TELKOMNIKA, Vol. 11, No. 8, August 2013, pp. 4556~4564
e-ISSN: 2087-278X

A Method of Discovering Tolerance Markov Blanket based on Completely Dependent Unknown Components

[1]Hongzhou He*, [2]Mingtian Zhou
[1]College of Mathematics & Computer Science, Mianyang Normal University, 621000, China
[2]School of Computer Science and Engineering, University of Electronic Science & Technology of China, Chengdu 611731, China
*Corresponding author, e-mail: zmoonmoonlhm@yahoo.com.cn

Abstract

A novel tolerance feature subset selection method from incomplete data set, denoted by MaxG-IIAMB, is proposed to pick out the Markov-boundary (MB), the minimal subset of features, of target variable but without making any assumption about the unknown component distribution. The classification experimental results of risk factors observed in a sample of 1841 employees of a Czech car factory demonstrate the practicability and superiority of our method over the classical expectation-maximization (EM) and available case technique (ACA).

Keywords: *feature subset selection (FSS), markov boundary (MB), probabilistic classification, unknown component distribution.*

1. Introduction

Feature subset selection (FSS), the precondition of supervised probabilistic classification, aims to select the necessarily features for many under-sample and high-dimension samples such as risk factor analysis [1] and information retrieval [2]. The Markov-boundary (MB) [3-4] is one of the most known and efficient solutions for FSS. Markov-boundary of a class variable *C*, MB(*C*), is defined as any minimal subset of full feature set **F** such that *C* is conditionally independent of the rest of **F** given MB(*C*). In information science field, there has been a growing interest in picking out the MB automatically from sample data set. The constraint-based (CB) algorithms, such as PCMB [5], IAMB [6] or its variants: Fast-IAMB [7] and Inter-IAMB [8] have been proposed to address the problem. These methods systematically check the data for independence relations and use those relationships to infer necessary features in the MB. When no unknown component in instances of the data set, they can estimate efficiently the MB (*C*). Unfortunately, when the database is not completed, i.e., some components are reported as unknown, these methods do not function any more.

The Gibbs sampling [9] and the EM algorithm [10-12] are known solutions to handle incomplete data sets, but they assume implicitly that the distribution of unknown component depends only on the observed values in the data set. Under this assumption, the unknown values can be inferred from the available data. However, this assumption does not hold and it is hard to test in practice on the one hand, on the other hand, the decrease in accuracy may be severe with EM-based methods once the assumption is not held.

Based on Ramoni's ideas in [13], we give a novel tolerance test method, denoted by MaxG-IIAMB, for discovering the Markov-boundary (MB) of target variable in an incomplete data set. Without any assumption about the unknown components distribution, MaxG-IIAMB gets over the shortcoming in [9-12] by maximizing the conditional dependence measure over all possible ways to restore the unknown components. According to the idea in [13], when no information about the distribution of unknown component is available, an incomplete data set contains the set of all possible estimates. In this paper, we provide the detail description of these constraints in obtaining the completion of an incomplete data set.

The remainder of this paper describes our approach. Section 2 gives the formal description of involved problem and reviews the background and motivation of the research.

Received March 15, 2013; Revised May 16, 2013; Accepted May 27, 2013

Section 3 describes the theoretical framework of the method, while Section 4 applies our method to conduct experiments on synthetic incomplete data sets in [1] and compare its performance with the other two methods: available case analysis (ACA) and maximum likelihood with EM.

2. Problem Description

Let D={d_1,d_2,…, d_N} be data set, d_s (s=1,2,…N) is a d-dimension(d≥3 for convenient in subsequent discussion) instance column feature vector with its each component valued from discrete random variable. We denote the two of these random variables as X, Y and a set of some other variable as **Z**. We will discuss the independence and/or dependence between X and Y given condition **Z.** Notations $X\perp_{\mathbf{Z}}Y$ and $X//_{\mathbf{Z}}Y$ mean that variable X is conditional independent and dependent on variable Y given condition **Z** in some distribution P, respectively. Constraint-based learning methods [5-8] systematically check the data for independence relations and use those relationships to infer necessary features included in the MB.

