International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 6, Volume 2 (June 2015) www.ijirae.com

L1 Variable Selection Based Causal Structure Learning

Rania N. Elkhateeb
Faculty of Science
Sohag University

Mahmoud A. Mofaddel
Faculty of Science
Sohag University

Marghany H. Mohamed
Faculty of computers & Information
Assiut University

Abstract- ***Causal graphs are the most suitable representation of causal relationships between variables in a complex system. Directed acyclic graphs (DAGs), also known as Bayesian networks, are the well known and frequently used to represent these causal relationships. Learning the structure of causal DAGs from data is very important in applications to various fields, such as medicine, artificial intelligence and bioinformatics. Recently L1-regularization technique has been used for learning causal DAGs model structure. In this paper we propose an improved version of Grow-Shrink algorithm (GS), presented by Margaritis 2000, where we used L1 variable selection for Markov blanket selection step. Then we compare the performance of the proposed method with the state of the art ones: PC (Peter & Clark), and GS algorithms. Finally, we discuss and analyze the results.***

Keywords: Bayesian networks; causal discovery; graphical models; Markov blanket; regularization.

I. INTRODUCTION

Bayesian networks (*DAGs*) are the most convenient graph for representing causal relationships; because they are directional thus are capable of displaying relationships clearly and intuitively. A *DAG* consists of a set of nodes N where each node i represents a variable X_i receives directed edges from its set of parent nodes P_i so They can be used to represent both direct and indirect causation and handle uncertainty though the established theory of probability [1].

The last decade witnessed great interest in learning causal Bayesian networks from observational data because it can be used to automatically construct Decision Support Systems. In addition learning Bayesian networks have been used in several applications, for example, in bioinformatics for the interpretation and discovery of gene regulatory pathways, variable selection for classification, designing algorithms that optimally solve the problem under certain conditions, information retrieval and natural language processing [3].

The algorithms for learning causal Bayesian networks can be classified into three main types. The first, constraint-based algorithms use tests to detect relationships among variables for example *SGS* and *PC* algorithms [7]. The second, score-based algorithms select the structure that has the highest score of a function that measures how well the structure fits the data for example the *GES* algorithm [8]. The third is hybrid algorithms that utilize both score-based and constraint-based methods to construct the graph for example *MMHC* (Max-Min hill climbing) [12].

L1 regularization has been used first to learn the structure in probabilistic undirected graphical models. However, it is more preferable to use it in learning directed acyclic graphs (*DAGs*); because *DAGs* are more efficient for computing joint probabilities, samples, and (approximate) marginal's, the likelihood in *DAG* models factorizes into a product of single variable conditional distributions in contrast to directed acyclic graphs where estimating single-variable conditional distributions is used as an approximation and *DAGs* allow mixing different types of variables in a straightforward way.

Some algorithms like *GS* [2] and *CS* (Collider Set) [4] use Markov blanket approach to improve scalability by using information obtained from feature selection algorithms. *GS* algorithm refers to the first class (constraint-based methods) addresses the disadvantages of this category, exponential execution times and proneness to errors in dependence tests used, by identifying the Markov blanket of each variable in the Bayesian net as a preprocessing step. This preprocessing step handles the exponential time to be polynomial under the assumption of bounded Markov blanket size.

In this paper we use *L1* regularization technique for feature selection [5] in the preprocessing step to improve the existing *GS* algorithm, because adding regularization to a learning algorithm avoids over fitting. *L1* norm is advantageous because it encourages the sparse solutions which in order reduce the computational requirements.

A background and the problem statement are shown in section 2. Then the *GS* algorithm is discussed in section 3. After that the proposed algorithm *L1GS* is described in section 4. The experimental results of applying *L1GS*, *GS* and *Pc* algorithms to a set of observational datasets and a comparison between them is shown in section 5. Finally, section 6 provides the conclusion and hints to future work.

II. BACKGROUND AND PROBLEM STATEMENT

Relatively recently (1980's), the idea of inferring causal relations from observational data took place in many practical cases [10, 11]. Since then, several algorithms have been developed to infer such causal relations which greatly reduce the number of experiments required to discover the causal structure. In general learning the causal structure [6, 7] is a multivariate data analysis problem that aims at constructing a directed acyclic graph (*DAG*) presents the direct causal relationships among the interesting variables of a given system.

The last decade witnessed more interest in local learning. Markov blanket learning methods were first proposed by Koller & Sahami (1996) and Cooper (1997) introduced incomplete local causal methods. The local causal discovery methods (direct cause and effect) were first introduced by Aliferis & Tsamardinos, 2002a and Tsamardinos et al., 2003b [13].

