2009 Ninth International Conference on Intelligent Systems Design and Applications

Improved Learning of Bayesian Networks in Biomedicine

Antonella Meloni, Luigi Landini
Department of Information Engineering
University of Pisa
Pisa, Italy
antonella.meloni@iet.unipi.it, llandini@ifc.cnr.it

Andrea Ripoli, Vincenzo Positano
"G. Monasterio" Foundation
CNR
Pisa, Italy
ripoli@ifc.cnr.it, positano@ifc.cnr.it

***Abstract*— Bayesian Networks represent one of the most successful tools for medical diagnosis and therapies follow-up. We present an algorithm for Bayesian network structure learning, that is a variation of the standard search-and-score approach. The proposed approach overcomes the creation of redundant network structures that may include non significant connections between variables. In particular, the algorithm finds which relationships between the variables must be prevented, by exploiting the binarization of a square matrix containing the mutual information (MI) among all pairs of variables. Four different binarization methods are implemented. The MI binary matrix is exploited as a pre-conditioning step for the subsequent greedy search procedure that optimizes the network score, reducing the number of possible search paths in the greedy search. Our approach has been tested on two different medical datasets and compared against the standard search-and-score algorithm as implemented in the DEAL package.**

Keywords-stuctural learning; bayesian network;biomedicine

I. INTRODUCTION

Bayesian Networks are used to represent knowledge about an uncertain domain [1] and they have emerged as one of the most successful tools for medical diagnosis, selection of optimal treatment alternatives and prediction of treatment outcome [2].

A Bayesian network (BN) is a graphical model that represents a joint probability distribution over a set of random variables [3] and it is defined by a pair B={G, P}. The network structure G is a directed acyclic graph (DAG) whose nodes represent random variables and whose edges represent direct dependencies among the variables and are drawn by arrows between nodes. The second component P is a set of numerical parameters, which represent conditional probability distributions.

In many practical settings the BN is unknown and its characteristics should be learned from the data. The learning task in a BN can be separated into two subtasks: *structural learning,* that is to identify the topology of the network, and *parameter learning,* that finds the numerical parameters for a given network topology. Our work focuses upon structural learning.

Structural learning of Bayesian Networks can be performed by using the score-and-search approach, that has been first implemented in R in the package DEAL [4]. However, this method often converges to a redundant network that may include arcs associated with variable couples not linked by a significant relationship [5].

In this paper we present a new method, based on the inclusion of the mutual information metric in the search and score strategy, able to prevent the inference of too many arcs. The developed method has been tested on two validated medical databases.

II. STRUCTURAL LEARNING

The score-and-search-based approach attempts to find a graph that maximizes the selected score or metric, which evaluates how well a given network matches the data. The BDe (Bayesian with Dirichlet prior and Equivalence) metric [6] has been used in this study. The network score is its posterior probability given the database. It can be efficiently calculated in closed form under the following five assumptions: multinomial sample, parameter independence, parameter modularity, complete data, and likelihood equivalence. Likelihood equivalence when combined with parameter independence implies Dirichlet assumption: all network parameters have a Dirichlet distribution.

The greedy search with random restarts [6] is used as the strategy for searching for DAGs with higher score. Greedy search starts at a specific point (a structure without any arcs). After, the algorithm considers all neighbors of the current point, and moves to the neighbor that has the highest score. The neighbors are the structures that can be generated from the current structures by adding, deleting or reversing a single arc, subject to the acyclicity constraint. If no neighbors have higher score than the current point, the algorithm stops. The application of random restarts allows to solve the problem of the premature convergence to local maxima [7]. The search is run until an optimum is reached. Then a new initial state is randomly chosen and the algorithm is run again. After n iterations the best solution is sought.