A very important measure is mutual information content I(X; Y) between X and Y and condition mutual information content I(X; Y |**Z**) between X and Y conditionally on **Z** as follow:

$$\mathrm{I}(X;Y)=\sum_{x}\sum_{y}p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

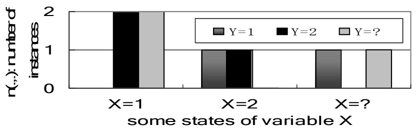
$$\mathrm{I}(X;Y\,|\,\mathbf{Z})=\sum_{\mathbf{z}}\sum_{x}\sum_{y}p(x,y,\mathbf{z})\log_2\frac{p(x,y\,|\,\mathbf{z})}{p(x\,|\,\mathbf{z})p(y\,|\,\mathbf{z})} \tag{2}$$

Where $p(x, y)$ and $p(x, y, \mathbf{z})$ is the joint probability distribution; $p(x|\mathbf{z})$, $p(y|\mathbf{z})$ and $p(x,y|\mathbf{z})$ are all condition probability distribution.

When D is completed, that is, all components of each d_s are known, the computation of I(X; Y |**Z**) is straightforward. Unfortunately, the value cannot be computed when D is not complete, i.e., some components in some d_s are reported as unknown. If **Z** is empty, we wish to estimate the maximum value of I(X; Y) from the incomplete data set D. These unknown components take on multiple types of incomplete cases that are relevant to the estimation of $n(i, j)$, the number of instances X takes on x_i, Y takes on y_j occurring simultaneously in D. For instance, Figure 1(a) shows a data set composed of eight instances with several unknown X and Y components by "?" (for simplicity, we suppose x_i=i and y_j =j). $n(i,?)$ will, for instance, denote the number of instances where X takes on i and Y is unknown. $n(1,?)$=2 in this example (see vector d_3 and d_6 in Figure 1(a)). Figure 1(b) gives the histogram of $n(.,.)$ distribution. The issues involved in estimating theses values from an incomplete data set D are better explained if we regard D as the result of a deletion process applied to a completed but unknown database D_c. We define a consistent completion of D to be any completed database D_c from which we can obtain D using some deletion process. The set of consistent completion D_c is given by all databases in which the unknown components are replaced by one of the possible values of the unobserved variables.

$$\begin{aligned}
&(\cdots x\cdots y\cdots)^T\\
d_1&:(\cdots 2\cdots 1\cdots)^T\\
d_2&:(\cdots 1\cdots 2\cdots)^T\\
d_3&:(\cdots 1\cdots ?\cdots)^T\\
d_4&:(\cdots ?\cdots 1\cdots)^T\\
d_5&:(\cdots 1\cdots 2\cdots)^T\\
d_6&:(\cdots 1\cdots ?\cdots)^T\\
d_7&:(\cdots 2\cdots 2\cdots)^T\\
d_8&:(\cdots ?\cdots ?\cdots)^T
\end{aligned}$$

(a)



(b)

Figure 1. Data Set with some Components of Variable X, Y Unknown (a) and Number of Instances with some X, Y Components (b)

In general, there have been three types of assumption about distribution of the unknown component:

a. unknown completely at random (UCR): the distribution of an unknown component neither depends on observed values nor unobserved values in the data set;

b. unknown partly dependency (UPD): the distribution of an unknown component is a function of the observed values in the data set;

c. unknown completely dependency (UCD): the distribution of an unknown component depends on both observed and unobserved values in the data set.

In order to specify the deletion processes, a dummy binary column vector $F=(F_1,F_2,\ldots,F_m)^T$ may be associated with each feature variable F, here m is the number of states for variable F. When component $F=f_k$ ($k\in\{1,2,\ldots,m\}$) is not observed in some probability in D, iff (if and only if) the F_k takes on value '1' in that probability. When the probability distribution of each F_k is independent of the F and other variables, the data may be seen as UCR; when probability distribution of each F_k is a function of the observed values in the data set, data are UPD; when probability distribution of each F_k is a function of the observed and unobserved values, data are UCD. The Gibbs sampling [9] and expectation maximization (EM) algorithm [14] are well known solutions to handle incomplete data sets but they rely on the assumption that data are UPD. The problem is that UCR and UPD assumptions are hard, if not impossible, to test. Most important, one cannot simply infer the unknown component from the observed ones anymore when the data is UCD. Hence, we need a general approach to deal with the UCD worst-case.

```
Phase I (forward growing)
MB = $\phi$,
While MB has changed
  Find the feature $X$ in $\mathbf{F}$-MB-$\{C\}$ that maximizes
     I($C$; $X$ | MB))
   If $C//_{\mathrm{MB}}X$
      Add $X$ to MB
   End If
End While
Phase II (backwards shrinking)
Remove from MB all variables $X$ for $C\perp_{\mathrm{MB}-\{X\}}X$
Return MB
```

