A New general learning structure approach of Bayesian networks from data

Heni Bouhamed [1], Afif Masmoudi [2], Thierry Lecroq[1], Ahmed Rebaï [3]

[1] University of Rouen, LITIS EA 4108, 1 rue Thomas Becket, 76821 Mont-Saint-Aignan cedex, France
Heni.bouhamed@yahoo.fr
Thierry.lecroq@univ-rouen.fr
[2] Department of Mathematics, Faculty of Science of Sfax, Soukra B.P 802 Sfax, Tunisia
Afif.masmoudi@fss.rnu.tn
[3] Bioinformatics Unit, Centre of Biotechnologie of Sfax, 3018 Sfax, Tunisia
Ahmed.rebai@cbs.rnrt.tn

Abstract. Nowadays, Bayesian Networks (BNs) have constituted one of the most complete, self-sustained and coherent formalisms useful for knowledge acquisition, representation and application through computer systems. Yet, the learning of these BNs structures from data represents a problem classified at an NP-hard range of difficulty. As such, it has turned out to be the most exciting challenge in the learning machine area. In this context, the present work's major objective lies in setting up a further solution conceived to be a remedy for the intricate algorithmic complexity problems imposed during the learning of BN-structure through a massively-huge data backlog. Our present work has been constructed according to the following framework; on a first place, we are going to proceed by defining BNs and their related problems of structure-learning from data. We, then, go on to propose a novel heuristic designed to reduce the algorithmic complexity without engendering any loss of information. Ultimately, our conceived approach will be tested on a car diagnosis as well as on a Lymphography diagnosis data-bases, while our achieved results would be discussed, along with an exposition of our conducted work's interests as a closing step to this work.

Keywords: Bayesian Network; structure learning; modeling; algorithmic complexity.

1 Introduction

The huge amounts of data, made recently available, pertaining to the various research fields, have made it crucially critical for the learning techniques to be efficient, in so far as the processing of complex data dependences is concerned. Owing to their flexibility and easily-recognizable mathematical formulations, BNs are most often the basic selected model opted for in a wide-array of application-fields whether astronomic, textual, bioinformatics and web-mining applications. Yet, with an incredibly huge number of variables, the learning BNs structure from data remains

a big challenge to be retained and considered in terms of calculation power, algorithmic complexity and execution time [1]. Most recently, however, various algorithms have been developed with respect to the BNs learning structures from data-bases [2, 3, 4]. A considerable class of these algorithms rests on the metric-scoring methods, excessively compared and exhaustively applied as approaches [5, 6]. Nevertheless, these algorithms and scoring methods remain still insufficient with regards to those cases in which the number of variables exceeds hundreds of thousands [7]. In so far as our work is concerned, these algorithms and metric scores are not going to be dealt with or questioned. Rather, we seek to further enrich them through a new heuristic based on clustering pertaining to structure learning, in a bid to further reduce the algorithmic complexity as well as the execution time, with the purpose of modeling some previously non-modelizeable information systems, by using, exclusively, the underway available algorithms.

Our work, we reckon, is critically important for a number of various reasons. First, we have managed to demonstrate, throughout its scope, that by wholly subdividing an information system into sub-sets, we tend to dramatically reduce the number of possible structures necessary for learning the BNs structures. Second, a special heuristic has been devised and proposed whereby this reduction could be exploited without engendering any significant loss of data. Ultimately, by combining our proper heuristic with the existing prevailing structure-learning algorithms, one can considerably reduce the extent of algorithmic complexity as well as the learning of BNs structure from data execution time, in such a way that even a large number of non-modelizeable variables could be treated or processed.

The remainder of this article has been arranged as follows. The next upcoming section deals with the BNs and their structure learning problems. In the following section, we are going to put forward a new heuristic which we shall test upon a car-diagnosis and Lymphography diagnosis data bases. Finally, we will close up our work by concluding and paving the way for certain potential perspectives relevant to future researches.

2 BNs and structure learning from data problems

It is worth highlighting that knowledge representation and the related reasoning, thereof, have given birth to numerous models. The graphic probability models, namely, BN, introduced by Judea Pearl in the 1980s, have been manifested in the practical tools useful for the representation of uncertain knowledge and reasoning process from incomplete information.

Hence, the BN graphic representation indicates the dependences (or independences) between variables and provides a visual knowledge representation tool, that turns out to be more easily understood by its users. Furthermore, the use of probability allows to take into account the uncertainty, by quantifying the dependences between variables. These two properties have been at the origin of the first terms allotted, initially, of BN, "probabilistic expert systems", where the graph used to be compared with some set rules pertaining to a classic expert system, and

conditional probability presented as a quantification measurement of the uncertainty related to these rules.

The number of all BN possible structures has been shown to ascend sharply as a super-exponential on the number of variables. Indeed, Reference [12] derived the following recursive formula for the number of Directed Acyclic Graph (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 in cases the number of nodes exceeds seven. In fact, most structure-learning methods use heuristics to search the DAGs space.