Using Markov blanket (*Mb*) definition , *Mb*(*X*) is a minimal variable subset conditioned on which all other measured variables are probabilistically independent of *X*, in causal structure learning achieves promising results, since it was first proposed by Margaritis (2000) [1]. This approach is quite efficient since it extracts the Markov blanket information for each variable $X \in V$ from observational data, *V* is a set of variables, and then constructs a *DAG* graph from it. Another attempt for using this technique was proposed by Pellet& Elisseeff (2008) [9].

The Markov blanket for each variable $X \in V$ is the smallest set containing all variables carrying information about *X* that cannot be obtained from any other variable, in a causal graph this is the set of all parents, children, and spouses of *X*, fig. (1) Presents an example of *Mb*(*X*). Formally, in the context of a faithful causal graph *G* we have: $X \in Mb(Y) \Leftrightarrow Y \in Mb(X)$. An important property of Markov blanket is that *Mb*(*X*) is unique in a faithful *BN* or a *CPN*.

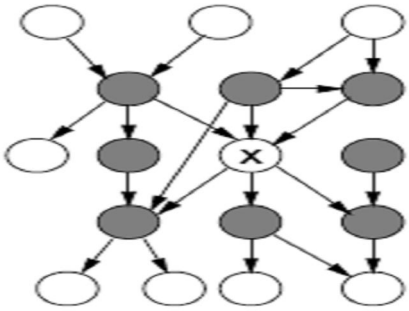

Fig. 1 An example of Markov blanket of variable *X*

As mentioned before *DAG* models are one of the best ways to model the joint distribution $P(X_1, X_2, \ldots, X_n)$ of a set of *n* random variables. If we repeatedly use the definition of conditional probability, $P(X, Y) = P(Y \mid X)\, P(X)$, in the order *n* down to 1, then we obtain the factorization of the joint distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid X_{1:i-1})$$

This factorization of the joint distribution is valid for any probability distribution.

The main goal of this paper is to use the Markov blanket approach to learn the structure of *DAG* that encodes causal relationships among a set of interesting variables *V* from observational data *D*. We propose an improved version of *GS* algorithm where we use *L1*regularization for feature selection to select the Markov blanket for each interesting variable in a preprocessing step before starting the causal structure learning phase. We call the proposed algorithm as *L1GS*.

III. GROW-SHRINK ALGORITHM (GS)

It is the first published sound Markov blanket induction algorithm that was proposed by Margirits & Thrun [2]. It was introduced with the intent to induce the Markov blanket for the purpose of speeding up global network learning. The whole algorithm can be split into two phases; the first: finds the Markov blanket for each variable which greatly facilitate the recovery of the local structure around each node, the second: performs further conditional-independence tests around each variable to infer the structure locally and uses a heuristics to remove cycles possibly introduced by previous steps. Table (1) lists the steps of the *GS* responsible for building the local structure using the Markov blanket information, *DN*(X) represents the direct neighbors to *X*. the complexity of step 1 for exploring Markov blanket is *O* (*n*) in the number of independence tests.

The *GS* algorithm passes each variable and its Markov blanket members twice. In the first pass it removes the possible spouse links between linked variables *X* and *Y* by looking for a d-separating set around *X* and *Y* (set *Z* d-separates *X* and *Y* if every path from *X* to *Y* is blocked by *Z*). In the second pass, the *GS* orients the arcs whenever it finds that conditioning on a middle node creates a dependency. While searching for the appropriate conditioning set, *GS* selects the smallest base search set (set *B* in table 1) for each phase. The search for the smallest set *B* has two advantages. The first, it reduces the number of tests, which is desirable because each phase contains a subset search, exponential in time complexity with respect to the searched set. The second advantage is reducing the average size of the conditioning set, which increases the power of the statistical tests, and thus helps reduce the number of type II errors [9].

It is obvious from table (1) that the whole algorithm requires $O(n^2 + nb^2 2^b)$ conditional independence tests, where $b=\max_x(|B(X)|)$. If we assume that *b* is bounded with a constant then the algorithm is $O(n^2)$ in the number of conditional independence tests.

The main core and advantage in *GS* algorithm is using a Markov blanket technique before starting the structure learning; because it restricts the size of the conditioning sets. We concentrate on this step to improve the *GS* performance and choose another Markov blanket algorithm *L1MB*; it will be shown in details in the next section.