Figure 2. IAMB Algorithm for Learning Markov-Boundary of Class Variable C in Feature Set **F**

In our implementation, we use the algorithm in [6], named IAMB (see Figure 2 above), and a statistically-oriented conditional independence test based on the G-statistical:

$$G=2\sum_{i=1}^{m}\sum_{j=1}^{p}\sum_{k=1}^{q}n(i,j,k)\ln\frac{n(i,j,k)n(.,.,k)}{n(i,.,k)n(.,j,k)} \tag{3}$$

Where m, p, q is the number of states for variable X, Y and condition $\mathbf{Z}$ respectively, $n(i,j,k)$ is number of X takes on x_i, Y takes on y_j and $\mathbf{Z}$ takes on $\mathbf{z}_k$ occurring simultaneously in the sample and $n(.,.,k)=\sum_{i=1}^{m}\sum_{j=1}^{p}n(i,j,k),\ n(i,.,k)=\sum_{j=1}^{p}n(i,j,k),\ n(.,j,k)=\sum_{i=1}^{m}n(i,j,k)$.

The algorithm in Figure 2 is a two phase approach. The growing phase attempts to add the most dependent variables to the Markov blanket (a superset of MB) and the shrinking phase attempts to remove as many irrelevant variables as possible.

Tsamardinos et al. prove in [8] that IAMB returns the correct Markov-boundary under the assumptions that independence test are reliable and that the learning data set is a sample from a distribution P faithful to a Directed Acyclic Graph (DAG) G. A distribution P is said faithful with respect to G with vertex set $\mathbf{F}$, if < G, P > satisfies the faithfulness condition as follow: a. the set of parents, children and parents of children of each variable $F\in\mathbf{F}$ is the unique Markov boundary of F, and b. F and variable F' are not adjacent in G iff there exists $E\in$ $\mathbf{F}\backslash\{F\cup F'\}$ such that $F\perp_E F'$ (see [15]).

3. Tolerance Test

3.1. Statistical Test

Following the principle in Section 2, we proposed a novel tolerance statistical test with no assumptions about the unknown component distribution. This test makes always the worst-case dependency when independency cannot be guaranteed in all the distributions associated with the consistent data completion D_c. Let $\mathrm{I}(X;\ Y\,|\mathbf{Z};\ D_c)$ be the value of $\mathrm{I}(X;\ Y\,|\mathbf{Z})$ evaluated on complete set D_c, we give the following definition.

Definition The notation $\mathrm{I}^t(X;Y|\mathbf{Z})$ is called a tolerance conditional mutual information content with respect to $\mathrm{I}(X;Y|\mathbf{Z})$ if $\mathrm{I}^t(X;Y|\mathbf{Z})$ is the supremum of $\mathrm{I}(X;Y|\mathbf{Z};D_c)$ for all incomplete data set D and for every consistent data completion D_c obtained from the D , i.e.

$$\mathrm{I}^{\mathrm{t}}(X;Y\,|\,\mathbf{Z}) = \sup_{\substack{\forall D\ and \\ D_c\ from\ D}} \mathrm{I}(X;Y\,|\,Z;D_c) \tag{4}$$

The tolerance statistical test resulted from the above $\mathrm{I}^t(X;Y|\mathbf{Z})$ always takes the worst-case assumption about the unknown component distribution to decide whether X and Y are conditionally independent given $\mathbf{Z}$. It is implicitly in the definition that $\mathrm{I}^t(X;Y|\mathbf{Z};D_c) = \mathrm{I}(X;Y|\mathbf{Z};D_c)$ for any completion D_c obtained from D. In general, as the request of faithfulness, we would add an edge in G to mean a direct dependency when the CB algorithm is run on these data during the course of completion. The following theorem shows that a tolerance Markov blanket can be obtained using IAMB(C, D, I^t) (i.e., IAMB run with the tolerance test).