III. LEARNING PRE-CONDITIONING BY MUTUAL INFORMATION

A well recognized limit of the previously described method is that it has a tendency to find too many arcs among the variables. In fact, the greedy search will add arcs to the network structure even if the contribute of the arc to the global value of the metric is very low and does not represent a real relationship established among variables. A new approach has been developed in order to overcome this drawback. First it requires the computation of the mutual

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.163

information (MI) among all pairs of variables. Mutual information measures the general dependence of random variables without making any assumptions about the nature of their underlying relationships [8]. The mutual information between the variables X_i and X_j is then defined as:

$$MI(X_i, X_j) = H(X_i) + H(X_j) - H(X_i, X_j) \geq 0 \quad (1)$$

where $H(X_i)$ represents the Shannon entropy of the empirical probability distribution [9]. Assume that the variable Xi has M possible states $x_{i1}, \ldots, x_{iM}$, each with its corresponding probability $p(x_{im})$, then the entropy can be calculated as:

$$H(X_i) = -\sum_{m=1}^{M} p(x_{im}) \log p(x_{im}). \quad (2)$$

The logarithm in the Equation (2) refers to the natural logarithm. The joint entropy $H(X_i, X_j)$ of two discrete variables X_i and X_j is defined analogously as:

$$H(X_i, X_j) = -\sum_{m=1}^{M} \sum_{l=1}^{L} p(x_{im}, x_{jl}) \log p(x_{im}, x_{jl}) \quad (3)$$

Here $p(x_{im}, x_{jl})$ denotes the joint probability that X_i is in state x_{im} and X_j is in state x_{jl}, calculated from a multivariate histogram. The number of possible states M and L may be different.

A mutual information matrix (MIM) can be computed as a square matrix whose i,j element is the mutual information between X_i and Xj.

MI is zero if X_i and X_j are statistically independent and increases the less statistically independent X_i and X_j are. In practice, since MI is always non-negative, its evaluation from random samples may give a positive value even for variables that are, in fact, mutually independent. Moreover, entropy estimation based on relative frequencies has several sources of error, such as finite number of observations [10].

Hence, our proposed approach, in its second step, finds out the significant relationships and returns as outcome a binarized MIM in which the ones represent those links. Four different binarization methods are implemented. The first one uses a threshold and the others have been taken by the field of reverse engineering, because widely used to infer genetic networks to microarray data.

Finally, the binarized MIM is used to establish which network structures are acceptable. If the element *i,j* in the matrix is equal to 0, an arc between the two correspondent variables X_i and X_j will be not allowed in the greedy search algorithm. Consequently, the DAG which contains any of these arcs will be disregarded in the search procedure.

A. MI thresholding

The elements of the MIM larger than the threshold I_0 are transformed to state 1 and the elements smaller than I_0 are transformed to state 0. Different threshold values were experimented based on the percentiles of the MI distribution.

The MIM is symmetric and the upper triangle is extracted. A vector constituted by the selected elements is created and the percentiles (10^{th}, 15^{th}, 20^{th}, 25^{th}, 30^{th}, 35^{th}, 40^{th}, 45^{th}, 50^{th}, 55^{th}, 60^{th}) of its distribution are calculated.

The percentiles are the 100-quantiles, namely the quantiles expressed as percentage. The quantiles are calculated using the algorithm type 8 discussed in Hyndman and Fan [11]. Using this algorithm, the resulting quantile estimates are approximately median-unbiased regardless of the distribution of data whose sample quantiles are wanted. Each calculated percentile has been tested as threshold.

B. CLR method

The CLR (Context Likelihood of Relatedness) algorithm [12] derives a score related to the empirical distribution of MI values. For each couple of variables X_i and X_j, it takes into account the score:

$$s_{ij} = \sqrt{s_i^2 + s_j^2} \quad (4)$$

where:

$$s_i = \max(0, \frac{MI(X_i, X_j) - \mu_i}{\sigma_i}) \quad (5)$$

and μ_i and σ_i are, respectively, the mean and the standard deviation of the empirical distribution of the mutual information values $MI(X_i, X_k)$, with $k = 1, \ldots, n$. A square and symmetric matrix whose i,j element is sij, the score of the pair $\{X_i, X_j\}$ is obtained and subsequently binarized by assigning 0 at all the null elements in the matrix and 1 at the remaining ones.

