International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:6, 2011

Data-organization before Learning Multi-Entity Bayesian Networks Structure

H. Bouhamed, A. Rebai, T. Lecroq, and M. Jaoua

***Abstract*—**The objective of our work is to develop a new approach for discovering knowledge from a large mass of data, the result of applying this approach will be an expert system that will serve as diagnostic tools of a phenomenon related to a huge information system. We first recall the general problem of learning Bayesian network structure from data and suggest a solution for optimizing the complexity by using organizational and optimization methods of data. Afterward we proposed a new heuristic of learning a Multi-Entities Bayesian Networks structures. We have applied our approach to biological facts concerning hereditary complex illnesses where the literatures in biology identify the responsible variables for those diseases. Finally we conclude on the limits arched by this work.

***Keywords*—**Data-organization, data-optimization, automatic knowledge discovery, Multi-Entities Bayesian Networks, score merging

I. Introduction

IT is worth highlighting that knowledge representation and the related reasoning, thereof, have given birth to numerous models. The graphic probability models, namely, Bayesian Network (BN), introduced by Judea Pearl in the 1980s, have been manifested in to practical tools useful for the representation of uncertain knowledge, and reasoning process from incomplete information. A major challenge in such modelling is the large number of variables that increases exponentially the computational complexity of learning Bayesian Networks structures [13].

Actually, there are several types of BN, e.g, Multi-Agent BN, Oriented-Object BN, Dynamic BN etc. [12]. Reference [7] proposed a formalism that unifies the first order logic and probability theory. This formalism is called Multi-Entities BN (MEBN). MEBN are composed of fragments (called MFrags) representing the joint distribution of a subset of variables. A fragment consists of a set of variables context, a set of variables input, a set of resident variables, a direct acyclic graph (DAG) on the input variables and the resident variables (in which the variables input are nodes root) and a set of conditional distributions for each local resident variables. An MFrags is very close to a BN for which context nodes are observed. A MEBN is a set of MFrags which must satisfy the properties that a variable should never be an ancestor of itself (no path) [12].

In this study our principle is based on the fact that the complexity of learning BN structure is exponential giving the exponential increase in the number of variables.

Heni Bouhamed is with Ecole Nationale d'Ingénieurs de Sfax Tunisia, Tunisia (e-mail:heni_bouhamed@yahoo.fr)

So there is a need for methods that avoid learning structure with all variables at the same time when the number of variables is large.

The solution that we will propose is based on the modulation of learning structure: each class has its own learning before forming the final structure containing all the variables. For this, we will learn from the MEBN formalism.

As for the remaining constituent sections of the present research work, they are organized as follows: the next section is allotted to the introductory exposition of BN structure learning problem. As for the following section, a novel approach for data-organization before learning MEBN structures is going to be presented and which is going to be applied and tested on a special biological data-base. As for the following section, a new heuristics of learning MEBN structure is going to be presented. As regards the last section, it depicts our conclusion and the limits arched by this work.

II. Problem of learning Bayesian network structure from data

The number of all possible structures for Bayesian networks has been shown to increase as a super-exponential on the number of variables. Indeed, Reference [10] derived the following recursive formula for the number of DAG with n variables:

$$r(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} r(n-i) = n^{2^{O(n)}} \quad (1)$$

which gives: r(1)=1, r(2)=3, r(3)=25, r(5)=29281, r(10)=4,2 10^{18}

This means that, it is impossible to perform an exhaustive search of all structures in a reasonable time when the number of nodes exceeds seven. Most structure learning methods use heuristics to search the space of DAGs [12].

III. New approach for data-organization before learning MEBN structures

It is well recognized that the strategy based on single variable analyses has a very limited value in elucidating the mechanisms involved in complex phenomena [4]. The approach we propose here is fundamentally a multivariate analysis and operates in four steps. It start by the calculation of a statistical score (test value or p-value) for each variable depicting its involvement in a phenomenon and then classifies variables according to their association to the studied phenomenon and the complementarity between them.

In the third step, we calculate a 'global' statistical score for each class or cluster of variables that is a function of the

correlation between the variables and their scores. Finally the classes will be ranked in a decreasing order based on their global score (after a logarithmic transformation in order to have a high score if the value of the score statistic is low) and a number of them are selected.