Theorem Suppose the independence tests are correct and that the learning data set D_c is an independent and identically distributed sample from a probability distribution P faithful to a DAG G. D is an incomplete data set obtained from D_c by some type of distribution of unknown component. Then the algorithm IAMB (C, D, I^t) returns a tolerance Markov blanket of C.

Proof If Y and C are parents or children of the same node X in G, then $C\perp_{\mathbf{Z}\cup\{X\}}Y$ for $\mathbf{Z}$=$\mathbf{F}$-$\{C\}\cup\{Y\}$. Recall that IAMB works in two stages. In the forward growing stage, candidate variables are added sequentially to the current MB(C) candidate set when they are not found independent on C conditioned on the current MB(C). In the backwards shrinking stage, the extra variable are removed from MB(C). So Y and X will enter this set during the first stage, but Y will be removed during the second stage because of $C\perp_{\mathbf{Z}\cup\{X\}}Y$ for $\mathbf{Z}$=$\mathbf{F}$-$\{C\}\cup\{Y\}$; b. If $Y\in$MB(C) and C are neither parents nor children of the same node $Z\in\mathbf{F}$-$\{C\}\cup\{Y\}$ in G, then Y is one of the parents and children of C in G according to the faithfulness assumption, thus Y remains dependent on C conditioned on any $Z\in\mathbf{F}$-$\{C\}\cup\{Y\}$. From the above definition, we have $\mathrm{I}^t(C;\ Y|Z)\geq\ \mathrm{I}(C;\ Y|Z;\ D_c)$ for all $Z\in\mathbf{F}$-$\{C\}\cup\{Y\}$. So feature variable Y is necessarily in the output of IAMB(C, D) run with the tolerance test.

3.2. Independent Test

We will show how to design practically a tolerance test based on the G-statistic. Let $n^D(i,j,k)$ be the number of instances in which X takes on x_i , Y takes on y_j and $\mathbf{Z}$ takes on $\mathbf{z}_k$ occurring simultaneously in incomplete data set D. let r_i , r_j , r_k, s_{ij}, s_{ik}, s_{jk} and t_{ijk} be the number of additional instances as contribution to n (i,j,k) in completion D_c owing to $n^D(?,j,k)$, $n^D(i,?,k)$, $n^D(i,j,?)$, $n^D(?,?,k)$, $n^D(?,j,?)$, $n^D(i,?,?)$ and $n(?,?,?)$ in incomplete data set D, respectively. The value that would be computed from the complete data set D_c (if known) is $n(i, j, k)$=$n^D(i, j, k)$+r_i+r_j+r_k+s_{ij}+s_{ik}+s_{jk}+t_{ijk} .

With the above result in mind, we devised an algorithm called *UtoMaxG* to approximate the maximum G, which is depicted in Figure 3. The general idea is to select sequentially the triple (i, j, k) that increases most G.

The incomplete information of X, Y and $\mathbf{Z}$ in D should impose several constraints on r_i , r_j , r_k, s_{ij}, s_{ik}, s_{jk} and t_{ijk}. We consider the following constraint maximum estimate of G related to equation (3): Max G subject to:

$$\begin{cases} \sum_{i=1}^{m} r_i = n^D(?,j,k) \ for \ \forall j,k; \\ \sum_{j=1}^{p} r_j = n^D(i,?,k) \ for \ \forall i,k; \\ \sum_{k=1}^{q} r_k = n^D(i,j,?) \ for \ \forall i,j; \\ \sum_{i=1}^{m}\sum_{j=1}^{p} s_{ij} = n^D(?,?,k) \ for \ \forall k; \\ \sum_{i=1}^{m}\sum_{k=1}^{q} s_{ik} = n^D(?,j,?) \ for \ \forall j; \\ \sum_{j=1}^{p}\sum_{k=1}^{q} s_{jk} = n^D(i,?,?) \ for \ \forall i; \\ \sum_{i=1}^{m}\sum_{j=1}^{p}\sum_{k=1}^{q} t_{ijk} = n^D(?,?,?); \end{cases} \tag{5}$$
$$n(i,j,k) = n^D(i,j,k) + r_i + r_j + r_k + s_{ij} + s_{ik} + s_{jk} + t_{ijk}.$$

```
Input: feature variables $X$, $Y$; conditioning set $\mathbf{Z}$; an incomplete data set $D$;
Output: the maximum for the $G$-statistic $MaxG$;
For all $i, j, k$ do
    Compute $n^D(i, j, k)$
    $n^D(.,.,k) = \sum_{i=1}^{m}\sum_{j=1}^{p} n^D(i,j,k),$
    $n^D(.,j,k) = \sum_{i=1}^{m} n^D(i,j,k),$
    $n^D(i,.,k) = \sum_{j=1}^{p} n^D(i,j,k),$
End for
For all $i, j, k$ do
    $\sigma(i,j,k) = \frac{n^D(i,j,k)n^D(.,.,k)}{n^D(.,j,k)n^D(i,.,k)}$
End for
/*no increasing sort to all $\sigma(i,j,k)$ and save the result to an  one-dimension array Q*/
Q=NISort {$\sigma(i,j,k)$}
$idx$ = 1 /* index of the first element of Q*/
Repeat
    Add as much as instances for the $i, j, k$ with respect to above idx.
    $idx = idx + 1$
Until $D$ is complete
$MaxG$ = $G$-statistic computed by equation (3) in the complete data set.
```

