

Complementary Analysis of High-Order Association Patterns and Classification

Thomas W.H. Lui and David K.Y. Chiu

Department of Computing and Information Science
University of Guelph, Guelph, Canada

Abstract

To facilitate more meaningful interpretation considering the internal interdependency relationships between data values, a new form of high-order (multiple-valued) pattern known as Nested High-Order Pattern (or NHOP) is recently proposed. This pattern satisfies a consistent statistical criterion when the pattern is iteratively extracted. The general form of High-Order Pattern (HOP), that NHOP is a subtype, is a set of multiple associated values (identified as variable outcomes) extracted from a random N -tuple. The pattern is detected by statistical testing if the occurrence is significantly deviated from the expected according to a prior model or null hypothesis. Here we extend our work of NHOP to the classification task. The rationale is that, meaningful association patterns, involving multiple values jointly and at the same time predict classification, can reinforce the underlying regularity, and hence provide a better understanding of the data domain. In this paper, we propose a Classification method based on the *Nested High-Order Patterns* (C-NHOP). In evaluating our method using 26 UCI benchmark datasets, the experiments show a highly competitive and interpretable result.

1. Introduction

From the analysis of data, mining various forms of association patterns that also predict the classification groupings often provides more reliable information about the underlying physical properties, and hence reflects a better understanding of the data domain. Association mining aims to discover descriptive patterns relating different values from database. Classification aims to build a classifier from a set of training data whose classes are known and useful in classifying new instances into one of the pre-defined classes. Recently, associative classification is proposed that often considers multiple values as a pattern in the design of the classification rules. This paper, following the same spirit, aims to produce a system in a more tightly-coupled manner that the extracted data patterns and classification can be jointly interpreted.

A desirable characteristic in the process of classification is that the classifier can be interpreted from the description of the extracted patterns, which are understandable by the human users. Empirical studies (Liu, Hsu, and Ma 1998;

Li, Han, and Pei 2001) show that associative classification systems can achieve competitive classification performance, together with standard AI methods such as decision trees (Quinlan 1993) and rule induction (Cohen 1995), and are as easily interpretable.

Some associative classification systems make use of association rule mining algorithms which aim to generate classification rules that satisfy the minimum support and minimum confidence (Liu, Hsu, and Ma 1998; Li, Han, and Pei 2001; Thabtah et al. 2004). However, the setting of minimum support and confidence can be difficult. For example, when minimal support is set too high, many useful rules for classification may be missed. In addition, a detected rule with high support usually involves only a few values and may not be detailed enough to classify precisely. On the other hand, when minimal support is set too low, the computational cost is much higher as the search space increases enormously. Furthermore, since the extracted rules with low support usually have low occurrence, they could overfit the data and thus may ignore the underlying properties of the domain. As a result, the extracted rule set may contain many irrelevant rules in the classification process. Pruning these irrelevant rules can improve the classification accuracy, but may not address the actual issue of capturing the underlying properties of the domain.

Similarly, when the minimum confidence threshold is set too high, the rules may involve too many values that can overfit the data. However, lowering the confidence threshold will also degrade the classification accuracy. As a result, the optimality often depends on the user who typically changes parameters while running the algorithm.

In this paper, we propose a novel association pattern based on Nested High-Order Pattern (NHOP) (Lui and Chiu 2007) which depends on statistical criteria that apply iteratively to the extracted pattern components. Intuitively, a high-order pattern (HOP) is a complex pattern that forms and reflects a complex interactive relationship among its composed values (Wong and Wang 1997). It is based on sampling outcomes of a random N -tuple, rather than frequent itemsets. Thus, it draws on the probabilistic properties by analyzing relationships in random variables. The high-order relationships are extracted from the outcome values to construct statistically significant association patterns that are jointly occurring. Thus the patterns' reliability is related directly to the statistical evaluation.

To explain the proposed high-order pattern more clearly, a high-order pattern is a set of jointly occurring interdependent (or associated) values that are statistically significant. For example, given sample outcomes of $(X_1, X_2, \dots, X_{20})$, a high-order pattern can be $(X_3=A, X_7=B, X_{12}=D)$. The order of a high-order pattern is defined as the number of interdependent values detected. Based on analysis of statistical significance, a high-order pattern can be defined as a set of values whose observed occurrences is statistically deviated from the prior model assumption, such as the independence assumption (Wong and Wang 1997; Chiu, Wong, and Cheung 1991).

