Optimizing the amount of models taken into consideration during model selection in Bayesian Networks

Robert Castelo Arno Siebes

Department of Information Systems (INS)

Research Group on Data Mining (INS1)

National Research Institute for

Mathematics and Computer Science

CWI, P.O. Box 94079

1090 GB Amsterdam, The Netherlands

{robert,arno}@cwi.nl +31.20.5924123/4139

Abstract

Graphical model selection from data embodies several difficulties. Among them, it is specially challenging the size of the sample space of models on which one should carry out model selection, even considering only a modest amount of variables. This becomes more severe when one works on those graphical models where some variables may be responses to other. This is the case of Bayesian Networks that are modeled by acyclic digraphs.

In this paper we try to reduce the amount of models taken into consideration during model selection. The less amount of models considered, the less amount of steps performed to end the model selection process, and therefore, the less computational effort required to fit data and models, We propose a simple idea: to select models from sample spaces of lower dimension and use them as starting models for sample spaces of an upper dimension.

Plots on experimental results are provided on four different synthetic datasets. They show that the main idea reduces substantially the steps taken during model selection, in comparison to a greedy model selection procedure.

1 Introduction

Bayesian Networks are a class of recursive graphical models based on acyclic digraphs (aka directed acyclic graphs – DAGs). The expressiveness of these models goes beyond decomposable graphical models, and they establish an order among variables where some are response to others. This turns into a severe problem when one tries to select a *good* set of models, since the dimension of the sample space is more than exponential in the number of variables considered [7].

It has been adopted as standard practice by many authors (e.g. Cooper and Herskovits [3] were one of the formers) to establish a complete order among the variables, that bans connections pointing to variables that appear before their source in the given order.

If we take a greedy bottom-up selection strategy, starting from an empty graph, and one considers all possible additions of arcs at every step of the selection strategy, the set of candidates has size n^2 (slightly less to avoid

directed cycles). Using the order we drop this quantity to $n(1+n)/2$. It has been proved empirically that this constraint leads the selection process to an end in a reasonable time. In our paper we raise the question whether we can do better without introducing any further constraint. More concretely, we look at the amount of models, among the considered candidates, that improve the fit at every step of a bottom-up selection strategy. This quantity is accumulated and it gives us at the end the amount of models that have been taken into consideration throughout the whole model selection process. Further, by plotting this quantity along the time line one gets a rough idea of the shape of the sample space of models given the selection criteria currently used.

We show that this total amount of models that have been taken into consideration, can be significantly reduced by selecting models from sample spaces of lower dimension, and using these models as starting models for sample spaces of an upper dimension. We provide plots of experiments on the Alarm dataset [2] which consists of 37 variables, and on three synthetic datasets we constructed ourselves, sampled from a common Bayesian Network of 48 nodes, that show reduction ratios of 7.38,8.56,10.87 and 8.93 respectively. This means, in the case of the Alarm dataset for example, that while the normal selection process needs to examine 2879 models, the proposed incremental selection strategy needs only 390, to select at the end the same model. A remaining issue of this approach, is how do we choose the successive subspaces of lower dimension, and in which order do we proceed to increase the dimension and reuse the selected models. A naive approach is to do this at random, and surprisingly it provides some gain. We give a solution that outperforms the naive approach by doing a hierarchical clustering of variables [1]. This is a procedure of polynomial cost, in concrete

$\mathcal{O}(n^2)$, that outputs a tree which contains in its leaves single variables, and aggregates at each higher level, variables into clusters.

We propose to traverse in depth-first the tree, and pick up those clusters of variables that suit a predefined size. Two successive nodes picked up in this way are likely to have a close relationship according to the tree, therefore the order in which we are picking up the clusters will determine also the order in which we increase the dimension of the sample space of models. Detailed algorithms for the complete procedure are provided.

In section 2 we will give a precise notion of what we call the set of models taken into consideration. In section 3 we will describe our main idea about optimizing the size of this set, giving a first naive approach to the implementation of this idea. In section 4 we will talk about hierarchical clustering of variables, and how this methodology is used to improve the approach described in section 3. In section 5 we will report experimentation on four different datasets, that supports our conjecture about reducing the amount of steps in model selection. Finally, in section 6 we will draw some conclusions and research questions.

2 Models taken into consideration

Let $d = (V, E)$ be an acyclic digraph, where V is the set of vertices (variables, nodes) with n elements and E the set of edges, in this case arcs, such that E does not induce any directed cycle.