Figure 3. UtoMaxG Algorithm for Completing D

4. Experimental

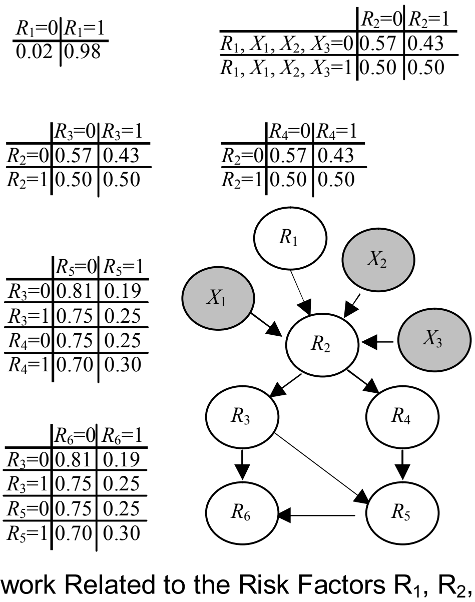
This section gives the results of an experiment based on synthetic data in [1] when UCR, UPD and UCD are all considered. The aim of these experiments is to show that the MB returned by our method can reveal interesting dependencies that may have lost by standard approaches: the available case analysis (i.e., using only the instances where X, Y, $\mathbf{Z}$ are known for the estimation of $X \perp_{\mathbf{Z}} Y$, denoted by ACA) and the EM maximum a posterior probability (EM-MAP) algorithm [15].

4.1. Method and Data Set

We consider the Interleaved Incremental Association Markov Boundary (Inter-IAMB) [6, 16] as our reference Markov boundary discovery algorithm. Inter-IAMB is a variant of IAMB that has been proposed to improve its data efficiency while still being correct under faithfulness assumptions. The difference between IAMB and Inter-IAMB is that the shrinking phase is interleaved into the growing phase in Inter-IAMB. We compare the accuracy of Inter-IAMB with the standard *G*-test using the ACA and the EM-MAP resample approach, denoted by stA-IIAMB and stE-IIAMB respectively, versus Inter-IAMB with the tolerance *G*-test based on *UtoMaxG*, denoted by MaxG-IIAMB. In our implementation, Inter-IAMB considers both tests to be reliable when the number of instances in *D* is at least ten times the number of degrees of freedom and skips it otherwise. Skipping the test means the variables are assumed to be independent without actually performing the test.

A data set is reported in [1] that involves six boolean risk factors $R_1,\ldots, R_6$ observed in a sample of 1841 employees of a Czech car factory. Ramoni and Sebastiani considered these data in [13] and used a standard scoring-based structure learning algorithm to output a structure that they used afterwards as a toy problem to learn the conditional probability tables from incomplete data sets. They assess the robustness of their method called Robust Bayesian Estimator (RBE) that produces probability intervals containing the estimates that can be learned from all completed data sets. In this section, we use the same toy problem to assess the performance of our feature selection method. The Bayesian network which represents the dependency relation of these variables is depicted in Figure 4.

