*2011 International Conference on Information and Network Technology*
*IPCSIT vol.4 (2011) © (2011) IACSIT Press, Singapore*

# Evaluation of Markov Blanket Algorithms for fMRI Data Analysis

Mr. S. A. Agrawal [1], Dr. M. S. Joshi [2], Prof. M. B. Nagori [3], Mr. T. N. Mane [4]

[1, 3, 4] Dept. of Computer Science, Govt. Engineering College, Aurangabad

[2] Dept. of Computer Science, Govt. Engineering College, Awasari

[1,2,3,4] India

**Abstract.** Aiming at the extraction and selection of features for localization of the areas of the brain that have been activated by a predefined stimulus ,this paper presents an approach to select features of fMRI datasets using Bayesian Network and Markov Blanket. When a large data set is of interest selecting relevant features is in demand in a classification task. It produces a tractable number of features that are sufficient and possibly improve the classification performance. The activity patterns in functional Magnetic Resonance Imaging (fMRI) data are unique and located in specific location in the brain. The Markov blanket contains a minimal subset of relevant features that yields optimal classification performance. This paper studies a statistical method of Markov blanket induction algorithm for filtering features. We point out an important assumption behind the Markov blanket induction algorithm and show its effect on the classification performance.

**Keywords:** fMRI, Feature selection, Dynamic causal modeling, Markov Blanket.

## 1. Introduction

### 1.1. Functional Magnetic Resonance Imaging

It is a specialized type of Magnetic Resonance Imaging scan. It helps to measure changes in the blood flow and blood oxygenation related to neural activity in the brain in response to a presented stimulus [5]. The brain requires a supply of oxygen to provide energy. According to the neural activity, supply of blood to the brain increases to provide enough oxygen.

The fMRI BOLD signal produces a 3D image composed of points called voxels, which are the analogous of pixels in a 2D image. Each voxel denotes the intensity of the signal from the brain. Statistical tests are performed in the data to determine which voxels show significant activation, then those voxels are clustered, and the resulting clusters are known as Regions of Interest (ROI).

Besides identifying which regions of the brain are active during a task, it is also helps to discover causal relationships among activity in those regions. In fMRI terminology, the causal relationship among brain regions is known as effective connectivity, and is defined as "the influenced one neural system exerts over another" [6].

FMRI data usually consist of recordings from multiple subjects, so analysis of such data requires the use of causal search methods that avoid combining the data in a straightforward way; otherwise the combined data might exhibit spurious associations. Additionally, different subjects might have different but overlapping set of ROIs and the signal strengths from the same brain regions may differ from subject to subject.

Corresponding author. Tel.: +919096234467;

*E-mail address:sanjay.07agr@gmail.com1; madhuris.joshi@gmail.com2; kshirsagarmeghana@gmail.com3; tanajisggs@gmail.com4*

### 1.2. Feature extraction

There are two main methods for reducing dimensionality: feature selection and feature extraction. Transforming the input data into a set of features is called feature extraction. In the context of neuroimaging this consists of transforming a 3 (or 4) dimensional brain scan into a long vector of features (voxels) within the brain. If the features are carefully chosen, it is then expected that the feature set will extract the relevant information from the input data in order to perform the desired classification task.

### 1.3. Feature selection

Feature selection, also known as subset selection or variable selection, is a process commonly used in machine learning, wherein a subset of the features available from the data are selected for application of a learning algorithm. In subset selection, we are interested in finding the best subset of the set of features. The best subset contains the least number of dimensions that most contribute to accuracy. We discard the remaining, unimportant dimensions. Using a suitable error function, this can be used in both regression and

classification problems. Feature selection is necessary either because it is computationally infeasible to use all available features, or because of problems of estimation when limited data samples (but a large number of features) are present [2].

There are two approaches: In forward selection, we start with no variables and add them one by one, at each step adding the one that decreases the error the most, until any further addition does not decrease the error (or decreases it only slightly). In backward selection, we start with all variables and remove them one by one, at each step removing the one that decreases the error the most (or increases it only slightly), until any further removal increases the error significantly. In either case, checking the error should be done on a validation set distinct from the training set because we want to test the generalization accuracy. Feature selection gives many benefits as follows:

- Improving prediction performance by eliminating useless noise features
- Finding measured variables relevant to target concept.
- Data efficiency because it speeding up the training and inference process.
- Time efficiency because it reducing the measurement and storage requirements
- Improving model generalization accuracy [3].

In Section 2, the Dynamic causal modeling, the Markov blanket itself is defined, as well as its relation with Bayesian Network. Then section 3 reviews all major algorithms on learning Markov blanket. It concludes with a summary about all algorithms.