The significance of a high-order pattern is evaluated using a statistical test (Wong and Wang 1997). An overview of the test can be described as follows. First, the observed frequency of a pattern candidate is obtained from the data. The expected frequency of the pattern candidate can then be estimated based on the presumed null hypothesis which normally assumes that all the values are mutually independent. It is rejected if the actual observed frequency of the pattern candidate statistically deviates from the expected frequency. In this case, the pattern candidate is accepted as a significant pattern.

The null hypothesis assumption of independence has been used extensively in many pattern discovery problems. When the independence hypothesis is rejected, the alternative hypothesis is accepted. The alternative hypothesis suggests that not all of the values in the pattern are independent. In other words, it suggests that at least some values are interdependent, without specifying which values are interdependent when more than two values are involved. That is, a comprehensive nature of the internal interdependency relationships can be identified and interpreted later.

In general, an n -order pattern ($2 < n \leq N$), extracted from the dataset of a random N -tuple, may still contain independent disjointed components. For example, in a 4-order pattern (A, B, C, D) , the value subset $\{A, B\}$ may be independent of the other subset $\{C, D\}$, even though the whole set $\{A, B, C, D\}$ deviates from its null hypothesis of the prior model. To facilitate interpretation and identify patterns that are important considering their internal interdependency relationships, a new form of restricted subtype known as Nested High-Order Patterns (or NHOP) is recently proposed (Lui and Chiu 2007). In (Lui and Chiu 2007), the analysis using aligned protein sequences shows that NHOP is useful in identifying key multiple sites at the 3-dimensional protein molecular core. Here, we extend the work on NHOP and apply it for the classification task that is reinforced from the extracted patterns.

By NHOP, we refer to a pattern such that its components are detected to be interdependent when the pattern is expanded iteratively. This pattern thus satisfies the statistical criterion of interdependency consistently for all its components. An illustration of a NHOP which consists of four components, denoted as $((A, B), C), D$, is shown in Figure 1.

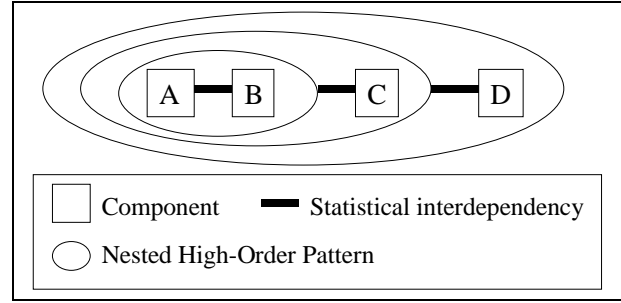


Figure 1. Illustration of a Nested High-Order Pattern, denoted as $((A, B), C), D$.

To construct a NHOP, an interdependent component is added iteratively to a previous lower order pattern. The motivation is that with interdependent components added iteratively, we avoid detecting patterns that are independently disjoint. At each expansion of pattern, NHOP composes components that are not simply due to the chance model as specified in the prior assumption. In other words, we avoid adding irrelevant value components into the constructed high-order pattern. NHOP is desirable for classification purpose because by avoiding components that are due to the chance model, NHOP can better represent the true underlying regularities in the data. It can then more likely be generalized to unknown testing data.

A NHOP pattern can be defined as monothetic or polythetic pattern. A monothetic pattern expands from a single value (Lui and Chiu 2007). A NHOP pattern that is polythetic can include more complex expansion with more values. A well-known example of polythetic pattern is the XOR pattern, where all the attribute values are independent to each other, which requires for more complex pattern relationships. In this paper, we generalize our algorithm to allow expansion of a pattern component to include multiple values when a single value pattern expansion is inadequate. In our experiments, we found that using one or two values, rather than a large number is already adequate for most benchmark datasets.

The rest of the paper is organized as follows. Section 2 describes the notation and basic concepts used in this paper. Section 3 presents in depth our proposed method relating to classification which we called C-NHOP (Classification based on NHOP). Section 4 presents the extensive experimental results.