TABLE I
PLAIN STEPS FOR *GS* ALGORITHM

1. compute Markov Blanket Mb.
   $\forall X \in V$ compute $Mb(X)$
2. Compute Graph structure.
   $\forall X \in V$ & $Y \in Mb(X)$
   determine Y to be a direct neighbour of X if X & Y are independent given S $\forall S \subset T$, where T is the smaller of $Mb(X) - \{Y\}$ and $Mb(Y) - \{X\}$.
3. Orient Edges.
   $\forall X \in V$ & $Y \in DN(X)$, orient $Y \rightarrow X$ if $\exists Z \in DN(X) - DN(Y) - \{Y\}$ such that Y & Z are independent given $S \cup \{X\}$ $\forall S \subseteq U$, where U is the smaller of $Mb(Y) - \{Z\}$ & $Mb(Z) - \{Y\}$.
4. Remove Cycles
   Do the following while there exist cycles in the graph :
   a. compute the set of edges $C = \{X \rightarrow Y$ such that $X \rightarrow Y$ is part of a cycle$\}$.
   b. remove the edge in C that is part of the greatest number of cycles, and put it in R.
5. Reverse Edges.
   insert each edge from R in the graph, reversed.
6. Propagate directions.
   $\forall X \in V$ & $Y \in DN\{X\}$ such that neither $Y \rightarrow X$ nor $X \rightarrow Y$, execute the following rule untill no longer applies : if there exists a directed path from X to Y, orient $X \rightarrow Y$.

IV. THE PROPOSED ALGORITHM *L1GS*

Using *L1* regularization for learning *DAGs* model structure has been studied before by Li & Yang (2004, 2005), Huang et al. (2006) and Levina et al. (2008). In this paper we use *L1* regularization for variable selection to explore the Markov blanket for each variable.

we propose an improved version of *GS* algorithm where we used *L1MB* algorithm[5] , introduced by Mark Schmidt (2007), in the first step in table (1) to compute the Markov blanket for each variable $X \in V$. We show that using regression to find the Markov blanket results in much lower false negative rates and it is also more statistically efficient because it does not need to perform conditional independency tests on exponentially large conditioning sets. Table (2) shows the steps of *L1MB* algorithm.

The problem of choosing the Markov blanket for a node *X* from a set *V* can be formulated as solving the following: $\hat{\theta}_j^{L1} | (V) = \arg\min_{\theta} nill\,(X, V, \theta) + \lambda \|\theta\|_1$, where λ is the scale of the penalty on the *L1* norm of the parameter vector, excluding θ_0, the simplest way for choosing λ is to use cross validation approach. Hence *L1MB* technique depends on regressing each node X_j on all others ($V = x–j$), using *L1* variable selection. This process takes $O(nd^3)$ time per node, and (ideally) finds a set that is as small as possible, but that contains all of *X*'s parents, children and co-parents (its Markov blanket)

TABLE II
L1MB ALGORITHM

Input : *Data* X_j^i *for* $i = 1,...,n$

and $j = 1,...,p$.

Output : *Markov Blanket*

MB_j *for each node* j.

for $j = 1$ *to* p *do*

// start with empty Markov blanket.

$MB_j \leftarrow \phi;$

$s \leftarrow score\left(x_j, x_0\right);$

// optimize for base variable.

$b \leftarrow \min_b \sum_{i=1}^{n} \log\left(1 + \exp\left(x_j^i b\right)\right);$

// regression weights gradient is zero.

$g \leftarrow \sum_{i=1}^{n} x_j^i x_{-j}^i \,/\left(1 + \exp\left(x_j^i b\right)\right);$

// regularization parameter maximum value

$\lambda_{\max} \leftarrow \max_i \left\{g_i\right\};$.

for $\lambda = \left((p\text{-}1)/p\right)\lambda_{\max}$

down to 0 *in increments of* $\lambda_{\max} / p$ *do*

$\{w, b\} \leftarrow \arg\min_{w,b} \sum_{i=1}^{n} \log\left(1 + \exp\left(-x_j^i \left(w^T x_{-j}^i + b\right)\right)\right) + \lambda \|w\|1;$

// non zero variables

$nz = \{u \mid w_u \neq 0\};$

// score with selected MB.

$s_{sub} = score\left(x_j, x_{nz}\right);$

if $s_{sub} > s$ *then*

$s \leftarrow s_{sub}$; // new maximum value.

$MB_j \leftarrow nz$ // higher scoring Markov blanket

V. EXPERIMENTAL RESULTS

In order to test the effectiveness of the proposed algorithm *L1GS* and compare the results with the original *GS* and the state of the art *PC* algorithms; we performed a series of experiments on different samples of well known datasets from the Bayes net repository [14], table (3) shows some statistical information about these datasets. We used *BNT* Matlab toolbox [15] to implement *PC* and we use Mark Schmidt implementation for *L1MB* [5] for the proposed algorithm *L1GS* and the GS algorithm also implemented in Matlab. The statistical tests were done using chi-square X^2 test, for *PC* and *GS*; we chose the default value of $\alpha = 0.05$.