3 A new clustering-based heuristic: theoretical framework and methodology

The idea lying behind our conceived proceeding lies in the rapid super-exponential surge of algorithmic complexity of learning BN structure from data with respect to the rise in the number of variables. To remediate this problem, our preconceived idea consists in subdividing the variables into subsets (or clusters), by treating each cluster's learning structure separately, while looking for a convenient procedure whereby the different structures could be assembled into a final structure version. In this regard, it has been noticed that in numerous information systems, so as not to say in most of them, there exists, at least, one single central variable of a global interest constituting the basis of the system's modelization. In this respect, we reckon to execute the processing of each cluster learning structure with the central interest variable, then, proceed by assembling the different various structures around this central variable as a next step.

In the upcoming part (3.1), we shall demonstrate, mathematically, that by subdividing the variables and by separately processing each cluster's learning structure with the interest variable, we dramatically reduce the number of possible DAG in respect of the simultaneous learning structure of the entire variables. After that, in part (3.2), we are going to explain our proposed framework procedure as well as the methodologies to be pursued.

3.1 Theoretical background

The below represented Robinson formula depicts the number of possible DAG in respect of the variables' number:

The Robinson formula is $r(n) = \sum_{i=1}^{n} (-1)^{i+1} 2^{i(n-i)} \binom{n}{i} r(n-i)$; $r(1) = 1$, where n stands for the number of variables.

In this section, we will prove that $r(n) > \sum_{l=1}^{k} r(J_l + 1)$, where *n-1= J_1+J_2+…+J_k*; *J_l+1<n* and $l = 1, \dots, k$.

Proposition 1

For all $n \geq 2$, we have:

i) $r(n) \geq 2^{n-2} n\, r(n-1)$

ii) $r(n) \geq 2^{\frac{(n-1-J)(n+J-2)}{2}} n(n-1) \dots (J+2) \times r(J+1)\ \forall\ (1+J) \leq n$

View Proof in Appendix A

We denote by: $n\text{-}1 = J_1 + J_2 + \dots + J_k$; $J_l + 1 < n$

$J_{l^-} = \min_{1 \leq l \leq k} (J_l)$

$J_{l^+} = \max_{1 \leq l \leq k} (J_l)$

Proposition 2

$\frac{r(n)}{\sum_{l=1}^{k} r(J_l+1)} \geq \rho\left(n, J_{l^-}, J_{l^+}, k\right) \gg 1$; Where

$$\rho\left(n, J_{l^-}, J_{l^+}, k\right) = \frac{2^{\frac{\left(n-1-(J_{l^+})\right)(n+(J_{l^-})-2)}{2}}\, n(n-1)\dots\left((J_{l^+})+2\right)}{k}$$

Proof

According to ii) of Proposition 1, we have

$r(n) \geq 2^{\frac{(n-1-J_l)(n+J_l-2)}{2}} n(n-1) \dots (J_l+2) r(J_l+1); l = 1, \dots, k.$

Hence,

$\sum_{l=1}^{k} \mathrm{r(n)} \geq \sum_{l=1}^{k} 2^{\frac{(n-1-J_l)(n+J_l-2)}{2}} n(n-1) \dots (J_l+2) r(J_l+1)$

$\mathrm{k} \times \mathrm{r(n)} \geq \sum_{l=1}^{k} 2^{\frac{\left(n-1-(J_{l^+})\right)(n+(J_{l^-})-2)}{2}} n(n-1) \dots (J_{l^+}+2) r(J_l+1)$

Where $2^{\frac{\left(n-1-(J_{l^+})\right)(n+(J_{l^-})-2)}{2}} n(n-1) \dots (J_{l^+}+2) = \text{constant}$

$\mathrm{k} \times \mathrm{r(n)} \geq 2^{\frac{\left(n-1-(J_{l^+})\right)(n+(J_{l^-})-2)}{2}} n(n-1) \dots (J_{l^+}+2) \sum_{l=1}^{k} r(J_l+1)$

$$\mathrm{r(n)} \geq \frac{2^{\frac{\left(n-1-(J_{l^+})\right)(n+(J_{l^-})-2)}{2}}\, n(n-1)\dots\left((J_{l^+})+2\right)}{k} \sum_{l=1}^{k} r(J_l+1)$$

Therefore: $\frac{r(n)}{\sum_{l=1}^{k} r(J_l+1)} \geq \rho\left(n, J_{l^-}, J_{l^+}, k\right) \gg 1,$

where $\rho\left(n, J_{l^-}, J_{l^+}, k\right) = \frac{2^{\frac{\left(n-1-(J_{l^+})\right)(n+(J_{l^-})-2)}{2}}\, n(n-1)\dots((J_{l^+})+2)}{k}$

Finally, we can conclude that $r(n) >> \sum_{l=1}^{k} r(J_l+1)$, where $n\text{-}1 = J_1 + J_2 + \dots + J_k$; $J_l + 1 < n$ and $l = 1, \dots, k$.