2. Notations and Basic Concepts

2.1 Notations and Data Representation

Our data are represented as sampling outcomes of a random N -tuple. Alternatively, we can describe them as an ensemble of a dataset represented as a relation in a relational database. The relation can then be represented as

$X=(X_1, X_2, \dots, X_N)$ with N variables. An instance of X is a realization that can be denoted as $x=(x_1, x_2, \dots, x_N)$. Each x_i ($1 \leq i \leq N$) can take up an attribute value denoted as $x_i=a_{ip}$. An attribute value $x_i=a_{ip}$ is a value taken from the attribute value set, $\Gamma_i=\{a_{ip}|p=1, 2, \dots, L_i\}$ where L_i is the number of possible values for the attribute X_i . A value a_{ip} in Γ_i is then defined as a primitive. Let E be a subset of X . E then has at least one variable, but less than or equal to N variables. The realization of E is denoted as e . A compound pattern event (or just compound event) associated with the variable set E is a set of attribute values instantiated by a realization e . The compound event can be represented by $e=\{X_j=a_{jq} | X_j \in E, q \in 1, 2, \dots, L_j\}$. The order of the compound event is $|E|$, which can be conveniently denoted as n ($\leq N$).

2.2 Evaluation of Statistical Significance

The expansion of NHOP is based on evaluating the statistical interdependency between its components, denoted, say, $E_1=e_1$ and $E_2=e_2$. Statistically significant patterns are considered in order to differentiate from those that are due to chance. To evaluate the statistical significance, a measure between the outcome values corresponding to their variables is used. This is considered to identify those values that are interdependent even though other values of the same variables are not.

We use the following method for evaluating the statistical interdependency between two compound events $E_1=e_1$ and $E_2=e_2$, using a statistics $z(e_1, e_2)$. The notation E_1 and E_2 refers to the joint variables composed of the corresponding variable subset that makes up the event of e_1 and e_2 respectively. E_1 and E_2 consists of distinct variable sets (i.e. $E_1 \cap E_2 = \emptyset$). The standardized residual $z(e_1, e_2)$ (Wong and Wang 1997) is defined as:

$$z(e_1, e_2) = \frac{obs(e_1, e_2) - exp(e_1, e_2)}{\sqrt{exp(e_1, e_2)}} \quad (1)$$

where $obs(e_1, e_2)$ is the observed frequency and $exp(e_1, e_2)$ is the expected frequency for (e_1, e_2) respectively. The expected frequency of (e_1, e_2) is calculated from the independence assumption where the null hypothesis assumes the two events considered to be independent to each other. If the deviation between the observed and expected frequency of the compound events is sufficiently large, then the pattern is found to be previously unexpected. Thus, a new pattern not previously known can be evaluated. It is known that the standardized residual is normally distributed only when the asymptotic variance of $z(e_1, e_2)$ is close to one, otherwise, standardized residual has to be adjusted by its variance using the formula below. The adjusted residual (Wong and Wang 1997) is expressed as:

$$d(e_1, e_2) = \frac{z(e_1, e_2)}{\sqrt{v(e_1, e_2)}} \quad (2)$$

where $v(e_1, e_2)$ estimates the maximum likelihood of the variance of $z(e_1, e_2)$. It is expressed as:

$$v(e_1, e_2) = 1 - P(E_1 = e_1)P(E_2 = e_2). \quad (3)$$

The adjusted residual $d(e_1, e_2)$ has an asymptotic normal distribution. Thus the significance level can be chosen to be 95%, or any other acceptable level. The test for statistical significance is based on the following inequality $d(e_1, e_2) \geq N_\alpha$, where N_α is the tabulated threshold with α being the confidence level.

2.3 Measure of Interdependency of NHOP

The amount of interdependency between two events is estimated using a measure of mutual information. The mutual information between the two compound events $E_1=e_1$ and $E_2=e_2$ is calculated as:

$$MI(E_1 = e_1, E_2 = e_2) = \log \frac{P(E_1 = e_1, E_2 = e_2)}{P(E_1 = e_1)P(E_2 = e_2)} \quad (4)$$