TABLE III
DATASETS FROM BAYES NET REPOSITORY

Name	*No. Nodes*	*No. edges*	*Max Parents*	*Fig.*
Alarm	37	46	3	2
Insurance	27	52	3	3
Hailfinder	56	66	4	4
Carpo	61	74	5	5

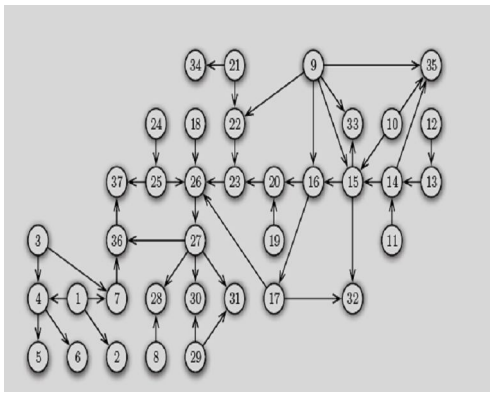

Fig. 2 Alarm Network

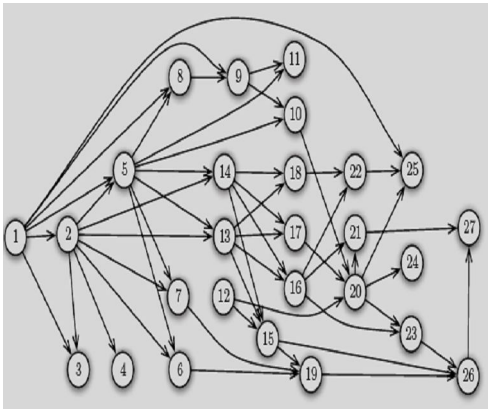

Fig. 3 Insurance Network

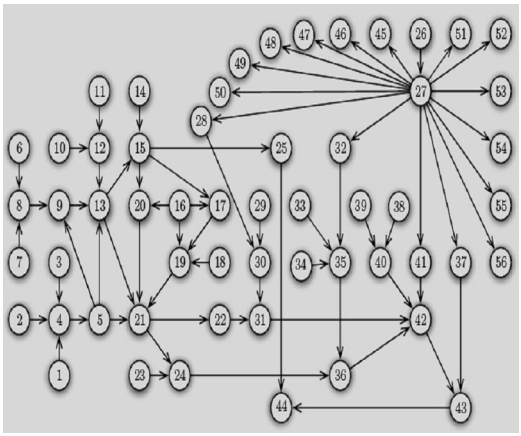

Fig. 4 Hailfinder Network

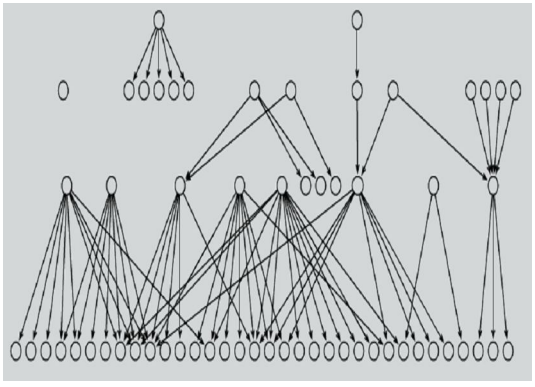

Fig. 5 Carpo Network

In this series of experiments, we compared the proposed algorithm *L1GS* to the original *GS*, where the graph being built is initialized with the Markov blanket information about each variable. However, *PC* algorithm is modified to start with the moral graph (instead of the full graph in the original version of *PC*). We tested the three algorithms on each network using chi-square X^2 test of the sample partial correlation coefficient as computed on artificial data, with significance $\alpha = 0.05$. Table (4) shows the results of these experiments.

TABLE IV
NUMBER OF TESTS AND SIZE OF THE CONDITIONING SETS AS PERFORMED BY VARIOUS ALGORITHMS

Algorithm / Network	L1GS	GS	Pc
Alarm			
Tests	1462	1485	11335
Average	2.21	2.61	4.32
Maximum	7	8	9
Insurance			
Tests	6379	6435	773567
Average	3.55	3.62	7.64
Maximum	10	11	15
Hailfinder			
Tests	2776	2809	11345
Average	2.43	2.66	5.77
Maximum	6	5	7
Carpo			
Tests	200618	209342	202525
Average	7.36	7.45	5.44
Maximum	8	8	6

The results for the modified *PC* algorithm are only shown for the comparison purposes; it is a general purpose algorithm which is not specialized in local structure learning using Markov blankets. From the comparison we note that, whenever the Markov blanket information is available or cheap to obtain, there are more efficient algorithms than the states of the art.