A. Single variable analysis and Classification

The chi-square test is a widely use test to measure the association between categorical variables. For binary variables (two categories) such as the disease status and a risk factor in epidemiological studies the chi-square is easily calculated [15].

Classification is the act of creating groups of variables by identifying those that share common characteristics (redundancy, correlation). The choice of a classification algorithm and whether it is supervised or not, depends on the nature and characteristics of data. Classification can also be done based on experts' knowledge in the field [16].

B. Merging scores of each class

In this step it is question on how to derive a score for each class based on the scores of variables within classes. Most of the methods used to combine scores in computer science literature and specifically in knowledge discovery in database, are those that consists in merging scores of independent variables such as: Average and Maximum (MAX) scores [1]-[9], Sum, Minimum(MIN) and product scores [5].

However, the statistical literature provides many methods that combine score by taking into account the correlations between variables. Two of these are the Truncated Product Method [11] that combines p-values of correlated tests and Length Heuristic (LH) who chooses the best tests to represent the class [14].

Whose algorithms are described below:

Truncated Product Method (TPM) Algorithm

For each class of variables, the following steps are to be undertaken:

1: Construct a correlation matrix for variables within the class.
2: Calculate the Cholesky matrix *C* for each correlation matrix
3: Choose the scores' maximum value π (p-values) to be selected.
4: Calculate $W_0 = \prod_i^L p_i^{I(p_i \le \pi)}$
Where *L* designates the number of variables in the class.
5: Put *A=0*
6: Randomly generate *L* independent values from a uniform distribution generating the vector R^*: $u_1^*,.....,u_L^* \in [0,1]$
7: Transform the vector R^* into another vector *R* having the values with equation:

$$R = 1 - \Phi\left\{C\Phi^{-1}(1-R^*)\right\} \tag{2}$$

$$//\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \ (3)\ \Phi^{-1}(x) = \int_{x}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{3}$$

$$W = \prod_i^L R_i^{I(R_i \le \pi)}$$

8: Calculate
9: If $W \le W_0$, then *A=A+1*
10: repeat steps 6 to 9, B times
11: obtain the combined score (p-value) by means of *A/B*.

Length Heuristic (LH) Algorithm

The objective is to choose the best tests (statistic score) among the other tests within a class.

1: We observe a sequence of a correlated statistical tests of one class *C: (T_1, T_2, …, T_j)*.
2: We compute a probability for each T_i according the following formula:

$$pr(T_{\max} > c) \le \Phi^{-1}(c) + e^{-c^2/2} \frac{L}{2\pi} \tag{4}$$

Where $L = \sum_{j=2}^{J} \arccos(p_j)$ and p_j is the correlation coefficient between T_{j-1} and T_j.
3: The tests with the largest value will be selected as the representative of the class.
4: Repeat 1, 2 and 3 steps for each class.

C. Ranking and Selection of classes closely related to a phenomenon

Classes of variables can then be ranked based on their scores. If a p-value is used as a score, ranking is based on sorting in increasing order (smaller p-values are indicative of higher significance). Often the score is calculated as the logarithmic transformation $-\text{Log}_{10}$(p-value) such that a high score value implies a high degree of significance (association).

The purpose of this step is to select classes of variables that are the most associated to the phenomenon. There are numerous methods for selecting most influential variables in statistical and computer science literature. However, to our knowledge there is few methods that deal with selecting classes of variables. Here we propose a new method inspired of [6].

D. Experimentation

Genome wide association studies are studies in which geneticists assess the association of thousands of molecular markers with a disease phenotype. The traditional way of analyzing the data is to compute chi-square association tests and corresponding p-value for each marker (variables) and then select those with the weakest p-values as indicative of interesting genome region on 213 Canadian patients with schizophrenia and 241 Canadian controls, genotyped. for 164 single nucleotide polymorphism (SNP) markers on chromosome 13.

The database is a text file formatted as follows:

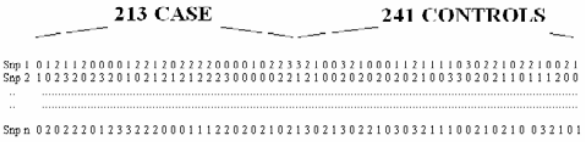

Fig. 1 Data base Format

Where 0: corresponds to the *aa* genotype,
1: corresponds to the *Aa* or *Aa* genotype
2: corresponds to the *AA* genotype
3: corresponds to missing data

Our first objective is to select genomic regions (classes of variable) which are the most significantly associated to the disease (schizophrenia). Secondly, we will try to model these regions by BN that provide a tool for disease diagnosis.