The goal here is to infer the Markov boundary of R_3, that consists of $\{R_2, R_4, R_5, R_6\}$, only one possible false positive here (namely R_1). In order to increase the possible number of false positives), we augment the problem by adding three extra independent variables, denoted by X_1, X_2 and X_3, which has identical distribution to R_1, as depicted in Figure 4. Note that the maximum number of false negatives now equals the maximum number of false positives.



$R_1=0$	$R_1=1$
0.02	0.98

	$R_2=0$	$R_2=1$
$R_1, X_1, X_2, X_3=0$	0.57	0.43
$R_1, X_1, X_2, X_3=1$	0.50	0.50

	$R_3=0$	$R_3=1$
$R_2=0$	0.57	0.43
$R_2=1$	0.50	0.50

	$R_4=0$	$R_4=1$
$R_2=0$	0.57	0.43
$R_2=1$	0.50	0.50

	$R_5=0$	$R_5=1$
$R_3=0$	0.81	0.19
$R_3=1$	0.75	0.25
$R_4=0$	0.75	0.25
$R_4=1$	0.70	0.30

	$R_6=0$	$R_6=1$
$R_3=0$	0.81	0.19
$R_3=1$	0.75	0.25
$R_5=0$	0.75	0.25
$R_5=1$	0.70	0.30

Figure 4: The Bayesian Network Related to the Risk Factors R_1, R_2, R_3, R_4, R_5, R_6 with Three Extra Independent Variables X_1, X_2, X_3

4.2. Procedure used to Remove Data

We associate several dummy feature column vector $Ri=(R_1(i), R_2(i),\ldots, R_{Mi}(i))^T$ with its every component takes on one of the two values 0 and 1 with some probability. Here *Mi* is the number of states for the variable R_i. For each case in the original data set, we generated a combination of values of *Ri* and removed the *k*th state r_{ki} of the variable R_i if the value of $R_k(i)$ was 1. X_1, X_2 and X_3 are also subject to the deletion. The variables X_j are associated with a dummy feature column vector $Xj=(X_1(j),X_2(j),\ldots,X_{M1}(j))^T$.

a. All variables $R_1,\ldots, R_6$ and X_1, X_2, X_3 in the data set are subject to the deletion process covered in section 2. The original network was augmented by the nine dummy feature vectors $Ri(i=1,\ldots,6)$ and Xj $(j=1,2,3)$, marginally independent of $R_1,\ldots, R_6$ and X_1, X_2, X_3, as shown in figure 5(a). Thus, data removed with this process were UCR. This process was repeated with two sets of probability values: $P(R_k(i)=1)=0.05$ (i=1,.., 6; k=1,…,Mi), $p(X_k(j)=1)$ = 0.05 (j=1,2,3;k= 1,…, M1) for the first set, and $P(R_k\ (i)=1)=0.1$ (i= 1,.., 6; k=1,…,Mi), $p(X_k(j)=1)$ = 0.1(j=1,2,3;k= 1,…, M1) for the second set.