C. ARACNE method

ARACNE (algorithm for the reconstruction of accurate cellular networks) uses a well-known information theoretic property: the data processing inequality (DPI) [13]. The DPI [8] states that if two variables X_i and X_z interact only through a third variable, X_j, then:

$$MI(X_i, X_z) \leq \min(MI(X_i, X_j), MI(X_j, X_z)) \quad (6)$$

A weight equal to their mutual information is assigned to each pair of nodes. Then, the algorithm examines each variables triplet and removes the edge with the smallest value, interpreted as an indirect interaction. A square and symmetric matrix, containing all MIs for pairs of variables considered directly interacting and 0 otherwise, is computed. The matrix is binarized by assigning 0 at all the null elements and 1 at the remaining ones.

D. MRMR method

The MRMR (Maximum relevance minimum redundance) method allows to select variables in a stepwise mode so that each new variable selected has the highest individual MI with the output (maximum relevance) and the lowest

possible average MI with the preselected variables (minimum redundancy) [14].

In a supervised learning task, the output is denoted by Y and V represents the set of input variables. The greedy search starts by selecting the variable X_i that has the highest mutual information to the target Y. After it selects the variable X_j that has a high information $MI(X_j, Y)$ to the target and at the same time a low information $MI(X_j, X_i)$ to the previously selected variable. In the following steps, given a set S of selected variables, the method updates S by choosing the variable X_j that maximizes the score:

$$S_j = MI(X_j;Y) - \frac{1}{|S|}\sum_{X_k \in S} MI(X_j;X_k) \quad (7)$$

This selection procedure is repeated considering at every turn a different variable as target.

For each pair $\{X_i, X_j\}$, MRMR returns, according to (5), two (not necessarily equal) scores s_i and s_j. The score of the pair is computed by taking the maximum between s_i and s_j. A square and symmetric matrix whose i,j element is the score of the pair $\{X_i, X_j\}$ can be computed. The matrix is binarized by assigning 0 at all the null elements in the matrix and 1 at the remaining.

IV. Performance Analysis

The proposed metrics, hereafter called "S&S (search and score) + threshold", "S&S + CLR", "S&S + ARACNE" and "S&S + MRMR", have been numerically investigated by means of two medical datasets, widely used in the literature: ASIA (Fig. 1) and CANCER (Fig. 2). The ASIA network, introduced by Lauritzen and Spiegelhalter [15] is a small network constituted by 8 discrete variables and 8 arcs. The CANCER network includes 5 discrete variables and 5 arcs [16]. Each network has been used to generate several databases by means of probabilistic logic sampling method [17]. The sample sizes considered for ASIA network are N = 1000, 5000 and 10000. The sample sizes considered for CANCER network are N = 1000, 2500 and 5000.

The results have been compared to those obtained by using the standard structural learning procedure described in subsection II. The algorithm has been built on the top of the package DEAL [4].

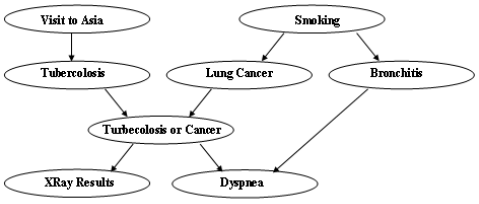

Figure 1. The ASIA network

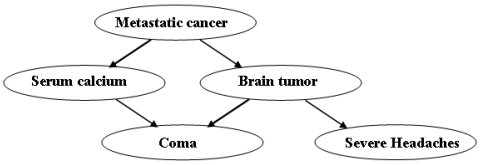

Figure 2. The CANCER network

Different criteria have been selected to gauge the quality of the reconstructed structure: the number of correct edges, the number of extra edges, the number of missing edges and the F-score. The F-score is a weighted harmonic average of precision (p) and recall (r), expressed as [18]:

$$Fscore = \frac{2 * p * r}{p + r} \quad (8)$$

The precision measures the fraction of real edges (present in the real network) among the ones inferred by the algorithm and the recall, also known as true positive rate, denotes the fraction of real edges that are correctly inferred.

A. Calculation of the best threshold

First, the dependence of the algorithm S&S + threshold (defined also as S&S + T) results on the threshold values has been investigated. For both networks, the algorithm has been repeated considering all possible threshold values for each sample size. Figure 3 and 4 show the F-scores obtained for ASIA and CANCER networks respectively.