To calculate the overall relevance of a detected NHOP based on combining the interdependency among the patterns expanded in all the previous iterations, we use a weighted product of the mutual information (Ω). It is defined as follows. Consider an n -component NHOP, denoted as $e=((e_1, e_2), e_3, \dots, e_n)$, where e_j is the j^{th} component append to the $(j-1)$ - component NHOP. The measure Ω of e , which indicates the total internal interdependency, is defined as:

$$\begin{aligned} \Omega(e = (((e_1, e_2), e_3), \dots), e_n)) \\ = P(e) \times [MI(e_1, e_2) \times MI((e_1, e_2), e_3) \times \dots \times \\ MI((e_1, e_2, \dots, e_{n-1}), e_n)] \end{aligned} \quad (5)$$

This measure calculates the summed total interdependency of all components identified. Notice that, Ω is high when the calculated mutual information for all the added components in the pattern is consistently high. Furthermore, the j^{th} component is added only when it is statistically significantly interdependent with the $(j-1)$ -component NHOP.

3. Classification based on NHOP

3.1 Overview

The classification algorithm proposed here is designed as a complementary consideration between the detected NHOP with the classification pattern. Note that the two types of patterns are both interpretable in understanding the data domain. Thus, their reliability reinforces each other. We call the algorithm that extracts patterns with these

properties C-NHOP algorithm (or Classification based on Nested **H**igh-**O**rd **P**attern Algorithm). C-NHOP consists of three phases including: 1) extracting patterns from the training data; 2) constructing the classifier; and 3) classifying new instances in the testing data.

3.2 Pattern Extraction

The algorithm for extracting the NHOP patterns from the training data is presented as follows. The patterns are generated from low to high order by using pattern expansions iteratively. Initially, the number of values for pattern expansion is set to one. In the first iteration, the 2-order patterns in the form of $(C=c_i, X_j=a_{jq})$ are detected using the adjusted residual for each class $C=c_i$, where the class label, $C=c_i$, is associated to one value only.

To have a more efficient search for higher order patterns, only a subset of detected patterns in current iteration is selected using a database coverage method (Liu, Hsu, and Ma, 1998). The basic idea of the method is to select a set of high ranking patterns to "cover" as many instances as possible in the training dataset.

An overview of the database coverage method is described as follows. The detected patterns are ranked using the objective function, as in Equation (5). The highest ranked pattern that can correctly cover at least one instance in the training dataset is first selected. A pattern covers an instance if the pattern's values are found in the instance. Once a pattern is selected, the covered instances are removed from the dataset. The next highest ranked pattern that covers the data is then considered. The pattern selection continues until as many instances as possible are covered.

If some instances in the data cannot be covered by these patterns, we increase the size of pattern expansion by one. The objective is to cover the remaining uncovered instances by patterns with multiple values expansion. We consider single value expansion first before using expansion with multiple values. When there are instances in the data cannot be covered by NHOP with single value expansion, a more complex NHOP based on multiple value expansion will be used with additional computational cost.

When all the instances are covered or when the expansion increases up to a predefined number (m), the selected patterns will be promoted for higher order search in next iteration. If no pattern selected in current iteration, the search terminates. In the next iteration, the expansion is reset to single value and instances are reset to be "uncovered".

Notice that the pattern is evaluated based on the whole set of training data rather than the remaining dataset as in many other methods at each search iteration. The advantage is that the detected patterns are reflective of the domain rather than biased for the classification task only.

3.3 Classifier Construction

To construct a classifier is to extract a subset of high quality patterns that can represent the training dataset. To achieve this goal, filtering out patterns that confuse between the classes or overfit the data is required.

Specific patterns with a large number of values tend to have higher accuracy. They, however, tend to have lower occurrence in the training data. With low occurrence, these patterns may not generalize well to the testing data. This is the problem of overfitting. In our method, we select patterns for the classifier using a pattern ranking scheme based on the order of the pattern and the degree of the patterns in satisfying the extracting criterion. A pattern ranking is defined as follows. Defining $confidence(c_i, e) = P(c_i, e)/P(e)$; given two patterns (c_i, e_1) and (c_i, e_2) , (c_i, e_1) is said to have a higher ranking than (c_i, e_2) , if:

- (1) $confidence(c_i, e_1) > confidence(c_i, e_2)$;
- (2) $confidence(c_i, e_1) = confidence(c_i, e_2)$
but order of $(c_i, e_1) < \text{order of } (c_i, e_2)$;
- (3) $confidence(c_i, e_1) = confidence(c_i, e_2)$,
order of $(c_i, e_1) = \text{order of } (c_i, e_2)$,
but $\Omega(c_i, e_1) > \Omega(c_i, e_2)$.