Let Ω^n be the sample space of models defined by V. Let t be a traversal operator that, given a Bayesian Network w creates a set $\mathcal{W} = t(w)$ of networks that contain all possible additions of arcs over w. The cardinality $|\mathcal{W}|$ of this set of candidates is close to n^2 (directed cycles are not allowed).

Using a complete order among the members of V: $v_0 < v_1 < \cdots < v_n$, the cardinality $|\mathcal{W}|$ drops to $n(1+n)/2$. The model selection is then leaded towards certain class of models, and the process ends in a reasonable time.

Let's assume we are at the step i of the model selection procedure where $\mathcal{W}_i = t(w_i)$ is the current set of candidates given the model w_i. Let $\mathcal{W}_i^* \subseteq \mathcal{W}_i$ be the subset that contains those models that improve the fit of w_i at step i given a certain selection criteria. When $\mathcal{W}_i^* = \emptyset$ we assume that the model selection process is finished.

The entire set of models that improve the fit at every step during model selection is then

$$\bigcup_i \mathcal{W}_i^*$$

This is the set of *models taken into consideration* and we will note it as $\mathcal{W}^*$. The amount of models taken into consideration is then the cardinality of this set

$$q = |\mathcal{W}^*|$$

By plotting q along the time line at each step i of the model selection process, we obtain a rough idea of the shape of the sample space of models given the current selection criteria.

3 Optimizing $|\mathcal{W}^*|$

The core idea about how to optimize the size of $\mathcal{W}^*$ relies in splitting up the initial set of variables V, in subsets V^i that define sample subspaces of models Ω^i. These Ω^i have then lower dimension than the original one Ω^n, and therefore one can expect to face less work than carrying out model selection on the whole Ω^n.

Of course, the ultimate goal is to select models in Ω^n, but one needs not to start from scratch if we already select models from Ω^i from a lower dimension. Those models selected in Ω^i may serve as *good* starting models for Ω^j such that $j > i$, until we reach the last dimension in Ω^n.

More formally, given sample spaces of models $\Omega^{n_1}, \ldots, \Omega^{n_m}$ of dimension n_i, from which we select *good* models $w^{n_1}, \ldots, w^{n_m}$, we define a *combining* function $g(w^{n_i}, w^{n_j})$ $i \neq j$ that takes two models as input and combines them in a single model $w^{n_i+n_j}$ as output.

The model $w^{n_i+n_j} = g(w^{n_i}, w^{n_j})$ is then used as starting model for the selection process in a sample space of models $\Omega^{n_i+n_j}$.

A naive implementation of this idea is to split up V at random given some arbitrary split size i. In section 5 we will see that we already achieved some gain with this approach. By gain, we mean that $|\mathcal{W}^*|$ is reduced. The reason for this gain, even using a random approach, is that by starting model selection from a model that has more fit than the one that contains all restrictions (the empty graph), one has to discard less models to end the process.

Logically, if the set of variables V is splitted in a more sensible way than at random, one can expect to increase the gain, and this is the subject in which we are going to elaborate on the following section.

4 Splitting Ω^n by means of a hierarchical clustering of variables

A sensible way of splitting the original sample space Ω^n should be, by finding out which subspaces Ω^{n_i} $n_i < n$ contain a Bayesian Network with as many edges as possible as the underlying model.

Carrying out model selection in such subspaces is the minimum requirement in order to expect that, afterwards, the selection process will take as less steps as possible. Therefore, a *good* splitting of Ω^n is in fact a *good* clustering of the set of variables V, according to a proper criteria.

Of course, if the model underlying the original sample space of models Ω^n is close to saturation, we will not find any way of splitting Ω^n that leads to a significant reduction of the amount of models taken into consideration. On the other hand, such underlying models are not as interesting as those more sparse, provide that the important graphical information (i-mapness [6]) relies on the missing edges.

A clustering process for variables is based on a pairwise measure of association as criteria to decide the boundaries of the clusters. In hierarchical clustering of variables, the clusters are organized as a tree where the root node contains the whole set of variables and the leaves contain one single variable each. The structure of the tree from the leaves to the root node aggregates at each level variables into clusters. Let the *radius* of each cluster be the minimum degree of association among the pairs contained within the cluster. It is clear that this *radius* will decrease as long as one considers higher clusters in the hierarchy.

Typically, operational criteria (e.g. power of prediction or entropy) are recommended [1], as measure of association provide their ability of describing the strength of an association.

However, an operational interpretation that renders a relationship completely disassociated, does not imply lack of significance of that relationship (i.e. marginal independence). We are not interested in the latter situation.