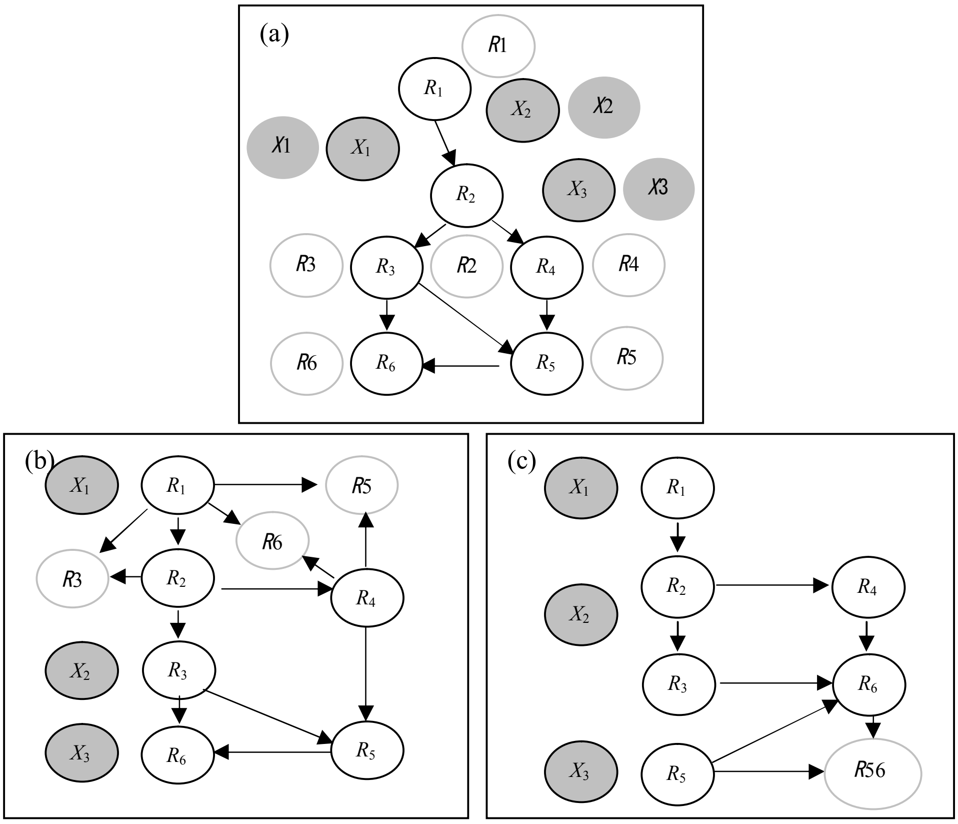


Figure 5: Graphical Representation of the UCR, UPD and UCD Unknown Component Distribution based on Figure 4

b. Only the variables R_3, R_5, and R_6 are subject to the deletion process covered in section 2. We associate these variables with dummy feature column vector R3, R5 and R6. The distribution for each of these dummy vectors was a function of the variables R_1, R_2, and R_4. Since R_1, R_2, and R_4 are fully observed and the distribution of R3, R5 and R6 is only dependent on the values observed in the incomplete data set, as shown in figure 5(b), data removed with this process are UPD. We have considered two different probability mass functions for the dummy vectors. For the first set, $P(R_k(i)=1)=0.1(i\in\{3,5,6\}$; k= 1,…,Mi), if its two parents have the same binary value and $P(R_k(i)=1)=0(i\in\{3,5,6\};k=1,\ldots,Mi)$, if its two parents take on different binary values. For the second set, $P(R_k(i)=1)=0.2(i\in\{3,5,6\};k=1,\ldots,Mi)$, if its two parents have the same binary value and $P(R_k(i)=1)=0(i\in\{3,5,6\};k=1,\ldots,Mi)$, if its two parents take on different binary values.

c. Only the variables R_5 and R_6 were subject to the deletion process covered in section 2. We associated the variables R_5 and R_6 with a M-dimension (M=min{M5,M6}) dummy feature vector R56. Again, we generated a value of the feature vector R56 and removed the kth state r_{k5} and r_{k6} of variables R_5 and R_6 respectively if the value of $R_k(56)$ was 1. Since the distribution of R56 depends on the unobserved values in the data set, as shown in figure 5(c), values removed with this process are UCD. We have considered two different probability mass functions for R56. For the first set, $P(R_k(56)=1)=0.1$ (k=1,…,M), if its two parents have the same binary value and $P(R_k(56)=1)=0$ (k= 1,…,M), if its two parents take on different binary values. For the second set, $P(R_k(56)=1)=0.2$ (k= 1,…,M), if its two parents have the same binary value and $P(R_k(56)=1)=0$ (k=1,…,M), if its two parents take on different binary values.

To evaluate the accuracy and data efficiency of the tolerance MaxG test on unknown component distribution, we used the deletion procedures described in section 2. For each deletion mechanism, we generated 100 data sets with 1000 instances in which the average proportion of unknown components were 5% and 10%. This makes a total of 600 hundreds data sets. Using R_3 as the target variable, we run stA-IIAMB, stE-IIAMB and our MaxG-IIAMB methods.

As a result, we measure the accuracy from perfect prediction and recall by combining precision (i.e. the number of true positives in the output divided by the number of nodes in the output) and recall (i.e., the number of true positives divided by the true size of the Markov Boundary) as $1-\sqrt{(1-precision)^2+(1-recall)^2}$. Figure 6 shows the accuracy and standard deviation values over 100 databases.

Obviously, the MaxG-IIAMB method yields a higher accuracy in all cases. The benefit is more apparent with 10% unknown rate. Interestingly, this is not only true for the UCD experiment for which both EM and ACA techniques are biased, but it also holds for UCR and UPD.