The first criterion based on confidence has been used by most associative classifiers such as (Liu, Hsu, and Ma 1998; Li, Han, and Pei 2001). Pattern with higher confidence may introduce less error in the classification phase. The second criterion said that when two patterns have the same confidence, the one with lower order is preferred. The rationale is that lower order pattern, which is more general, is less likely to overfit the data. When confidence and order of two patterns are the same, the third criterion chooses the one with a higher Ω value.

The algorithm for constructing the classifier is presented as follow. First, selected patterns in previous phase are sorted according to the ranking scheme. A refined set of patterns is then selected based on the database coverage method.

3.4 Classifying new instance

In classifying a new instance of unknown class in the testing phase, the class label of the first NHOP such that it is observed in the instance will be used. If there is no matched NHOP pattern, lower order patterns of NHOP with fewer components will be considered. If there is still no matched pattern, the majority class with the most number of samples will be used.

4. Experimental Results

4.1 Description of Data

The following experiments are evaluated based on 26 benchmark datasets in UCI Machine Learning Repository, which can be obtained from (Liu, Hsu, and Ma 1998). As

shown in Table 1 (column 1 to 4), the datasets differ in number of attributes (column 2), number of classes (column 3), as well as sample size (column 4). We use the same method as in (Liu, Hsu, and Ma 1998) to discrete the continuous attributes.

Each dataset is evaluated using 10-fold cross-validation. To ensure the class frequency distributions are more suitable for the cross-validation, each obtained dataset has been shuffled using the shuffle utility in C4.5 (Quinlan 1993). In cross-validation, a dataset is divided into 10 folds. Each fold is in turn used as the testing dataset while the remaining data is used as the training dataset. The average accuracy of all 10 trials is reported.

4.2 Experiment 1: Compare classification rates considering different sizes of pattern expansion

In this experiment, we compare the classification rates using different maximum number of expanded values allowed. This parameter is denoted as m .

The result using maximum size of expansion, $m=1$, $m=2$, and $m=3$ is shown in Table 1 (columns 5 to 7). The overall average classification rates of using $m=1$, 2, 3 are 85.0%, 86.6%, and 86.7% respectively. The rates using $m=1$ is lower than using $m=2$ or $m=3$ as expected. Since only monothetic patterns can be discovered when using $m=1$, the classification rate is poor in some cases where the dataset consists of many polythetic patterns.

For example, in the Tic-Tac-Toe dataset, the rate using $m=1$ is very poor, only with 70.3%. Using $m=2$, the rate can be immediately increased to 98.5%. Obviously, monothetic patterns cannot model the data well. Similarly, the rate using naïve bayesian method (NB) is also poor for this dataset as expected. As shown in Table 1 (column 11), the rate is only 69.9%.

Using $m=2$ gives higher average rates than using $m=1$ in general. By increasing m from one to two, 10 datasets get higher rates, 11 datasets get the same rates, and 5 datasets get lower rates. Using $m=3$ gives similar average rate as using $m=2$. By increasing m from two to three, 7 datasets get higher rates, 15 datasets get the same rates, and 4 datasets get lower rates. The rates based on $m=3$ is not significantly higher than that of $m=2$ using the t-test ($P=0.12$).

In summary, our proposed algorithm based on extracted NHOP patterns that allow two values expansion can be a better model than that based on single value expansion. However, there is no significant improvement when three values expansion is allowed.

In this experiment, the confidence level used for the adjusted residual statistical test is 90%. Same conclusions can be drawn in using 80% or 95% confidence level. Hence the result is not sensitive to the selected level.

Table 1. Classification rates on 26 benchmark datasets.