In Bayesian Networks, missing edges, introduce (conditional) independencies. Thus, variables interact by means of (marginal) dependencies. This

situation suggests that we should cluster together those variables that show a significant relationship among them. The χ^2 statistic is rather accurate characterizing the significance of a relationship, when enough data is available. Therefore we are going to use a χ^2 based measure of association introduced by Cramér [4] that maps the χ^2 value into the interval $[0, 1]$, where 0 means independence and 1 means perfect association. It has been claimed that such measures still lack of an operational interpretation ([5], p. 740), but that poses no problem for us, given the situation we have described.

More concisely, let N be the sample size and i, j the cardinalities of the variables, then the measure of association given by Cramér [4] is defined as:

$$\sqrt{\frac{\chi^2}{N \min(i-1, j-1)}}$$

In order to perform the hierarchical clustering of variables one should compute the above measure of association for every pair of variables, obtaining in that way a triangular similarity matrix. The clustering process uses this matrix and a complete linkage strategy [1], that guarantees that all the variables lie within a certain maximum radius, in order to output the tree describing the hierarchy of variables.

Anderberg [1] proves that this procedure has a polynomial algorithmic cost of $\mathcal{O}(2n^2 - 9n/2)$. Given that his process is only performed once, this cost puts no effective limit on the number of variables. For further details about cluster analysis in general, and this procedure in particular, the reader may consult [1].

So far, we have discussed how the variables are clustered. Now we will show how do we extract the necessary clusters from the hierarchy, providing at the same time the ordered sequence of subspaces Ω^i in which the incremental model selection will be carried out.

Figure 1 contains the procedure in pseudocode. It requires from us to specify a cluster-size (cs), which determines size of the extracted clusters, which will be in the range $[cs-1, cs, cs+1]$.

```
algorithm incr_strategy(tree t, node top_root, int cs, list small) return list
    node r, s, w
    list inc_strat, children, sub_inc_strat
    r := root_node(t)
    children := t.children(r)
    children.gotop()
    while ¬children.end() do
        s := children.next()
        if |s| ≥ cs − 1 and |s| ≤ cs + 1 then inc_strat.add(s)
        else if |s| < cs − 1 then small.add(s)
        else
            t_s := t.sub_tree(s)
            sub_inc_strat := incr_strategy(t_s, top_root, cs, small)
            inc_strat.concatenate(sub_inc_strat)
    if r = top_root then
        small.gotop()
        while ¬small.end() do
            s := small.next()
            w.add_content(s)
            if |w| ≥ cs − 1 then
                inc_strat.add(w)
                w.empty_content()
        if ¬w.isempty() then inc_strat.add(w)
    return inc_strat
```

Figure 1: Algorithm to flatten a hierarchy of variables.

The algorithm explores recursively a tree t of a hierarchical clustering of variables in depth-first. At the moment it finds a node with a number of variables of, plus/minus one cluster size (cs), it does not go further in that branch. It stores that set of variables in an ordered list (inc_strat), and continues the exploration.

If a node contains less variables than the given cluster size minus one, then the algorithm stores these variables in a temporal ordered[1] list ($small$) This list containing the small sets of variables is used globally by the successive

[1]note, this order simply follows the depth-first exploration order

calls of the algorithm by passing it by reference. When the top call of the algorithm has explored all its branches, it creates sets of the proper size out of the *small* list, by joining its elements. Every set created in this way is then added to the *inc_strat* list, which it will contain finally the whole set of variables in clusters of a certain size. The ordered list *inc_strat* returned by the algorithm, indicates which subnetworks must be recovered first, and in which order the incremental algorithm, that discovers the whole Bayesian Network, should operate. On figure 2 we may see an example of a hierarchy describing a clustering for a set of eight variables.

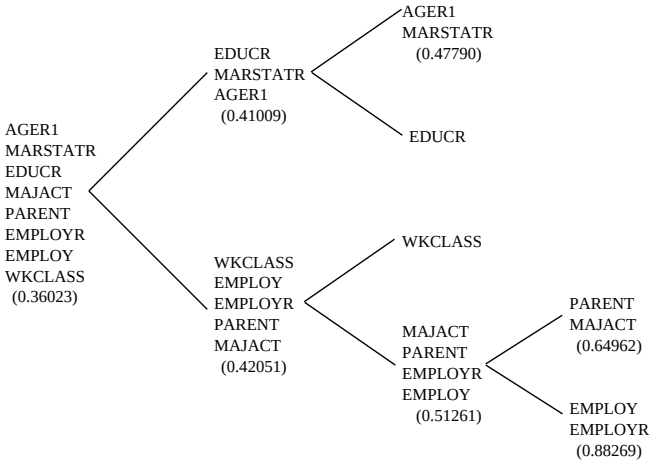


Figure 2: Hierarchical clustering on the variables of the root node of the tree. The radius of each cluster is specified between parenthesis