3.2 Procedure and applied methodologies

3.2.1 Choice of a global interest variable

The aim of such a step is to select a diagnosis variable, or a global interest variable, of the information system to be modelled. This variable could be a status variable, for instance: status of individuals (ill/sound), cars' status (start/not start), customer status (solvent/not solvent) etc. In such cases, the choice is an easy and immediate one. As for those cases in which the choice is not evident, due to the analyst's ignorance of the studied variables' nature, one might eventually resort to some classical automatic

data-exploring methods. For instance, to the principal component analysis in a bid to dismantle, from the first resulting axis, the mostly intervening variable in the information system subject of study.

3.2.2 The variables' clustering

The automatic type of clustering is the most frequently used and widespread technique among the data-analysis and data mining descriptive techniques. It is often used when we get a huge amount of data, within which we intend to distinguish some homogeneous subsets suitable for processing and for differential analyses [13].

Actually, there exist two major well-known algorithm classifying families in the literature, namely, the partition methods as well as the ascending hierarchical-clustering ones. The advantage of the ascending-hierarchical methods, as compared to the partitioning one, lies in the fact that they enable to choose, appropriately, the optimum number of clusters. Nevertheless, the partitioning criterion is not global; it exclusively depends on the already-obtained clusters, since two variables placed in different clusters could by no means be compared any more. Contrary to the hierarchical methods, the partitioning algorithms might continuously improve the clusters-quality [13], in addition to the fact that their algorithmic complexities are linear (for the most popular algorithms). Regarding our present work, however, we have chosen to use the K-means algorithm, as it is the most popular and applied in the literature, added to fact that its algorithmic complexity is linear (O(n)) [14]. We also propose to use a hierarchical clustering algorithm along with the bootstrap technique to obtain the optimal number of clusters that will be introduced as entries in the K-means algorithm. To note, the databases that will be applied to test our approach, in the experimentation section, consist of categorical variables, and regarding the performance of clustering we will use the toolbox ClustOfVar with the software R [15]. In particular, we will use the variant K-means for categorical variables [16] and the link-likelihood approach [17] (hierarchical clustering algorithm for categorical variables). To assess the stability of all possible partitions, 2 to p-1 (where p is the total number of variables) clusters from the hierarchical clustering, we will use a feature called "Stability" (also developed in the ClustOfVar toolbox) based on the "bootstrap" technique. The result is a graph which is then a tool to help to select the number of clusters. The user can be choosing the number *K* of clusters to the heights of the first increase in the stability.

3.2.3 Structure learning

A structure learning of each cluster's variable with the interest variable, will be undertaken. The ultimate structure would be the *n* structures obtained from each cluster around the interest variable.

Numerous algorithms have been devised with regards to the learning of BNs structure, noteworthy among which is the algorithm PC [18], Maximum weight spanning tree (MWST) [19], the algorithm K2 [3], Greedy Search (GS) [4] etc. Still, the most frequently used algorithm, according to the specialized literature, remains

the algorithm K2. It is characterized by its rapidity, promptness and the stability of constancy of its results. Yet, its major problem remains the initial order required for the entries, which is very influential on the final results. As a remedy to their problem, the most frequently used solution consists in applying the upstream MWST algorithm [20], to obtain a certain order of nodes useful to be introduced as entries for the K2 algorithm. Less sensitive to the data-base size variation, the MWST algorithm yields a graph quite similar to the original one. Nevertheless, this method runs exclusively through the (very poor) trees space [20]. It, therefore, turns out to us to be the most exclusive effective tool necessary for getting an initial order of nodes very accurate and close to the data, useful to be used in entry with the K2 algorithm.

To note that in our work, we will use the BNT toolbox [23] running on Matlab software (2010 version) to apply the MWST and K2 algorithms for learning structure. We will use also the BNT toolbox to learning parameters and inference.

4 Experimentations procedures

4.1 Data-bases

We are going to test our designed approach, firstly, on a car diagnosis data-base dubbed "Car Diagnosis 2". It is made up of eighteen variables (see Table 1), among which is a statue variable called "Car starts", the global interest variable of the information system. The parameters' generating file of this data base is available on the site http://www.norsys.com/downloads/netlib/. According to these parameters, we have been able to generate some 10000 examples, among which thirty two have been left aside for the references' testing phase. Secondly, the model will be applied on a Lymphography diagnosis data-base dubbed "Lymphography". It is made up of nineteen variables (see Table 2), among which is a statute variable called "Diagnosis", the global interest variable of the information system. This lymphography domain has been obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. In this respect, we would like to thank Mr. Zwitter and Mr. Soklic for providing the data. Among the 148 instances of data, thirty two have been left aside for the references' testing phase.