Dataset	Att	Cls	Size	Experiment 1			Experiment 2				
				m=1	m=2	m=3	CBA	CPAR	C4.5	NB	C-NHOP
Anneal	38	6	898	99.2	99.3	99.3	96.4	99.3	94.8	97.3	99.3
Australian	14	2	690	86.2	86.8	86.4	86.6	86.1	84.7	86.0	86.8
Auto	25	7	205	82.9	85.2	85.2	72.8	82.9	80.1	67.9	85.2
Breast-w	10	2	699	96.1	95.9	95.9	95.8	96.7	95.0	97.6	95.9
Cleve	13	2	303	82.7	82.7	83.3	83.3	83.0	78.2	82.9	82.7
Crx	15	2	690	86.7	86.4	86.7	85.9	84.8	84.9	85.4	86.4
Diabetes	8	2	768	77.5	77.5	77.5	74.7	76.6	74.2	75.6	77.5
German	20	2	1000	71.6	71.5	71.5	73.5	72.8	72.3	75.4	71.5
Glass	9	7	214	74.3	76.2	76.2	72.6	77.6	68.7	70.6	76.2
Heart	13	2	270	82.6	82.6	82.6	81.5	83.0	80.8	81.9	82.6
Hepatitis	19	2	155	83.8	86.9	86.3	84.9	83.8	80.6	85.0	86.9
Horse	22	2	368	85.9	84.6	85.7	81.3	84.9	82.6	79.4	84.6
Hypo	25	2	3163	97.1	97.1	96.9	98.3	98.3	99.2	98.5	97.1
Ionosphere	34	2	351	94.0	94.0	94.0	91.8	94.0	90.0	88.1	94.0
Iris	4	3	150	93.3	93.3	93.3	92.9	94.7	95.3	94.0	93.3
Labor	16	2	57	96.7	96.7	96.7	83.0	95.0	79.3	86.0	96.7
Led7	7	10	3200	71.8	71.8	71.8	72.2	73.5	73.5	73.3	71.8
Lymph	18	4	148	80.7	81.3	84.0	80.4	83.3	73.5	75.6	81.3
Pima	8	2	768	77.3	77.3	77.3	72.4	77.0	75.5	75.5	77.3
Sick	29	2	2800	96.4	96.4	96.5	97.3	96.7	98.5	96.1	96.4
Sonar	60	2	208	82.4	83.3	83.8	78.3	82.9	70.2	77.0	83.3
Tic-tac-toe	9	2	958	70.3	98.5	98.5	100	100	99.4	69.9	98.5
Vehicle	18	4	846	68.4	69.9	69.5	68.7	71.1	72.6	59.9	69.9
Waveform	21	3	4500	79.8	78.9	79.1	79.4	80.9	78.1	80.7	78.9
Wine	13	3	178	99.4	99.4	99.4	91.6	98.3	92.7	90.5	99.4
Zoo	16	7	101	93.0	97.0	97.0	94.6	95.0	92.2	86.3	97.0
Average accuracy				85.0	86.6	86.7	84.2	86.6	83.3	82.2	86.6

4.3 Experiment 2: Compare C-NHOP with other classifiers

In this experiment, we compare our algorithm with two associative classifiers including CBA (Liu, Hsu, and Ma 1998) and CPAR (Yin and Han 2003); a decision tree method, C4.5 (Quinlan 1998); and naïve bayesian (NB) method. The results for CBA, C4.5, and NB are taken from (Liu, Ma, and Wong 2000). We run CPAR downloaded from (Yin and Han 2003) with the default settings.

From Table 1 (column 8 to 12), the average rates for CBA, CPAR, C4.5, NB, and C-NHOP (using $m=2$) are 84.2%, 86.6%, 83.3%, 82.2%, and 86.6% respectively. Among the 26 datasets, the number of won-loss-tied datasets of C-NHOP against CBA, CPAR, C4.5, and NB are 19-0-7, 10-1-14, 19-0-7 and 19-0-7, respectively. When all the five classifiers are compared, CBA achieves the best accuracy in 2 datasets; CPAR wins in 9 datasets; C4.5 wins in 5; NB wins in 2; and C-NHOP wins in 12. In comparing the results, C-NHOP is significantly higher than CBA, C4.5, and NB when compared using the t-test with confidence level of $P<0.01$. The difference between C-NHOP's rates and CPAR's rates is not statistically significant. Here, we conclude that C-NHOP is competitive to these classifiers using these benchmark datasets.

Not only NHOP patterns can be applied to achieve high accuracy, they are also immediately interpretable. Here, we use the results from the Zoo dataset to demonstrate the interpretability of the NHOP patterns. The Zoo dataset is

Table 2. Top ranked NHOP animal class patterns (most frequent in the 10-fold cross-validation). [Note that the patterns closely fit our general understanding of the classes.]