## 2. Dynamic causal modeling, Bayesian Network and Markov Blanket

### 2.1. Dynamic causal modeling

DCM is a technique that originated from the neuroimaging community [8]. DCM is proposed as a confirmatory rather than as an exploratory technique for causal modeling of effective connectivity. DCM models the effective connectivity in the brain as a nonlinear and dynamic process. These characteristics make DCM models more complicated requiring a large number of free parameters and making the estimation of the parameters more dependent on constraints. The parameters of the model are not estimated. Instead, a probability distribution of the parameters is estimated by expected maximization. In a DCM model, the inputs are deterministic and are explicit from the experimental point of view; these inputs influence regions of interest, which in turn may influence other regions of interest producing a measured output that corresponds to the observed BOLD signal [8].

DCM for fMRI uses a simple (deterministic) model of neural dynamics in a network or graph of $n$ interacting brain regions or nodes. It models the change of a neuronal state-vector $x$ in time, where each region is represented by a single hidden state. It models the change of a neuronal state-vector $x$ in time, where each region is represented by a single hidden state, using the following bilinear differential equation:

$$\dot{x} = f(x, u, \theta) = Ax + \sum_{j=1}^{m} u_j B^{(j)} x + C_u \quad (1)$$

$$A = \frac{\partial f}{\partial x}|_{u=0} \quad B = \frac{\partial^2 f}{\partial x \partial u} \quad C = \frac{\partial f}{\partial u}|_{x=0} \quad (2)$$

Where $\dot{x} = dx/dt$ . This equation, results from a bilinear Taylor approximation to any dynamic model shows how changes in neuronal activity in one node $x_i$ are caused by activity in the others. More precisely, this bilinear form is the simplest low-order approximation that accounts both for endogenous and exogenous causes of system dynamics. The matrix *A* represents the fixed (or *Average*) coupling among nodes in the absence of exogenous input $u(t)$ . This can be thought of as the latent coupling in the absence of experimental perturbations. The *B* matrices are effectively the change in latent coupling induced by the $j^{th}$ input. They encode context-sensitive changes in *A* or, equivalently, the modulation of coupling by experimental manipulations. Because $B^{(j)}$ are second-order derivatives they are referred to as *Bilinear*. Finally, the matrix *C* embodies the influences of exogenous input that *Cause* perturbations of hidden states. The parameters $\theta \supset \{A, B, C\}$ are the connectivity or coupling matrices that we wish to identify. These define the functional architecture and interactions among brain regions at a neuronal level.

## 2.2. Bayesian Network and Markov Blanket

Bayesian network is a graphical tool that compactly represents a joint probability distribution P over a set of random variables U using a directed acyclic graph (DAG) G annotated with conditional probability tables of the probability distribution of a node given any instantiation of its parents. Therefore, the graph represents qualitative information about the random variables (conditional independence properties), whiles the associated probability distribution consistent with such properties that provides a quantitative description of how the variables related to each other. One example of Bayesian network is shown in Fig.1. The probability distribution P and the graph G of a Bayesian network are connected by the Markov condition property: a node is conditionally independent of its non descendants, given its parents [3].

Definition *1 (Faithfulness):* A Bayesian network G and a joint distribution P are faithful to one another if. Every conditional independence entailed by the graph G and the Markov condition is also presented in P .

Given the faithfulness assumption, the Markov blanket of T is unique, and it becomes trivial to retrieve it from the corresponding Bayesian network over the problem domain U. It is known as composed of T's parents, children and spouses (Fig.1). However, this requires the Bayesian network to be ready in advance to get the Markov blanket of some variable, but the structure learning of Bayesian network is known as NP-complete problem. Therefore, an ideal solution will allow to induce the Markov blanket but without having to have the whole Bayesian network ready first, which help to reduce the time complexity greatly .

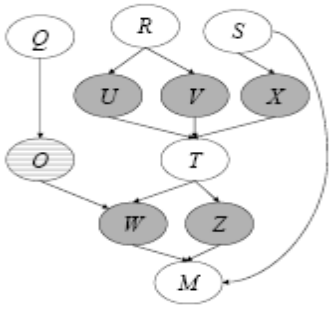


Fig.1: An example of a Bayesian network. The PC(T*)* are the variables in gray, while MB(T) additionally includes the texture-filled variable O[3].