Table 1 "Car diagnosis 2" variables.

Variables' names	possible states	Variables names	possible states
AL : Alternator	(Okay, Faulty)	HL: Head lights	(bright, dim, off)
CS : Charging System	(Okay, Faulty)	SP: Spark plugs	(okay, too_wide, fouled)
BA : Battery age	(new, old, very_old)	SQ: Spark Quality	(good, bad, very_bad)
BV: Battery voltage	(strong, weak, dead)	CC: Car cranks	(True, False)
MF: Main fuse	(okay, blown)	TM: Spark timing	(good, bad, very_bad)
DS: Distributor	(Okay, Faulty)	FS: Fuel system	(Okay, Faulty)
PV: Voltage at plug	(strong, weak, none)	AF: Air filter	(clean, dirty)
SM: Starter Motor	(Okay, Faulty)	AS: Air system	(Okay, Faulty)
SS: Starter system	(Okay, Faulty)	ST: Car starts	(True, False)

Table 2 "Lymphography" variables.

Variables' names	Possible states
V1: Lymphatics	(normal, arched, deformed, displaced)
V2: Block of affere	(no, yes)
V3: bl. of lymph. C	(no, yes)
V4: bl. of lymph. s	(no, yes)
V5: by pass	(no, yes)
V6: extravasates	(no, yes)
V7: regeneration of	(no, yes)
V8: early up take in	(no, yes)
V9: lym.nodes dimin	(0, 1, 2, 3)
V10: lym.nodes enla	(1, 2, 3, 4)
V11: changes in lym.	(bean, oval, round)
V12: defect in node	(no, lacunar, marginal, lac_central)
V13: changes in node	(no, lacunar, marginal, lac_central)
V14: changes in stru	(no, grainy, draplike, coarse, diluted, reticular, stripped, faint)
V15: special forms	(no, chalices, vesicles)
V16: dislocation	(no, yes)
V17: exclusion	(no, yes)
V18: no. of nodes	(1, 2, 3, 4, 5, 6, 7, 8)
VI: Diagnosis	(normal, metastases, malign_lymph, fibrosis)

4.2 Clustering

Regarding the clustering, we are going to use the stability function (bootstrap approach using the mean of corrected rand criterion) of the toolbox ClustOfVar [16] after the application of an hirarchical ascendant algorithm, in order to estimate, approximately, the number of clusters to be entered in the algorithm K-means.

Using the stability graphics, the optimal number of clusters selected, for "Car diagnosis 2" database, has been equal to three and the clustering result of variables is presented in "Table 3".

Table 3 Clustering results of the "Car diagnosis 2" data base.

Cluster 1	Cluster 2	Cluster 3
AL : Alternator	DS: Distributor	FS: Fuel system
CS : Charging System	TM: Spark timing	AF: Air filter
BA : Battery age		AS: Air system
BV: Battery voltage		
MF: Main fuse		
PV: Voltage at plug		
SM: Starter Motor		
SS: Starter system		
HL: Head lights		
SP: Spark plugs		
SQ: Spark Quality		
CC: Car cranks		

Using the stability graphics, the optimal number of clusters selected, for "Lymphography" database, has been equal to two and the variables clustering results is presented in "Table 4" below.

Table 4 Clustering results of the "Lymphography" data base.

Cluster 1	Cluster 2
V1, V8, V9, V10 V11, V12, V13, V14, V15, V16, V17, V18	V2, V3, V4 V5, V6, V7

4.3 The classical learning structure compared to our new heuristic

For the "Car diagnosis 2" database, "Figure 1" below depicts the classical structure learning result of the entire variables after applying the K2 algorithm, with as entry, the obtained order reached via the tree resulting from the implementation of the MWST algorithm (to note: we have chosen the interest variable as an initial variable during the application of the MWST algorithm). The execution time has been *3.45* seconds.

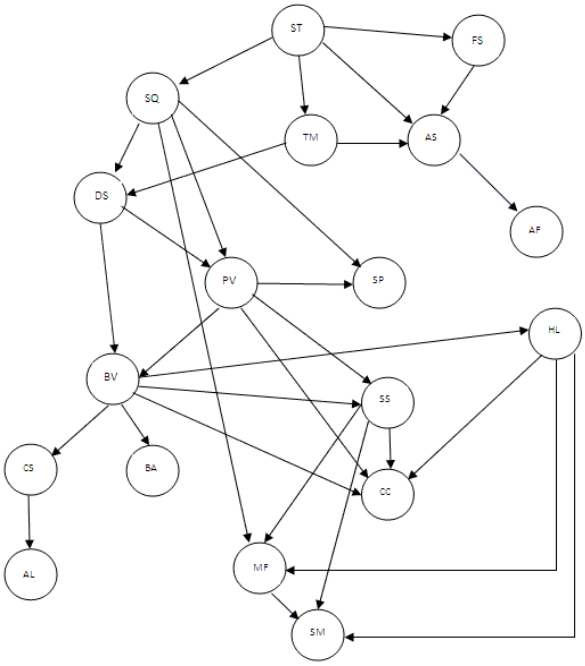

Figure 1 The classical structure learning result.