Order	Patterns
2-order	(class=mammal, milk=yes)
2-order	(class=birds, feathers=yes)
7-order	(((((class=reptiles, milk=no & toothed=yes), predator=yes & domestic=no), fins=no), breathes=yes)
3-order	((class=fishes, fins=yes), breathes=no)
5-order	(((((class=amphibians, aquatic=yes), legs=4), eggs=yes & toothed=yes)
6-order	(((((class=insects, legs=6), backbone=no), tail=no), toothed=no), airborne=yes)
4-order	(((((class=others(crest-bearing animals/worms), backbone=no), tail=no), breathes=no)

chosen because the extracted patterns can easily be verified. The Zoo dataset contains 101 samples with 16 attributes and a class attribute. Each sample represents an animal with its characteristics. The 16 attributes are hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic, and size. There are 7 classes of animals including 41 mammals, 20 birds, 5 reptiles, 13 fishes, 4 amphibians, 8 insects, and 10 others (including worms and crest-bearing animals).

Table 2 lists the most frequent top ranked NHOP for each class in the 10-fold cross-validation. All these NHOP are found meaningful. As shown in Table 2, not only values in the NHOP patterns are shown explicitly, the statistical associations between the pattern's components are also displayed. For example, consider the 5-order NHOP pattern (((class=amphibians, aquatic=yes), legs=4), eggs=yes & toothed=yes) in Table 2, the first component associated with the amphibian class is the aquatic characteristic. The next associated characteristic is "legs=4". Finally, the last associated component, which based on two values expansion, consists of two characteristics including egg-laying and toothed.

5. Conclusion

In this paper, we propose a novel classification system which we called C-NHOP (Classification based on Nested High-Order Pattern). It is designed to be highly interpretable, closely related to patterns observed from the dataset, and associated with their class labels such that classification predictability is meaningful. In addition, it also has the following characteristics:

- (1) It constructs NHOP patterns by adding statistically interdependent components one at a time which aims to eliminate associations due to random variations.
- (2) Extending from (Lui and Chiu 2007), we also identify polythetic patterns when monothetic patterns with single value expansion is inadequate.

(3) Generating patterns from a small to larger size, its predictive relationship is refined from a general to a specific description.

(4) It uses the whole dataset in evaluating patterns at each iteration, rather than using the remaining unclassified set, thus generated patterns are based on complete information of the observed data and is more interpretable.

References

Chiu, D.K.Y., A.K.C. Wong, Cheung, B.. 1991. Information discovery through hierarchical maximum entropy discretization and synthesis. *Knowledge Discovery in Databases*, eds. G. Piatetsky-Shapiro and W.J. Frawley, 125-140, MIT/AAAI Press.

Cohen, W. 1995. Fast effective rule induction. In *ICML'95*, pp.115-123, Tahoe City, CA, July.

Li, W., Han, J., Pei, J. 2001. CMAR: Accurate and efficient classification based on multiple-class association rule. In *Proceedings of the International Conference on Data Mining (ICDM'01)*, San Jose, CA, pp.369-376.

Liu, B., Hsu, W., Ma, Y. 1998. Integrating classification and association rule mining. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. New York, NY: AAAI Press, pp.80-86.
<<http://www.comp.nus.edu.sg/~dm2>>

Liu, B., Ma, Y., Wong, C.K. 2000. Improving an association rule based classifier. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*. Lyon, France, pp.504-509.

Lui, W.H.T., Chiu, D.K.Y. 2007. Discovering maximized progressive high-order patterns in biosequences. In *Proceedings of the 10th Joint Conference on Information Sciences*. Utah, USA, pp.110-115.

Thabtah, F., Cowling, P., Peng, Y. 2004. MMAC: A new multi-class, multi-label associative classification approach. In *Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04)*, Brighton, UK, pp.217-224.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Wong, A.K.C., Wang, Y. 1997. High-order pattern discovery from discrete-valued data. *IEEE Trans. on Knowledge & Data Engineering*. 9(6). pp.877-893.

Yin, X., Han, J. 2003. CPAR: Classification based on predictive association rules. *Proc. SIAM Int. Conf. on Data Mining (SDM'03)*, San Fransisco, CA, pp.331-335.
<<http://illimine.cs.uiuc.edu/download/index.php>>