*Definition 2 Markov Blanket (Probability viewpoint):* Given the faithfulness assumption, the Markov blanket of T is a minimal set conditioned on which all other nodes are independent of T, i.e.

$$\forall X \in U \backslash MB(T) \backslash \{T\}, I(X, T \mid MB(T)) \tag{3}$$

*Definition 3 Markov Blanket (Graphical viewpoint):* Given the faithful assumption, the Markov blanket of T is identical to T's parents, children and children' parents (spouses), i.e.

$$MB(T) = Pa(T) \cup Ch(T) \cup Sp(T) \tag{4}$$

*Theorem 1:* If a Bayesian network G is faithful to a joint probability distribution P, then: (1) there is an edge between the pair of nodes X and Y iff. X and Y are conditionally dependent given any other set of nodes; (2) for each triplet of nodes X, Y and Z in G such that X and Y are adjacent to Z but X is not adjacent to Y, X→Z←Y is a subgraph of G iff. X and Y are dependent conditioned on every other set of nodes that

contains Z . Given the faithfulness assumption, Definition 2 and Definition 3 define the Markov blanket from probability and graphical view respectively.

Definition 3 and Theorem 1 are the topology information as referred by more recent and finer algorithms such as MMPC/MB, HITON-PC/MB, PCMB and IPC-MB. Of course, faithful assumption is the basis for all, including GS, IAMB and its variants. Lucky enough, the vast majority of distributions are faithful in the sample limit [9].

# 3. Algorithms for learning Markov Blanket

## 3.1. KS

Koller-Sahami is the first algorithm for feature selection to employ the concept of Markov blanket. Although it is theoretically sound, the proposed algorithm itself doesn't guarantee correct outcome. KS algorithm requires two parameters: (1) the number of variables to retain, and (2) the maximum number of variables the algorithm is allowed to condition on. These two limits are helpful to reduce the search complexity greatly, but with a sacrifice of correctness [10,11].

## 3.2. GS

As its name indicates, GS proceeds in two steps, firstly it grows rapidly then shrinks by removing false positives. GS depends on two basic assumptions, faithfulness and correct/reliable conditional independence (CI) test. Here, the second assumption is required in practice since only when the number of observations are enough, the result of one statistical testing would be trustable. Actually, these two assumptions are also the basis of the following algorithms.

## 3.3. IAMB and Its Variants

In IAMB, it reorders the set of attributes each time when a new attribute enters the blanket in the growing phase based on updated CI testing results, which allows IAMB to perform better than GS since fewer false positives will be added during the first phase [10, 12].

Several variants of IAMB were proposed, like interIAMB, IAMBnPC and their combined version interIAMBnPC [14]. InterIAMBnPC employs two methods to reduce the possible size of the conditioning sets: (1) it interleaves the growing phase of IAMB with the pruning phase attempting to keep the size of *MB (T)* as small as possible during all steps of the algorithm's execution; (2) it substitutes the shrinking phase as implemented in IAMB with the PC algorithm instead. InterIAMB and IAMBnPC are similar to InterIAMBnPC but they only either interleave the first two phases or rely on PC for the backward phase respectively.

## 3.4. MMPC/MB

The overall MMMB algorithm is composed of two steps. Firstly, it depends on MMPC to induce which are directly connected to *T*, i.e. *PC(T)*. Then it attempts to identify the remaining nodes, i.e. spouses of *T*. The spouses of *T* are the parents of the common children of *T*, which suggests that they should belong to $\mathrm{U}_{X \in PC(T)} PC(X)$ So, MMPC is applied to each $X \in PC(T)$ to induce X's parents and children, which are viewed as spouse candidates which contains false ones to be filtered out with further checking. To determine if $Y \in \mathrm{U}_{X \in PC(T)} PC(X)$ is a spouse, actually there is a need to recognize the v- structure, i.e. $Y \rightarrow X \leftarrow T$

## 3.5. HITON-PC/MB

It works in a similar manner as MMPC/MB, with the exception that it interleaves the addition and removal of nodes, aiming at removing false positives as early as possible so that the conditioning set is as small as possible.

## 3.6. Fast-IAMB

Fast-IAMB speculatively adds one or more attributes of highest $G^2$ test significance without resorting after each modification as IAMB does, which (hopefully) adds more than one true members of the blanket. Thus, the cost of re-sorting the remaining attributes after each Markov blanket modification can be amortized over the addition of multiple attributes.

### 3.7. PCMB

PCMB claims to scale to thousands of features as IAMB does [11], but it is able to achieve much higher accuracy performance than IAMB given the same amount of data [16], which exactly reflects its data efficiency advantage.