Let's set a cluster size (cs) of $n_i = 4\ \forall i$, then the incremental strategy obtained by applying the algorithm of figure 1 on the tree of figure 2 will pick up first a subspace Ω^3 defined over the variables on the first (at top) child node of the root. The second subspace will be Ω^5 defined over the variables on the second (at bottom) child node of the root. Finally the third subspace will be the entire space Ω^n (in this case $n = 8$, formed by the previous two subspaces.

5 Experiments

In this section we are going to show that the split of the sample space of models of n variables Ω^n, in subspaces Ω^i of $i < n$ variables reduces the set of models taken into consideration along the model selection process. We defined this set on section 2, and noted it as $\mathcal{W}^*$.

To show clearly this reduction along this section, we plot the cardinality $q = |\mathcal{W}^*|$ along the time line. Steep lines mean fast convergence of the model selection process. In all the experiments a topological order that matches the underlying model has been used, therefore we are always selecting the same model at the end, which will be in fact the model underlying the data.

The time line is normalized with respect a greedy model selection strategy, that picks up at each step the model that best fits the data. The traversal operator $t(w)$ in this strategy creates all feasible models (acyclic digraphs matching the topological order used) with one extra arc, one less arc and one arc reversed.

The combining function $g(w^{n_i}, w^{n_j})$ that we will use here is a simple one, which outputs the union of the two input set of edges as set of edges of the combined Bayesian Network to be used as starting model in the subspace $\Omega^{n_i+n_j}$.

In first place we will report plots on the naive implementation described in section 3. That is to split up the initial set of variables V at random, given some arbitrary split size i. We have used the Alarm dataset [2], which among others has been used by Cooper and Herskovits [3], that has 37 variables and ten thousand records. This dataset is a synthetic dataset sampled from the Bayesian Network of figure 3.

As we said, variables are picked up at random in the naive implementation. In order to have a reasonable picture of the behavior of this implemen-

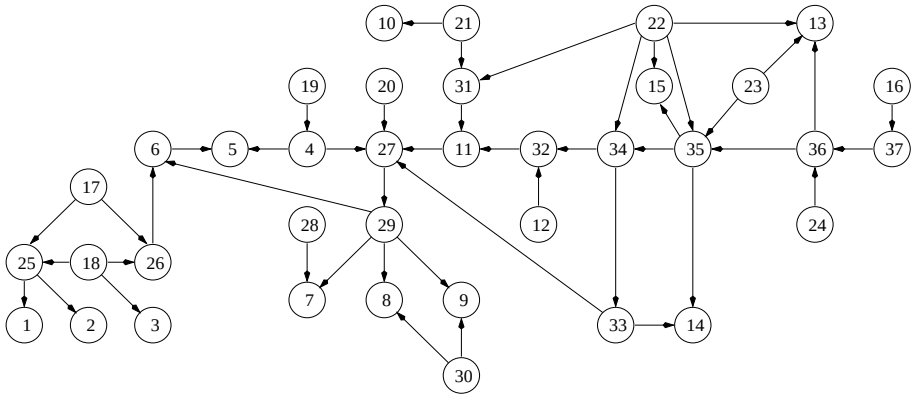


Figure 3: Bayesian Network underlying the Alarm dataset

tation, we have sampled twice the random groups of variables, providing two times the plots, one for each sample of the groups. Further we have experimented with three different split sizes, they are 4,5 and 6. In total we can see on figure 4 12 plots that compare an incremental model selection strategy using a naive implementation (grouping at random), with the greedy model selection carried out on the whole sample space of models Ω^n. The first block of six plots correspond to the first sample of random groups of variables, and the second block of six plots correspond to the second sample. Within each block the second row shows the total amount of models taken into consideration $q = |W^*|$ at every step of the model selection process, which is obtained by summing over the plot of the incremental strategy of the first row.

Along these experiments we see that q is reduced. At the bottom of each block of plots we have the ratio of q on the greedy model selection (2879) respect to q on the incremental strategy. For instance, the plots for the experiment using Ω^4 as starting subspaces and the first of the two random sample of groups of variables, show that 625 models sufficed to select the same model as in the greedy strategy which took into consideration 2879, and that is about 4 times less models.