The Figures 2, 3 and 4, appearing below, depict the structures resulting from the learning structure pertaining to every cluster of variables after applying the K2 algorithm, with, as an entry, the order obtained from the MWST resulting tree (we have selected the interest variable as being the initial variable during the MWST algorithm application to each cluster). The final structure is automatically represented by reassembling the clusters' structures around the interest variable (see Figure 5). The global execution time has been *1.45* seconds (over *1.32* seconds for cluster 1; *0.05* seconds for cluster 2 and *0.09* seconds for cluster 3). The sum of these executions' time (*1.45* seconds) remains significantly inferior to the structure learning of the entire variables simultaneously, which equals *3.45* seconds.

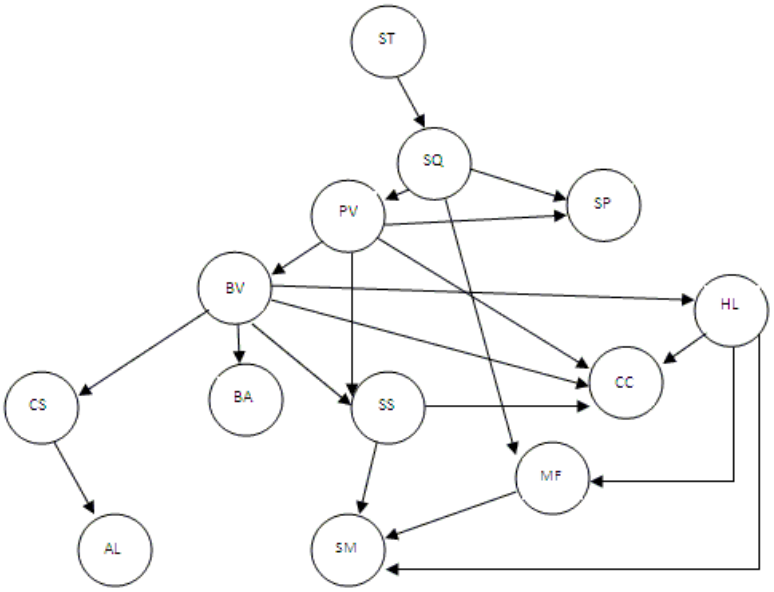

Figure 2 Cluster 1 structure.

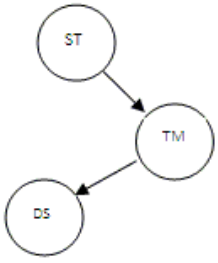

Figure 3 Cluster 2 structure.

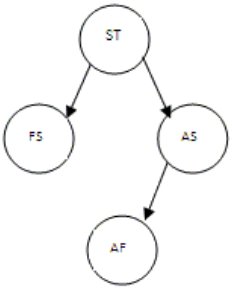

Figure 4 Cluster 3 structure.

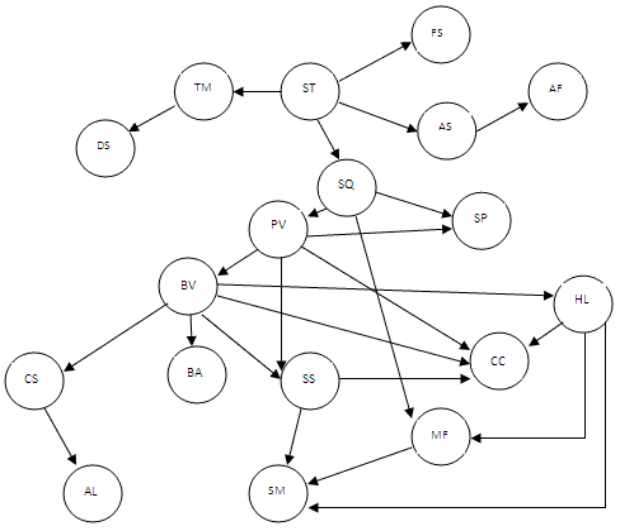

Figure 5 The ultimate Structure.

For the "Lymphography" database, the same treatment and the same algorithms are applied. The sum of learning structure of "cluster 1" and "cluster 2" executions' time (equal to *1.65* seconds) remains significantly inferior to the structure learning of the entire variables, simultaneously, which equals *2.67* seconds.

