tables," *Biometrics*, vol. 29, pp. 205–220, 1973.
[35] ——, *Analysis of Qualitative Data*. New York: Academic, 1978, vol. 1.
[36] Y. Wang, "High-order pattern discovery and analysis of discrete-valued data sets," Ph.D. dissertation, University of Waterloo, Waterloo, Ont., Canada, 1997.
[37] S. J. Haberman, *The Analysis of Frequency Data*, ser. Statistical Research Monographs. Chicago, IL: Univ. Chicago Press, 1974, vol. 4.
[38] R. O. Duda and P. H. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[39] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large database," in *Proc. ACM SIGMOD Conf. Management Data*, Washington, DC, Aug. 1993, pp. 207–216.
[40] R. Agrawal, H. Mannila, R. Srikant, H. Toivonon, and A. I. Verkamo, "Fast discovery of association rules," in *Advances In Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Cambridge, MA: MIT Press, 1996, ch. 12, pp. 307–328.
[41] B. Liu, W. Hsu, and Y. Ma, "Integration classification and association rule mining," in *Proc. 4th Int. Conf. Knowledge Discovery Data Mining*, New York, NY, Aug. 1998, pp. 80–86.
[42] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proc. Nat. Acad. Sci.*, vol. 87, no. 23, pp. 9193–9196, 1990.
[43] P. M. Murph and D. W. Aha, *UCI Repository of Machine Learning Databases*. Irvine, CA: Dept. Inform. Comput. Sci., Univ. California, 1991.
[44] D. Wallace, H. Pinto, D. Fisher, and Y. Wang, "Application of knowledge discovery techniques to improve efficiency of oil sands characterization and extraction operations," Alberta Research Council Pattern Discovery Software Systems Ltd., 2002.

Andrew K. C. Wong (SM'00) received the Ph.D. degree from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1968,

He taught at CMU for several years, and from 1976 to 2002, was a Professor of systems design engineering. He was the Founding Director, from 1980 to 2001, of the Pattern Analysis and Machine Intelligence (PAMI) Laboratory at the University of Waterloo, Waterloo, ON, Canada. In 2002, he became a Distinguished Professor Emeritus of the University of Waterloo and continues to conduct research in the PAMI Laboratory. He has authored and coauthored chapters and sections in a number of books on engineering and computer sciences, has published over 90 articles in scientific journals, 150 in conference proceedings, and holds five U.S. patents. Since 2000, he has been the Distinguished Visiting Chair Professor of the Computing Department, Hong Kong Polytechnic University. In technology transfer, he was a founder of Virtek Vision International Inc., Waterloo, a public company trading on TSE and served as its president from 1986 to 1993, and chairman from 1993 to 1997. In 1997, he founded Pattern Discovery Software Systems Ltd., Waterloo, with Y. Wong and has since served as Chairman.

Dr. Wong is the FCCP Winner of the 1991 Award of Merit.

Yang Wang (M'00) received the B. Eng. degree, in 1989, in electronic engineering, and the M. Eng. degree, in 1991, in systems engineering from Tsinghua University, Beijing, China. He received the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1997.

He is the co-founder of Pattern Discovery Software Systems Ltd., Waterloo, a software company specialized in data mining solutions, and has been the CTO of the company since 1997. His current research interests include exploratory data mining, knowledge discovery, and intelligent data analysis and their applications. He is an Adjunct Professor of systems design engineering at the University of Waterloo, where he co-supervises graduate students and conducts academic research.

Dr. Wang is a member of the IEEE Computer Society and ACM.