From the results in table (4), *L1GS* outperforms the others in nearly all the experiments or at least have the same results. *L1GS* and the original *GS* are close to one another in all scores, and outperform *PC* in the number of tests and in average and maximum size of the conditioning sets because they uses the Markov blanket information. However, our approach is a bit better than the others in terms of number of tests, while still using smaller average and maximum conditioning set sizes in all tested networks.

VI. CONCLUSION AND FUTURE WORK

The Markov blanket approaches enable constrained-based learning methods to learn the local causal structure of *DAGs*. It improves the algorithms' scalability by limiting the sizes of conditioning sets to Markov blankets. Hence, given restricted Markov blanket information with a theoretically correct conditional independence testing and firmest search procedure, the learning algorithm will guarantee the best *DAGs* structure.

Finally, Causal discovery and feature selection are strongly linked, since optimal feature selection discovers Markov blanket as sets of strongly relevant features, and causal discovery discovers Markov blankets as direct causes, direct effects and common causes of direct effects. The undirected moral graph (approximation of the causal graph) can be obtained by performing perfect feature selection on each variable. Then an extra step, computing the graph structure, is needed in order to transform the Markov blankets into parents, children and spouses. This step is exponential in the worst case, but is actually efficient provided the graph is sparse enough; this sparsity is achieved when using *L1* variable selection and this is what we implement in this paper.

The future challenges may be with learning causal structure including robust and consistent distribution-free structure learning with continuous and potentially highly nonlinear data. And we are looking forward to extend *L1GS* technique to handle multi-state discrete variables as well as modeling parent interactions and nonlinear effects.

An apparent weakness of the two-stage approaches is that if a true parent is missed in Stage 1, it will never be recovered in Stage 2. Another weakness of the existing algorithms is computational efficiency, i.e., it may take hours or days to learn a large-scale *BN* such as one with 500 nodes.

REFERENCES

[1] D. Margaritis, *Learning Bayesian Network Structure From Data.* CMU-CS, PhD thesis, 2003.
[2] D. Margaritis, S. Thrun, *Bayesian network induction via local neighborhoods*. In Advances Neural Information Processing Systems 12, 505–511, 2000.
[3] N. Friedman, M. Linial, I. Nachman, & D. Peter, *Using Bayesian Networks to Analyze Expression Data.* Computational Biology, 7, 601–620, 2000.
[4] J.P. Pellet, A. Elisseeff, *Using Markov blankets for causal structure learning*. J. Machine Learn. Res., 9, 1295–1342, 2008.
[5] Mark Schmidt, Alexandru Niculescu-Mizil & Kevin Murphy, *Learning Graphical Model Structure using L1-Regularization Paths*. Association for the Advancement of Artificial Intelligence, www.aaai.org, 2007.
[6] J. Pearl, & Verma T., *A theory of inferred causation*. Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference. San Francisco, California: Morgan Kaufmann Publishers, pp. 441-452, 1991.
[7] Peter Spirtes, Clark Glymour & Richard Scheines, *Causation, Prediction and Search.* Springer verlag, http://hss.cmu.edu/html/departments/philosophy/tetrad.book/book.html, 1993.
[8] D. M. Chickering, *Optimal structure identification with greedy search*. The Journal of Machine Learning Research, 3:507–554, 2002.
[9] Jean-Philippe Pellet & Andr´e Elisseeff, *Using Markov Blankets for Causal Structure Learning*. Journal of Machine Learning Research (9), 1295-1342, 2008.
[10] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, U.K, 2000.
[11] P. Spirtes, C. N. Glymour & R. Scheines. *Causation, Prediction, and Search*, volume 2nd. MIT Press, Cambridge, Mass, 2000.
[12] I. Tsamardinos, L. E. Brown & C. F. Aliferis , *The max-min hill-climbing Bayesian network structure learning algorithm*. Machine Learning, 65(1):31–78, 2006.
[13] F. Aliferis Constantin, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani & Xenofon D. Koutsoukos, Local *Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation*, Journal of Machine Learning Research (11), 171-234, 2010.
[14] G. Elidan, Bayes net repository. Website, URL http://compbio-.cs.huji.ac.il/Repository/. 2001.
[15] P. Leray and O. Francois, *BNT structure learning package*. URL http://banquiseasi.insa-rouen.fr/projects/bnt-slp/