For both networks, regardless of the sample size, by increasing the threshold, the performances of the algorithm tend to improve or, sometimes, to hold steady. Moreover, by increasing the sample size, a best score can be obtained. The best threshold is the 50th percentile. As can been seen, by using the 50th percentile as threshold and a sample size of at least 2500, the true structure of the CANCER network can been inferred.

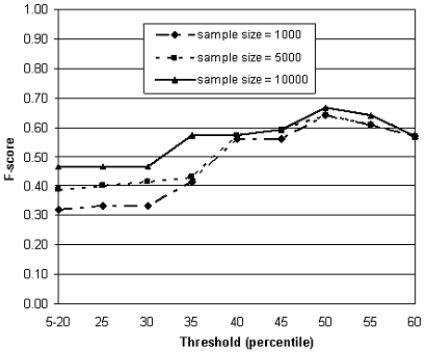

Figure 3. F-scores of the method S&S + T by varying the MI thresholds for the ASIA network. For all sample size, the 5th, 10th, 15th and the 20th percentiles are the same.

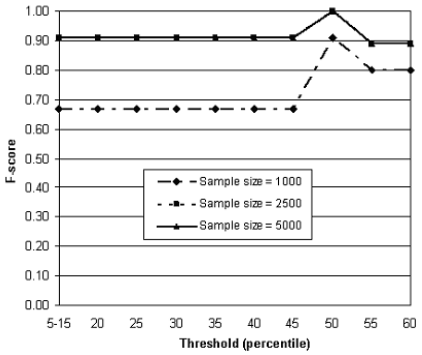

Figure 4. F-scores of the method S&S + T by varying the MI thresholds for the CANCER network. For all sample size, the 5th, 10th and the 15th percentiles are the same.

B. Comparison among the algorithms

Results of the learned network using the standard score-and-search algorithm implemented in DEAL and our variations are shown in Table 1 for ASIA network and in Table 2 for CANCER network. For the MI thresholding algorithm, only the outcomes obtained by using the best threshold evaluated in the previous paragraph are presented.

The general trend for all the algorithms is that the learned networks by them are more and more accurate as the size of the datasets gradually enlarges. For the CANCER network, our methods allows to find the true network if the sample size is major than 2500.

Our implemented variations improve the overall performance of the metric implemented in DEAL.

For the ASIA network, the application of the ARACNE method to binarize the MIM gives better results than the other approaches. For the CANCER network, there isn't difference between the four binarization methods in term of capability to find a good network.

TABLE I. EXPERIMENTAL RESULTS OF THE ALGORITHMS ON ASIA NETWORK BY VARYING THE SAMPLE SIZE.

Performance	Algorithm	Sample size 1000	5000	10000
Correct arcs	DEAL	4	5	7
	S&S+T	8	8	8
	S&S+CLR	8	8	8
	S&S+ARACNE	8	8	8
	S&S+MRMR	8	8	8
Incorrect added arcs	DEAL	18	18	15
	S&S+ T	9	9	8
	S&S+CLR	1	1	1
	S&S+ARACNE	1	0	0
	S&S+MRMR	4	2	1
Missing arcs	DEAL	4	3	1
	S&S+T	0	0	0
	S&S+CLR	0	0	0
	S&S+ARACNE	0	0	0
	S&S+MRMR	0	0	0
F-score	DEAL	0,27	0,32	0,47
	S&S+T	0.64	0.64	0.67
	S&S+CLR	0.94	0.94	1
	S&S+ARACNE	0.94	1	0.94
	S&S+MRMR	0.80	0.89	0.94

TABLE II. EXPERIMENTAL RESULTS OF THE ALGORITHMS ON CANCER NETWORK BY VARYING THE SAMPLE SIZE.

