Pattern Discovery as Event Association

Andrew K. C. Wong
University of Waterloo, Canada

Yang Wang
Pattern Discovery Technology, Canada

Gary C. L. Li
University of Waterloo, Canada

INTRODUCTION

A basic task of machine learning and data mining is to automatically uncover **patterns** that reflect regularities in a data set. When dealing with a large database, especially when domain knowledge is not available or very weak, this can be a challenging task. The purpose of **pattern discovery** is to find non-random relations among events from data sets. For example, the "exclusive OR" (XOR) problem concerns 3 binary variables, A, B and C=A⊗B, i.e. C is true when either A or B, but not both, is true. Suppose not knowing that it is the XOR problem, we would like to check whether or not the occurrence of the compound event [A=T, B=T, C=F] is just a random happening. If we could estimate its frequency of occurrences under the random assumption, then we know that it is not random if the observed frequency deviates significantly from that assumption. We refer to such a compound event as an event association pattern, or simply a **pattern**, if its frequency of occurrences significantly deviates from the default random assumption in the statistical sense. For instance, suppose that an XOR database contains 1000 samples and each primary event (e.g. [A=T]) occurs 500 times. The expected frequency of occurrences of the compound event [A=T, B=T, C=F] under the independence assumption is $0.5\times0.5\times0.5\times1000 = 125$. Suppose that its observed frequency is 250, we would like to see whether or not the difference between the observed and expected frequencies (i.e. 250 – 125) is significant enough to indicate that the compound event is not a random happening.

In statistics, to test the correlation between random variables, **contingency table** with chi-squared statistic (Mills, 1955) is widely used. Instead of investigating variable correlations, pattern discovery shifts the traditional correlation analysis in statistics at the variable level to association analysis at the event level, offering an effective method to detect statistical association among events.

In the early 90's, this approach was established for second order event associations (Chan & Wong, 1990). A higher order **pattern discovery** algorithm was devised in the mid 90's for discrete-valued data sets (Wong & Yang, 1997). In our methods, patterns inherent in data are defined as statistically significant associations of two or more primary events of different attributes if they pass a statistical test for deviation significance based on **residual analysis**. The discovered high order patterns can then be used for classification (Wang & Wong, 2003). With continuous data, events 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 (patterns) in the form of high dimension intervals from which probability density can be estimated by Gaussian kernel fit (Chau & Wong, 1999). Classification can then be achieved using Bayesian classifiers. For data with a mixture of discrete and continuous data (Wong & Yang, 2003), the latter is categorized based on a global optimization discretization algorithm (Liu, Wong & Yang, 2004). As demonstrated in numerous real-world and commercial applications (Yang, 2002), pattern discovery is an ideal tool to uncover subtle and useful patterns in a database.

In pattern discovery, three open problems are addressed. The first concerns learning where noise and uncertainty are present. In our method, noise is taken as inconsistent samples against statistically significant patterns. Missing attribute values are also considered as noise. Using a standard statistical **hypothesis testing** to confirm statistical patterns from the candidates, this method is a less ad hoc approach to discover patterns than most of its contemporaries. The second problem

concerns the detection of polythetic patterns without relying on exhaustive search. Efficient systems for detecting monothetic patterns between two attributes exist (e.g. Chan & Wong, 1990). However, for detecting polythetic patterns, an exhaustive search is required (Han, 2001). In many problem domains, polythetic assessments of feature combinations (or higher order relationship detection) are imperative for robust learning. Our method resolves this problem by directly constructing polythetic concepts while screening out non-informative pattern candidates, using statistics-based heuristics in the discovery process. The third problem concerns the representation of the detected patterns. Traditionally, if-then rules and graphs, including networks and trees, are the most popular ones. However, they have shortcomings when dealing with multilevel and multiple order patterns due to the non-exhaustive and unpredictable hierarchical nature of the inherent patterns. We adopt **attributed hypergraph** (AHG) (Wang & Wong, 1996) as the representation of the detected patterns. It is a data structure general enough to encode information at many levels of abstraction, yet simple enough to quantify the information content of its organized structure. It is able to encode both the qualitative and the quantitative characteristics and relations inherent in the data set.

BACKGROUND

In the ordinary sense, "discovering regularities" from a system or a data set implies partitioning the observed instances into classes based on similarity. Michalski and Stepp (1983) pointed out that the traditional distance-based statistical clustering techniques make no distinction among relevant, less relevant and irrelevant attributes nor do they render conceptual description of the clusters with human input. They proposed CLUSTER/2 as a conceptual clustering algorithm in a noise-free environment. It is effective for small data sets containing no noise yet computationally expensive for a large data set even with its Hierarchy-building Module. To deal with noise, COBWEB was introduced by Fisher (1987). However, the concept tree generated by COBWEB might be very large. For deterministic pattern discovery problems such as the MONK, COBWEB does not work well when compared with other AI and connectionist approaches (Han, 2001).

The Bayesian methods provide a framework for reasoning with partial beliefs under uncertainty. To perform inferences, they need to estimate large matrices of probabilities for the network during training (Pearl, 1988). When going to high-order cases, the contingency table introduces a heavy computation load.

Agrawal and Srikant (1994) proposed association rule mining to detect relationship among items in transactional database. It is well-suited to applications such as market basket analysis but not applicable in some other applications such as capturing correlations between items where association rules may be misleading (Han, 2001). Hence, Brin, Motwani and Silverstein (1997) proposed to detect correlation rules from the contingency table. However, correlation rule mining is not accurate when the contingency table is sparse or larger than 2×2 (Han, 2001) since it is designed for testing the correlation of two random variables.

Pattern discovery shifts the statistical test from the entire **contingency table** to the individual cells in the table. A **hypothesis test** at each individual cell is formalized by **residual analysis**. Therefore, it handles sparse and high dimensional contingency table much more effectively.

Recent development in pattern discovery methodologies includes building classifiers based on the discovered patterns. A typical example is Liu, Hsu and Ma 's CBA (1998) that uses association rules to classify data. More recent works include HWPR (Wang & Wong, 2003) and DeEPs (Li et. al., 2004). HWPR employs event associations as classification rules whereas DeEPs uses emerging patterns, a variation of association rules, to classify data. A detailed evaluation and comparison of these methods can be found in (Sun et. al., 2006).

MAIN FOCUS

Event Associations

Consider a data set D containing M data samples. Every sample is described by N attributes, each of which can assume values from its own finite discrete alphabet. Let $\mathbf{X} = \{ X_1, \ldots, X_N \}$ represent this attribute set. Then, each attribute, X_i, $1 \le i \le N$, can be seen as a random variable taking on values from its alphabet $\alpha_i = \{\alpha_i^1, ..., \alpha_i^{m_i}\}$, where m_i is the cardinality of the alphabet of the ith attribute. Thus, a realization of $\mathbf{X}$ can be denoted by $\mathbf{x} = \{ x_1, \ldots, x_N \}$, where x_i can assume