4.4 Both attained structures' relevant inferences and result comparisons

As our approach favors the preservation of data, principally for the interest variable's sake, we will learn the parameters of the two structures found for each of the databases studied (structure found after learning all the variables simultaneously and structure found after assembling the various structures of the clusters around the interest variables). As for the interest variable, we are going to calculate the probabilities of its different possible corresponding states, bearing in mind the states of the network's other nodes in respect of the two obtained BN structures. Thus, a thirty-two-example database will be used for experimenting the interest variables of both databases. Naturally, the experimentation examples have been excluded during the structures' learning. The differential statistical significance between the obtained probabilities, with respect to both structures, will be measured via the "*Z*" test (comparing the two observed means belonging to two different samples), according to the following formula: $Z = \frac{P1-P2}{\sqrt{variance(P1)+variance(P2)-2\times covariance(P1,P2)}}$ [24]

Hypothesis H_0: the difference between both probabilities is significant *($|Z|>1.96$)*.

HypothesisH_1: the difference between both probabilities is non-significant *($|Z|\leq 1.96$)*.

The two tested variables are "Car starts" of "Car Diagnosis 2" database and "Diagnosis" of "Lymphography" database. "Appendix B" contains two graphs

showing the variation of the *Z*-test for each variable studied according to its different possible states.

4.5 Discussion

Based on the achieved experimental results, the pairs of probabilities for the variable “diagnosis” of the "Lymphography" database are identical; the preservation of information has been complete (see Appendix B, Figure 7). As for the variable “Car Start” of “Car Diagnosis 2” database, the probabilities pairs are very similar but not identical; the hypothesis H_0 has always been rejected, even with very small Z values, not exceeding the value of $|0.46|$, very distant from the threshold of $|1.96|$, as set by the Z test theory (see Appendix B, Figure 6). It can, therefore, be deduced that the inference results, regarding both of learning structures approaches, are very similar even at eye sight, and without applying any statistical tests to measure the difference’s significance. Through our heuristic, we have managed to reduce, considerably, the algorithmic complexity of the BN structure learning without any significant loss of information, especially with regards to the interest variable. The clustering constancy and trustiness plays a determining role in the accuracy of the resulting structure. In fact, the more independent the obtained clusters are, the more the number of inter-cluster edges to be lost would shrink; consequently, the more independent the clusters are, the more negligible the lost information would be.

Throughout the present study, we have, firstly, demonstrated mathematically that the algorithmic complexity of the BN from data-base structure learning decreases dramatically in the cases when the variables’ subsets are treated in a separate way. In a second place, a heuristic has been proposed whereby the demonstrated conduct could be exploited by adding a solution serving to reassemble the sub-sets’ structures into a single structure framework. This solution has been based on the implementation of the information system’s interest variable as a linking variable among the subsets’ different structures. Through our proper experimentation procedure, we have proved that by implementing this undertaking, we can be immune against the information loss problem while achieving a considerable gain in terms of execution time. Our original solution has been improved; firstly because no criterion has been defined for the applicability of our approach on a certain database (possibility of having clusters sufficiently independent to avoid losing information). Secondly, the method applied for determining the optimal number of clusters is known to be greedy in computational complexity (in the order of $O(n^3)$). So, a heuristic, less complicated yet effective would be among our aim in future research. Inversely, however, with the help of our newly-devised concept, new large-scale horizons have been opened, paving the way for other more global solutions, taking advantage of the fact that the possible number of DAG decreases incredibly by treating the variables and subsets during the BN structure learning from data-base.

5 Conclusion

Within the scope of the present work, we have set up a new well-defined approach for the BN structure learning from data-base, so useful that it can be jointly applied with the already existing algorithms and underway heuristics. As a first step, we have demonstrated, mathematically, that the BN structures' possible space decreases, dramatically, by subdividing the relevant variables into clusters before processing the BN structure learning corresponding to each cluster apart. In the second step, a specially-devised heuristic has been proposed with the aim of joining each cluster's different structures. Actually, through a specially-conducted experimentation administered over tow data-bases, we have proved that loss in data turns out to be so negligible that it does not affect the extracted BNs stemming results during the inference stage, while saving a great deal of execution time.

In a potential future research, we reckon to make a serious attempt to investigate other possible alternatives, useful and fit to exploit the considerable reduction of algorithmic complexity during the BN structure learning by examining and treating variables' sub-sets, developing some structure-retrieving oriented heuristics, encompassing the already achieved sub-structures, a framework that would be the closest possible to the discovered structure, while simultaneously treating the whole set of variables in their entirety.