Performance	Algorithm	Sample size 1000	2500	5000
Correct arcs	DEAL	4	5	5
	S&S+T	5	5	5
	S&S+CLR	5	5	5
	S&S+ARACNE	5	5	5
	S&S+MRMR	5	5	5
Incorrect added arcs	DEAL	3	1	1
	S&S+T	1	0	0
	S&S+CLR	1	0	0
	S&S+ARACNE	1	0	0
	S&S+MRMR	1	0	0
Missing arcs	DEAL	1	0	0
	S&S+T	0	0	0
	S&S+CLR	0	0	0
	S&S+ARACNE	0	0	0
	S&S+MRMR	0	0	0
F-score	DEAL	0.67	0.91	0.91
	S&S+T	0.91	1	1
	S&S+CLR	0.91	1	1
	S&S+ARACNE	0.91	1	1
	S&S+MRMR	0.91	1	1

V. CONCLUSIONS

In this paper we have defined a new algorithm for Bayesian network structure learning, that is an evolution of the standard score-and-search-based approach. The algorithm first reconstructs a sort of skeleton of a Bayesian network, by finding the only arcs admitted, and then performs the Bayesian-scoring greedy search to infer the best network. This network doesn't have arcs classified as impossible. Our metric has been tested on two different medical datasets and compared against the standard score-and-search algorithm as implemented in the DEAL package. Our algorithm outperforms that metric and the successful numerical findings suggest that it could be very useful in medical domains

A possible limit of this work is that we have used 2 simple datasets, whereas in the domain of BioMedicine, BNs can be used for applications that involve large and complex network structures. Unfortunately, computational constraints forbid wide numerical testing in large networks using the R environment. In this first phase we are more interested in the methodology, but we will implement the code under other programming languages, in order to perform learning of larger networks.

REFERENCES

[1] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1988.

[2] D. J. Spielgelharter, "Probabilistic Expert Systems in Medicine," Statistical Science, vol. 2, pp. 3-44, 1987.

[3] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, vol. 29, pp. 131-163, 1997.

[4] S. G. Bøttcher, and C. Dethlesfen, "DEAL: A package for Learning Bayesian Networks," Journal of Statistical Software, vol. 8, pp. 1-40, 2003.

[5] S. G. Bøttcher, Learning Conditional Gaussian Networks. Tecnical Report R2005-22, Aalborg University, Denmark, 2005.

[6] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," Machine Learning, vol 20, pp. 197-243, 1995.

[7] M. Chickering, D. Geiger, and D. Heckerman, "Learning bayesian networks: Search methods and experimental results," Preliminary papers of the 5th Intl. Workshop on Artificial Intelligence and Statistics, 1995, pp. 112–128.

[8] T. M. Cover, and J. A. Thomas, Elements of Information Theory. John Wiley & Sons, New York, 1991.

[9] C. E. Shannon, "A mathematical theory of communication,", The Bell System Technical Journal, vol. 27, pp. 623-656, 1948.

[10] F. Bazsó, L. Zalányi, and A. Petroczi, "Accuracy of joint entropy and mutual information estimates," Proc. IEEE International Joint Conference on Neural Networks, 2004, pp. 2843-2846, doi: 10.1109/IJCNN.2004.1381108.

[11] R. J. Hyndman, and Y. Fan, "Sample quantiles in statistical packages," American Statistician, vol. 50, pp. 361–36, 1996.

[12] J. J. Faith, B. Hayete, J. T. Thaden, et al., "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," PLoS Biology, vol. 5, 2007, doi:10.1371/journal.pbio.0050008.

[13] A. A. Margolin, I. Nemenman, K. Basso, et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networksin a mammalian cellular context," BMC Bioinformatics, vol. 7, supplement 1, p. S7, 2006.

[14] G. D. Tourassi, E. D. Frederick, M. K. Markey, et al., "Application of the mutual information criterion for feature selection in computer-aided diagnosis," Medical Physics, vol. 28, pp. 2394–2402, 2001.

[15] S. L. Lauritzen, and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application on expert systems," Journal of the Royal Statistical Society, vol. 50, pp. 157–224, 1988.

[16] R. E. Neapolitan, Probabilistic reasoning in Expert Systems, Theory and Algorithms. John Wiley & Sons, New York, 1990.

[17] M. Henrion, "Propagating uncertainty in bayesian networks by probabilistic logic sampling," Uncertainty in artificial intelligence, vol. 2, pp.149–163, 1988.

[18] P. E. Meyer, K. Kontos, F. Lafitte, et al., "Information-theoretic inference of large transcriptional regulatory networks," EURASIP Journal on Bioinformatics and Systems Biology, 2007, doi:10.1155/2007/79879.