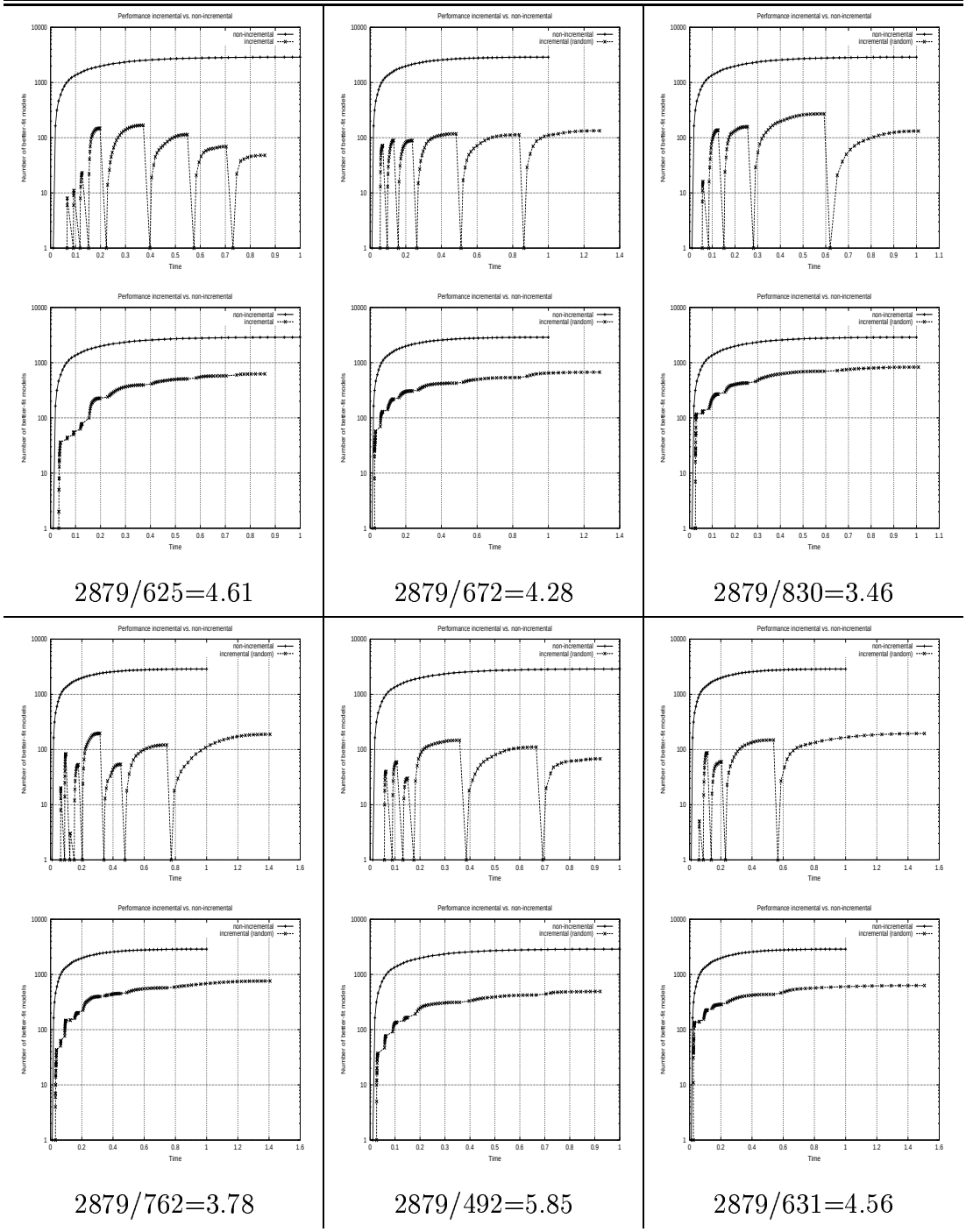
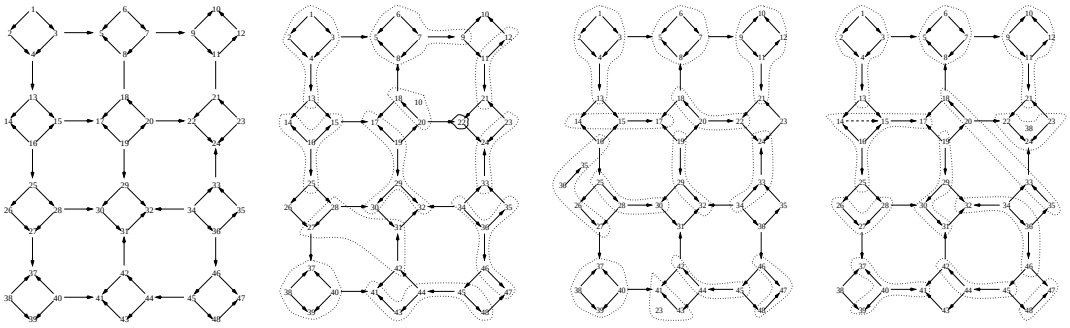