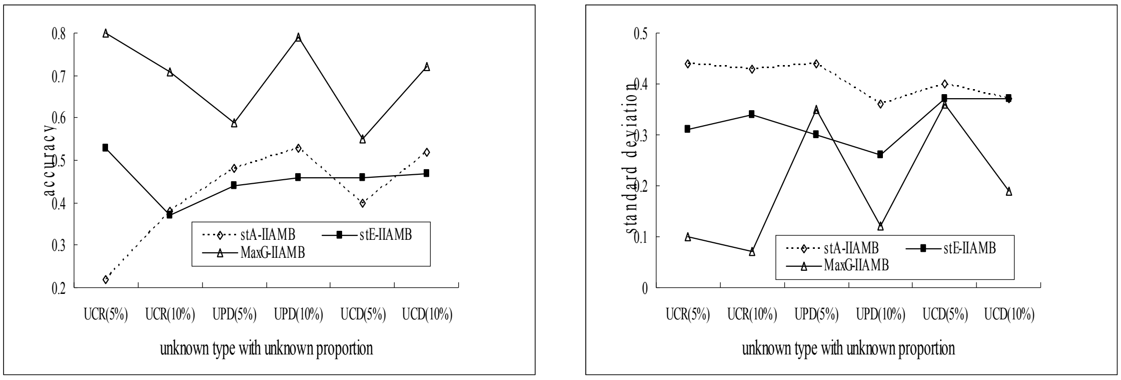


Figure 6. Accuracy and Standard Deviation from Perfect Precision and Recall

5. Conclusion

In this paper, we induced a tolerance constraint-based MB learning method from incomplete data. An application on synthetic incomplete data was carried out to illustrate its practical relevance and benefit compared to EM and available case analysis techniques.

Acknowledgements

This research was supported by the Education Foundation of Sichuan, China under Grant No. 12ZB070.

References

[1] J Whittaker. Graphical Models in Applied Multivariate Statistics. *John Wiley & Sons* (USA). 2009.

[2] DL Swets, J Weng. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*. 1996; 18(8): 831-836.

[3] D Koller, M Sahami. *Toward Optimal Feature Selection.* Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996), Bari (Italy). 1996: 284–292.

[4] D Margaritis, S Thrun. Bayesian network induction via local neighborhoods. *Advances in Neural Information Processing Systems*, MIT Press. 2000: 505–511.

[5] J Pena, J Bjorkegren, J Tegner. *Scalable, efficient and correct learning of markov boundaries under the faithfulness assumption.* Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning under Uncertainty (ECSQARU 2005), Lecture Notes in Artificial Intelligence 3571. 2005; 21: 136–147.

[6] I Tsamardinos, CF Aliferis, A Statnikov. *Algorithms for large scale markov blanket discovery*. Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference(FLAIRS). St. Augustine (USA). 2003: 376–381.

[7] S Yaramakala. Fast markov blanket discovery. MSThesis, Iowa State University(USA). 2004.

[8] S Yaramakala, D Margaritis. *Speculative markov blanket discovery for optimal feature selection.* Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), Houston (USA). 2005: 809–812.

[9] S Geman, D Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 1984; 6(6): 721–741.

[10] N Friedman. *Learning belief networks in the presence of unknown values and hidden variables.* Proceedings of the Fourteenth International Conference on Machine Learning. Nashville (USA).1997: 125–133.

[11] N Friedman. *The bayesian structural Emalgorithm.* Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence(UAI'98). University of Wisconsin Business School (USA). 1998: 129–138.

[12] O Francois, P Leray. *Generation of Incomplete Test-Data using Bayesian Networks.* Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN), Orlando (USA). 2007: 1-6.

[13] M Ramoni, P Sebastiani. Robust learning with missing data. *Machine Learning*. 2001; 45(2): 147–170.

[14] AP Dempster, NM Laird, DB Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Roy. Statist. Soc. Ser. B*. 1977; 39(1): 1–38.

[15] RE Neapolitan. Learning Bayesian Networks. *Prentice Hall* (UK). 2004.

[16] CF Aliferis, I Tsamardinos, A Statnikov. *Hiton: a novel markov blanket algorithm for optimal variable selection.* Proceeding of the 2003 Symposium of the American Medical Informatics Association (AMIA 2003). Marriott Wardman Park, Washington, DC (USA). 2003: 21–26.