1) Data Processing Steps

- Calculating a score for each variable (SNP) which is here the p-value from the chi-square test statistic.
- Variable classification: we class variable according to the genetics experts that suggest a gene might best represent a class of SNPs.
- Combine scores from each class using different strategies proposed in Sub-Section *B* and compare results.
- Rank classes according to their scores.
- Select classes involved in the disease (schizophrenia) using the approach described in Sub-section *C*.

2) Results of different fusion methods score

To compare p-value combining methods we took as reference the region G72 described by [2] as the region responsible for Schizophrenia. TABLE I gives results of TPM, LH and MIN Method. The genes found by these three methods are very similar and are contained in the G72 region.

TABLE I
RESULTS FOR THREE METHODS

	Rank	Gene name	Region	Score	Stati Sum	P-value
MIN	1	FOX01	151	1.33	1.33	0.70
	2	NARG1L	140-141	1.20	2.58	0.63
TPM	1	NARG1L	140-141	1.52	1.52	0.15
	2	FOX01	151	1.09	2.62	0.10
LH	1	FOX01	151	1.33	1,33	0.64
	2	NARG1L	140-141	1.20	2.58	0,59

In terms of complexity it is clear that the algorithm using the MIN method is the best but to check the reliability of the results of the three methods we will study empirical distribution of the observed minimum p-value ($P_{\min}^{obs}$) with the Monte Carlo simulations whose principle is as follows:

We simulate B times the data by calculating each time minimum p-value ($P_{\min}^{i}$) and at the end we calculate the overall p-value of each step by the following formula:

$$P_G = \frac{cardinality\left\{P_{\min}^{(i)} \le P_{\min}^{obs}\right\}}{B} \quad (5)$$

The p-value for the overall process using TPM is equal to 0.09, while it was 0.41 for MIN method and 0.33 for LH method. We can conclude that the results using the TPM are more significant and that this method is preferable in subsequent work.

3) Discussion

Using our approach we have successfully identified the most significant genes involved in disease schizophrenia that were found to be consistent with published results. The number of initial variables was decreased from 164 to 3, which dramatically decreases the computational complexity of learning Bayesian network structure; in fact the number of possible structure go from $r(164)>10^{406}$ down to $r(3)=25$ possible graphs, without large information loss.

IV. HEURISTIC TO LEARNING MEBN STRUCTURE OF SELECTED CLASSES

In the first parts of our work we have already defined methods for filtering the number of classes according to the degree of their implications in a given phenomenon. In this part we will present a new method for learning BN structure, building on the formalism introduced by [7] called MEBN.

Consider an example of ten nodes with a node phenomenon (Ph) and the other nodes are divided into four classes as follows:

Class 1: V1, V2, V3
Class 2: V5, V6, V7
Class 3: V8, V9
Class 4: V4

A structure will be learned for the variables of each class to which we add the node phenomenon (Ph), always as a cue node (have only children) (Figures 2, 3, 4 and 5).

Then a learning structure will be made between the phenomenon variable and the variables from all classes that have a direct edge with Ph node (Ph's children) (Figure 6).

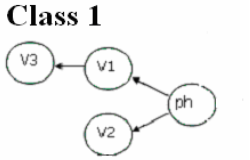

Fig. 2 Structure learning of a class 1

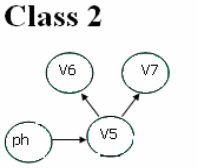

Fig. 3 Structure learning of a class 2

Fig. 4 Structure learning of a class 3

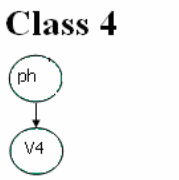

Fig. 5 Structure learning of a class 4

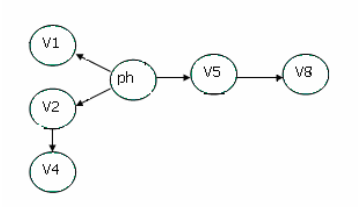

Fig. 6 Phenomenon variable and the variables that have a direct relationship

Finally, the remaining nodes of classes will be added to the final graph according to the class structures previously inferred (Figure 7).