Figure 4: Greedy model selection compared with incremental model selection using a naive implementation

Next, we are going to experiment with the approach described in section 4. In order to show that a hierarchical clustering of variables is able to capture clusters containing variables which are more likely to interact among each other, than with others, we have designed a Bayesian Network with a honeycombed structure that will help to see clearly how the groups of interacting variables are observed, see figure 5.a.



(a) (b) (c) (d)

Figure 5: Honeycombed structures: (a) original network, (b) (c) (d) clusters of networks obtained from the three samples

From this honeycombed Bayesian Network we sample three datasets, each of ten thousand records. For each sample the set of probabilities of the network is newly generated. Thus the three datasets reflect different amounts of evidence in the model. By doing this we can check empirically that the approach works as we expected.

Using the algorithm described in section 4 to flatten the hierarchy of clusters, with an input cluster size of four, we obtain the clusters of variables and the order in which subspaces of models must be explored, that is the incremental model selection strategy.

Every initial Bayesian Network resembles a connected part of the original network. Figures 5.b, 5.c and 5.d outline the clusters and networks found for

each dataset.

On figure 6 we may see the first four steps of the incremental model selection strategy over the first sampled dataset (clusters specified in figure 5.b), that show how the successive subspaces Ω^i of models are joined incrementally.

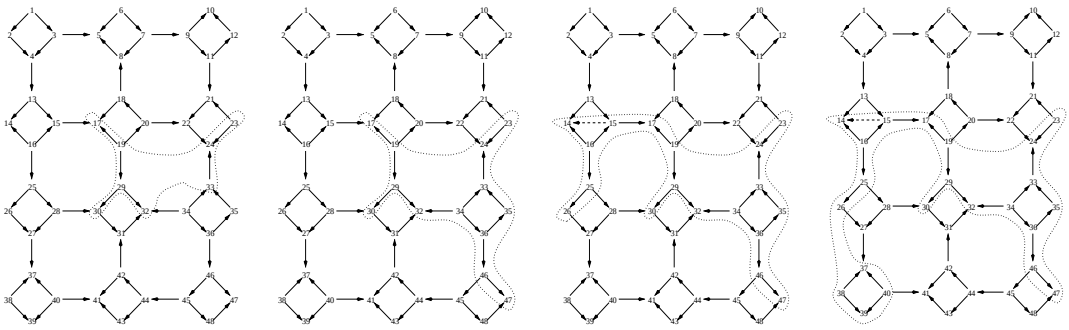
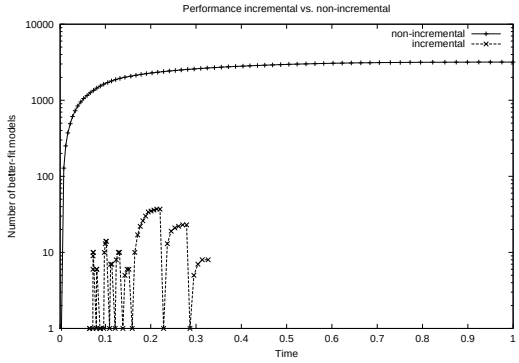


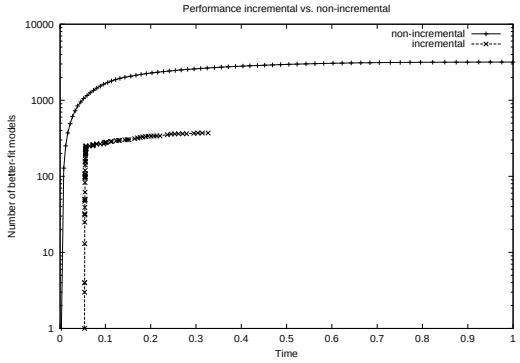
Figure 6: First four steps, from left to right, of the incremental strategy for the first sample of the honeycombed Bayesian Network. A dot line encloses the part of the network that at each step is being discovered.