References

1. Nefian, V.: Learning SNP using embedded Bayesian Networks. IEEE Computational Systems Bioinformatics Conference. (2006)
2. Cooper G., Hersovits E.: A Bayesian method for the induction of probabilistic networks from data. Machine learning. 9, 309--347 (1992)
3. Neapolitan R. E.: Learning Bayesian Networks. Newyork USA Prentice Hall. (2003)
4. Spirtes, P., Glymour, C., Scheines R.: Causation, Prediction, and Search. The MIT Press, 2nd edition. (2000)
5. Shulin Y., Chang K.: Comparison of Score Metrics for Bayesian Network Learning. IEEE Transactions on Systems, Man and Cybernetics-part A: Systems and Humans. 32(3) 419--428 (2002)
6. Bouchaala L., Masmoudi A., Gargouri F., Rebai A.: Improving algorithm for structure learning in Bayesian Networks using a new implicit score. Expert Systems with Application. 37, 5470--5475 (2010).
7. Mourad, R. Sinoquet C., Leray P.: A hierarchical Bayesian Network approach for linkage disequilibrium modelling and data dimensionality reduction prior to genome-wide association studies. BMC Bioinformatics. 16, (2011)
8. Zhang Y., Ji L.: Clustering of SNPs by structural EM algorithm, in proceeding of International Joint Conference on Bioinformatics. Systems Biology and Intelligent Computing. 147--150 (2009)
9. Hwang K., Kim B.H., Zhang B.T.: learning hierarchical Bayesian Networks for large-scale data analysis. In ICONIP. 670--679 (2006)
10. Mourad, R. Sinoquet, C. Leray P.: Learning hierarchical Bayesian Networks for genome-wide association studies. 19th International Conference on computational statistics, COMPSTAT. 549--556 (2010)

11. Judea P., Tom V.: A theory of inferred causation. In James Allen, Richard Fikes and Erik Sandewall, editors, KR' 91, Principles of knowledge representation and reasoning. 441--452 (1991)
12. Robinson R. W.: Counting unlabeled acyclic digraphs. Combinatorial Mathematics. 622, 28--43 (1977)
13. Tufféry S.: Data mining et statistique décisionnelle: l'intelligence des données. Editions TECHNIP. (2010)
14. Jain A.K.: Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 31, 651--666 (2010)
15. Chavent M., Kuentz V., Liquet B., Saracco J.: ClustOfVar: an R package for the clustering of variables. The R user conference, University of Warwick Coventry UK. (2011)
16. Chavent M., Kuentz V., Saracco J.: A partitioning method for the clustering of categorical variables. In classification as a tool for Research, Herman locarek-Junge, claus Weihs (Eds), Springer, International Federation of Classification Societies Conference (2009)
17. Lerman I.C.: Likelihood linkage analysis (LLA) classification method : An example treated by hand. Biochimie. 75, (5), 379--397 (1993)
18. Spirtes P., Glymour C., Scheines R.: Causation prediction and search. (1993)
19. Chow C., Liu C.: Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory. 14, (3), 462--467 (1968)
20. Francois O., Leray P.: Evaluation d'algorithmes d'apprentissage de structure pour les réseaux bayésiens. 14ème Congrès Francophone Reconnaissance des Formes et Intelligence Artificielle. 1453--1460 (2004)
21. Heckerman D., Geiger D., Chickering M.: Learning Bayesian Networks: The combination of knowledge and statistical data. 10th conference on uncertainty in artificial intelligence. 293--301(1994)
22. Kruskal J.: On the shortest spanning subtree of a graph and traveling salesman problem. The American Mathematical Society. Vol 7, 48--50 (1956)
23. Murphy K.: The BayesNet Toolbox for Matlab. Computing Science and Statistics: Proceedings of Interface. 33, http ://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html.
24. Sprinthall R.C., Basic Statistical Analysis. Seventh Edition. (2003)

Appendix A

Proof of Proposition 1 (i)

For all $n \geq 2$, we have:

i) $r(n) \geq 2^{n-2} n\, r(n-1)$

Proof by induction on n

For $n=2$; $r(2)= 3 \geq 2$ is verified

Tel $n \in IN$, we assume that $\forall\, i \leq n;\ r(i) \geq 2^{i-2}\ i\ r(i-1)$. (1)

By applying Robinson formula, we have:

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

We set $Vi = 2^{i(n+1-i)} \binom{n+1}{i} r(n+1-i)$

We will firstly prove that $(V_i)_{1 \leq i \leq n+1}$ is decreasing

$$\frac{Vi}{Vi+1} = \frac{2^{i(n+1-i)} \binom{n+1}{i} r(n+1-i)}{2^{(i+1)(n+1-i-1)} \binom{n+1}{i+1} r(n-i)}$$

$$= \frac{2^{i} \frac{1}{(n+1-i)!}\, r(n+1-i)}{2^{n-i} \frac{1}{(i+1)(n-i)!}\, r(n-i)}$$

By using (1),

$\frac{Vi}{Vi+1} \geq \frac{2^{2i-n}(i+1)2^{n+1-i-2}\,(n+1-i)r(n-i)}{(n+1-i)}$

Which imply, $\frac{Vi}{Vi+1} \geq 2^{i-1}(i+1) > 1$. This means that $(V_i)_{1\leq i\leq n+1}$ is decreasing.