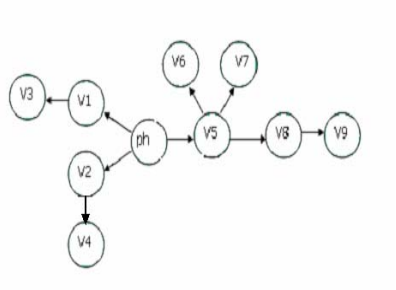

Fig. 7 Remaining nodes added to the final graph

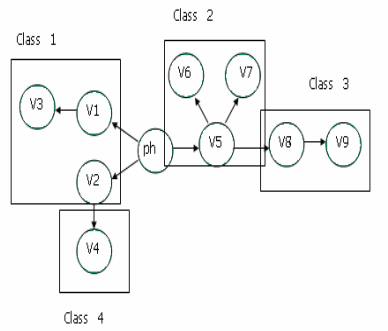

Fig. 8 Multi entities Bayesian network learning structure final graph

V. Conclusion and Limitations of Our Work

The approach proposed here first allows identifying classes of variables that are the most involved in a given phenomenon using several steps. Then, in order to get a graphical model that depict the relationships between selected variables and the phenomenon, we use MEBN, although the interest of using them compared to complexity classical structure learning algorithms remain to be demonstrated. Our approach for variable class selection was illustrated on an example from a genetic study of schizophrenia. The comparison of the proposed approach as a whole with other available similar methods will be the objective of another publication.

References

[1] M. L. Damian and F. H. Donald, "Combining multiple scoring systems for target tracking using rank-score characteristics," Information Fusion, 10, 124-136, 2009.
[2] S. Detera-Wadleigh and F. McMahon, "G72/g30 in schizophrenia and bipolar disorder: review and meta-analysis," Biological Psychiatry, 60(2): 106-114, 2006.
[3] P. Dempster, N. Laird and B. D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Stat Soc B 39: 1-38, 1977.
[4] M. Geudj, J. Wojcik, D. Robelin, M. Hoebeke, M. Lamarine and G. Nuel, "Detecting Local High-Scoring Segments: a First-Stage Approach for Genome-Wide Association Studies," Statistical Applications in Genetics and Molecular Biology, Vol. 5, Iss. 1, Article 22 2006.
[5] A. Jain, K. Nandakumar and A. Ross, "Score normalization in multimodal biometric systems," Pattern Recognition, volume 38 Issue 12, Pages 2270-2285, Dec 2005.
[6] S. Karlin and S. Altshul, "Applications and statistics for multiple high-scoring segments in molecular sequences," Proceedings of the National Academy of Science USA 90, 5873-5877, 1993.
[7] K. B. Laskey, "MEBN: A language for first-order Bayesian knowledge bases," Artificial Intelligence, 172, 140-178, 2007.
[8] O. Francois, and P. Leray, "Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens," In Proceedings of 14ème Congrès Francophone Reconnaissance des Formes et Intelligence Artificielle, RFIA, pages 1453-1460, Toulouse, France, 2004.
[9] H. N. Parkash and D. S. Guru, "Offline signature verification: An approach based on score level fusion," International journal of computer applications, 0975-8887, Article 10, No.18, 2010.
[10] R. W. Robinson, "Counting unlabeled acyclic digraphs," Combinatorial Mathematics, 622, 28–43, 1977.
[11] D. Zaykin, L. Zhivotovsky, P. Westfall and B. Weir, "Truncated product method for combining P-values," Genet Epidemiol, 22(2), 170-85, Feb 2002.
[12] O. François, "De l'identification de structure de réseaux bayésiens à la reconnaissance de formes à partir d'informations completes où incompletes, "Thèse de doctorat. Institut National des Science Appliquées de Rouen, 2006.
[13] P. Leray, "Réseaux Bayésiens: apprentissage et modélisation de systèmes complexes," habilitation à diriger les recherches, Université de Rouen, 2006.
[14] B. Efron, "The lenght heuristic for simultaneous hypothesis tests," Biometrica, 84, 143-157, 1997.
[15] C. Herman and E. L. Lehman, "The use of Maximum Likelihood Estimates in chi-square tests for goodness of fit," The annals of Mathematical Statistics volume 25, Number 3, 579-586, 1954.
[16] X. Rui, and C. W. Donald, "Clustering," IEEE Press/Wiley, oct 2008.