Finally, it is only left to show the plots of q along the time line, and realize that a more sensible split of the initial sample space of models Ω^n improves the naive implementation of the incremental model selection strategy. In figure 7 each row corresponds to the first, second and third random sample of the honeycombed Bayesian Networks. As before, q appears in every plot for the greedy model selection strategy. The first column outlines the evolution of q in every subspace Ω^i and the second column accumulates this quantity giving the total q. We have performed this experiment only for starting subspaces Ω^4.

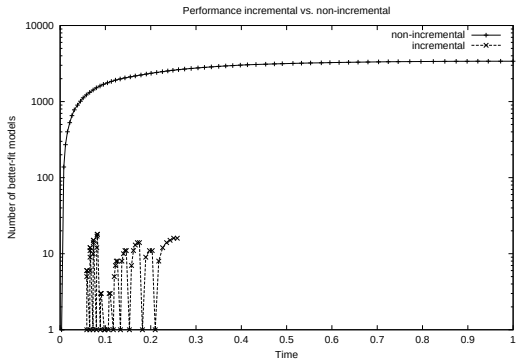
In figure 8 we may see the previous experimentation on the Alarm dataset. In this case this dataset is a unique sample, but we have carried out the experiments for three different starting subspaces $\Omega^4, \Omega^5, \Omega^6$, in order to compare it clearly with the naive approach. This comparison may be summarized