Secondly we will prove that $r(n+1) - 2^{n-1}(n+1)\,r(n) \geq 0$

Observe that,

$2^{n-1}(n+1)\,r(n) - 2^{2(n-1)}\binom{n+1}{2} r(n-1) \geq$

$2^{n-1}(n+1)2^{n-2}n\,r(n-1) - 2^{2(n-1)}\frac{(n+1)n}{2} r(n-1) \geq$

$2^{2n-3}(n+1)n\,r(n-1) - 2^{2n-3}(n+1)n\,r(n-1) = 0$

So, $2^{n-1}(n+1)\,r(n) + 2^{2(n-1)}\binom{n+1}{2} r(n-1) \leq 2 \times 2^{n-1}(n+1)\,r(n) = 2^n(n+1)\,r(n)$

=> The sum of the tow first elements of $r(n+1)$ minus $2^{n-1}(n+1)\,r(n)$ is positive

=> $r(n+1) - 2^{n-1}(n+1)r(n) = $ *positive element* $+ S$; where $S = \sum_{i=3}^{n+1}(-1)^{i+1}Vi$

Thirdly we will prove that $S \geq 0$

Case 1: n is impair: $n=2p-1;\ p \in IN^*$

$S=\sum_{i=3}^{n+1}(-1)^{i+1}Vi$

$=\sum_{i=3}^{2p}(-1)^{i+1}Vi$

$=\sum_{1\leq j\leq p}(-1)^{2j+1}V_{2j} + \sum_{1\leq j\leq p-1}(-1)^{2j+2}V_{2j+1}$

(Where in the first sum $i=2j$ and in the second $i=2j+1$)

$=\sum_{1\leq j\leq p-1} V_{2j+1} - \sum_{2\leq j\leq p} V_{2j}$

$=\sum_{1\leq j\leq p-1} V_{2j+1} - \sum_{2\leq j\leq p-1} V_{2j+2} = \sum_{j=1}^{p-1}(V_{2j+1} - V_{2j+2})$. Since $(V_i)_{1\leq i\leq n+1}$ is decreasing we can conclude that $\sum_{j=1}^{p-1}(V_{2j+1} - V_{2j+2}) \geq 0$

Case 2: n is pair: $n=2p$

$S=\sum_{i=3}^{n+1}(-1)^{i+1}Vi$

$=\sum_{i=3}^{2p+1}(-1)^{i+1}Vi$

$=\sum_{1\leq j\leq p}(-1)^{2j+1}V_{2j} + \sum_{2\leq j\leq p+1}(-1)^{2j}V_{2j-1}$

(Where in the first sum $i=2j$ and in the second $i=2j-1$)

$=\sum_{2\leq j\leq p}(V_{2j-1} - V_{2j}) + V_{2p+1}$. Since (V_i) is decreasing and $V_{2p+1} \geq 0$ then $\sum_{2\leq j\leq p}(V_{2j-1} - V_{2j}) + V_{2p+1} \geq 0$

Therefore $S \geq 0$, then

$r(n+1) - 2^{n-1}(n+1)r(n) \geq 0$, then

$r(n+1) \geq 2^{n-1}(n+1)r(n)$; Which proves the proposition.

Proof of Proposition 1 (ii)

ii) $r(n) \geq 2^{\frac{(n-1-J)(n+J-2)}{2}} n(n-1)\dots(J+2) \times r(J+1)\ \forall\ (1+J) \leq n$

Proof:

By using i) of proposition 1, we have the desired result.

$r(n) \geq 2^{n-2} \times 2^{n-3} \times \dots \times 2^J n(n-1)\dots(J+2) \times r(J+1)$

$r(n) \geq 2^{\frac{(n-1-J)(n+J-2)}{2}} n(n-1)\dots(J+2) \times r(J+1)$

where $(n-2)+(n-3)+\dots+J = \frac{(n-1-J)}{2}(J+n-2)$

We can conclude that our proposition 1 (ii) is confirmed.

Appendix B

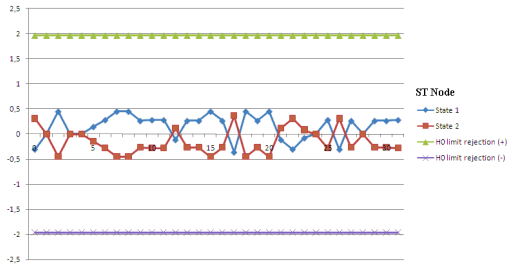

Figure 6 *Z*-test variation for the "Car starts" variable ("Car Diagnosis 2" database).

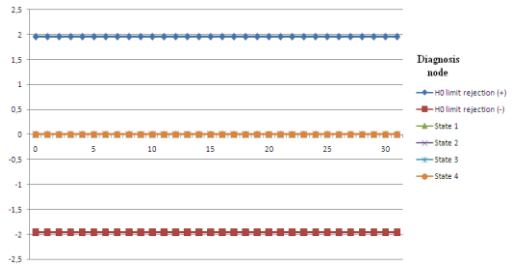

Figure 7 *Z*-test variation for the "Diagnosis" variable ("Lymphography" database).