### 3.8. IPC-MB

The overall heuristic as followed by IPC-MB is described as below:

- IPC-MB proceeds by checking and removing false positives. Considering that the size of *MB (T)* is normally much smaller than U, filtering out negatives is believed to be much easier job than directly recognizing positives.
- Recognizing and removing as many, and as early, negatives as possible is an effective way to reduce noise and to avoid conditioning on unnecessarily large conditioning set, which is the precondition for reliable CI tests and for the success of learning. Besides, it saves the computing time by preventing needless tests.
- IPC-MB filters negatives by conditioning on empty set on. Then, one variable is allowed for the conditioning set, and the checking continues on. This procedure iterates with increased conditioning set, resulting with more and more negatives are removed. So, it is obvious that the decision on a negative is made with as small conditioning set as possible, and as early as possible as well, which is the most important factor for the success of IPC-MB considering that the reliability of CI test is the most factor to influence the performance of such kind of algorithms.

Table 1: Comparison on the related algorithms for learning Markov Blanket

| | **KS** | **GS** | **IAMB** | **MMPC/ MB** | **HITON PC/MB** | **Fast IAMB** | **PCMB** | **IPC-MB** |
|---|---|---|---|---|---|---|---|---|
| **Publication Year** | 1996 | 1999 | 2003 | 2003 | 2003 | 2005 | 2006 | 2007/08 |
| **Sound in theory** | No | Yes | Yes | No | No | Yes | Yes | Yes |
| **Data Efficiency** | Poor | Poor | Very Poor | Good | Good | Poor | Good | Best |
| **Time Efficiency** | Good | Poor | Best | Poor | Poor | Best | Poor | Good |
| **Implication Difficulty** | Simple | Simple | Simple | Difficult | Difficult | Simple | Difficult | Simple |
| **Scalability** | No | Ignored | Ignored | Ignored | Ignored | Applicable | Applicable | Applicable |

## 4. Conclusion

Algorithms for causal discovery can be used over the fMRI data to try to find those interactions between brain regions. But doing causal discovery over fMRI poses several challenges and difficulties that causal discovery algorithms should address. The fMRI BOLD signal is an indirect, slower and delayed measurement of the neural activity, and that delay of the signal might vary between subjects, trials and even among brain regions of the same subject. Considering all aspects, IPC-MB achieves the best trade-off as compared with others, in term of effectiveness, time efficiency, data efficiency and topology information inferred.

## 5. References


[1]. Kenneth A. Norman, et,al, ”Beyond mind-reading: multi-voxel pattern analysis of fMRI data”, TRENDS in Cognitive Sciences, vol . 500, 2006.

[2]. Kohavi, R.and G.H. John, “Wrappers for Feature Subset Selection”, Artificial Intelligence. 97(1-2),1997.

[3]. Shunkai Fu and Michel C. Desmarais, “Markov Blanket based Feature Selection: A Review of Past Decade”, WCE ,

vol. 1,2010.

[4]. Guyon, I. and A. Elisseeff, “An Introduction to Variable and Feature Selection” Journal of Machine Learning Research, 3, 2003.

[5]. Clare, S. ,”Functional MRI: Methods and applications” University of Nottingham,1997.

[6]. Friston, K. J. Harrison, “Functional and effective connectivity in neuroimaging: A synthesis” Human Brain Mapping, 2(1-2), 56-78,1994.

[7]. Pearl, J., “Probabilistic reasoning in expert systems”, San Matego: Morgan Kaufmann, 1988.

[8]. Friston, K. J., Harrison, L., & Penny, W., “Dynamic causal modeling” NeuroImage, 19(4) ,1273-1302,2003.

[9]. Pearl, J., “Causality: Models, Reasoning, and Inference” Cambridge University Press, 2000.

[10]. Tsamardinos, I., C.F. Aliferis, and A.R. Statnikov. “Time and sample efficient discovery of Markov blankets and direct causal relations”, International Conference on Knowledge Discovery and Data Mining. ACM, 2003.

[11]. Peña, J.M., et al., “Towards scalable and data efficient learning of Markov boundaries”, International Journal of Approximate Reasoning, 45(2)., 2007.

[12]. Yaramakala, S. and D. Margaritis, “Speculative Markov Blanket Discovery for Optimal Feature Selection”, in ICDM. 2005.

[13]. Mohamed L. Seghier, Peter Zeidman**, “**Identifying abnormal connectivity in patients using dynamic causal modeling of fMRI responses**”,** Frontiers in Systems Neuroscience 2010, Vol. 4, No. 142

[14]. Tsamardinos, I., C.F. Aliferis, and A.R. Statnikov, “Algorithms for Large Scale Markov Blanket Discovery”, St. Augustine, Florida, USA: AAAI Press, 2003.

[15]. Carlos Arturo Perez, “Discovery of Causal relationship from fMRI data”, University of West Florida,2009.

[16]. Fu, S.-K. and M.C. Desmarais., “Tradeoff Analysis of Different Markov Blanket Local Learning Approaches”, in Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conference (PAKDD). 2008.