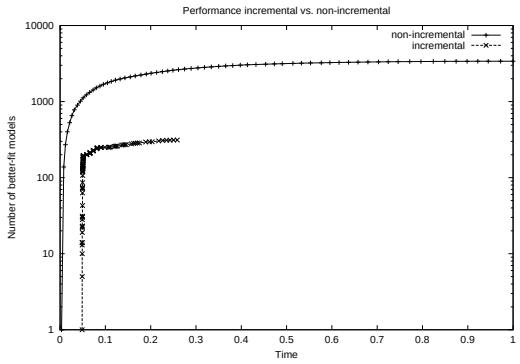
first sample



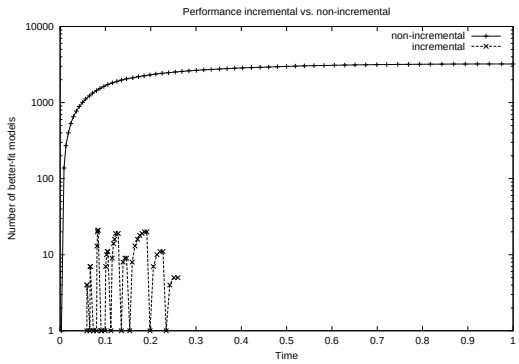
$3174/371 = 8.56$



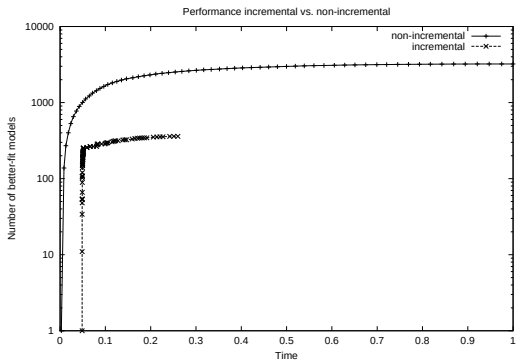
second sample



$3391/312 = 10.87$

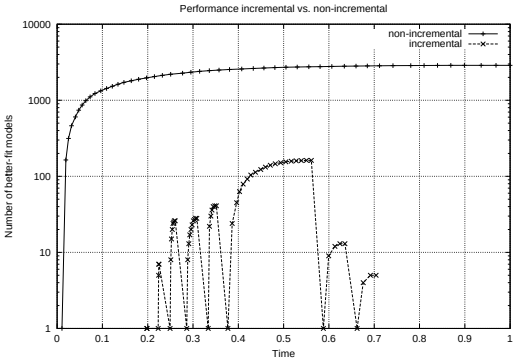


third sample

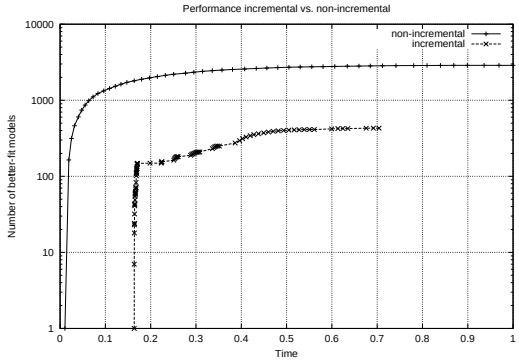


$3217/360 = 8.93$

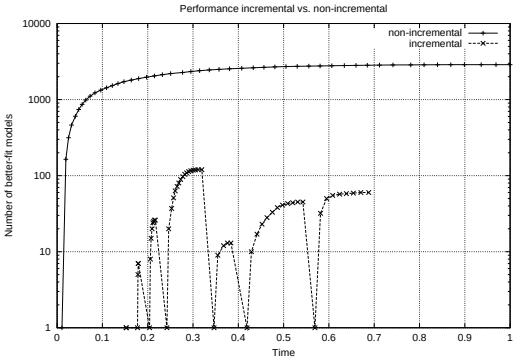
Figure 7: Greedy model selection compared with incremental model selection using an implementation based on a hierarchical clustering of variables



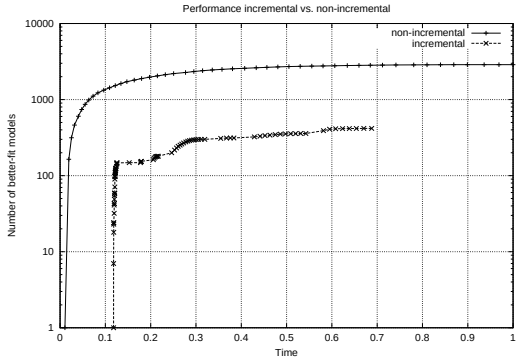
Ω^4



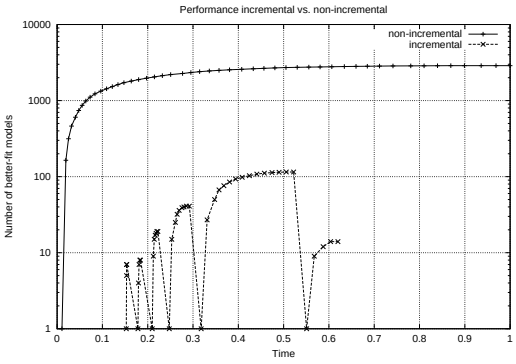
$2879/430 = 6.70$



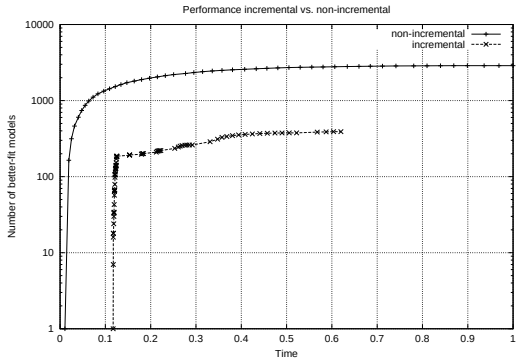
Ω^5



$2879/418 = 6.89$



Ω^6



$2879/390 = 7.38$

Figure 8: Greedy model selection compared with incremental model selection using an implementation based on a hierarchical clustering of variables

n_i	Naive	HC
4	3.78,4.61	6.70
5	4.28,5.85	6.89
6	3.46,4.56	7.38

Table 1: Comparison of reduction rations between the naive approach and using a hierarchical clustering of variables (HC).

in the table 1. In this table it is immediate that splitting by means of a hierarchical clustering improves the naive approach (randomization).

Throughout this section we have been using a topological order that matches the underlying Bayesian Network. This leads the selection strategies that we compare to the same end. We have experimented drooping this order out of the incremental strategy and the amount of models taken into consideration grow by a factor of 2. This matches the difference between the two traversal operators since using an order generates $n(1+n)/2$ models at each execution. But then of course, the incremental strategy does not select the right model anymore but with some extra arcs. We have experimented also reusing the topological order that may be inferred from models selected from sample spaces of lower dimension, and then the amount of models taken into consideration grows by a factor smaller than two, and the selected models are slightly better than those from the previous case.

6 Conclusions and Discussion

The simple idea of splitting the sample space of models Ω^n in smaller subspaces, where model selection can be carried out more efficiently, and then reuse the selected models to build a good initial one for a larger sample s-

pace, certainly appears more sensible than trying to select models directly on a vast sample space of models.

As a matter of fact, if one has a very good initial model, to start model selection with, then it is clear that model selection will not take many steps to find which model fits best the data. Of course, by *very good* here we mean that our subject-matter knowledge about the model is close to the model underlying the data.

It is of no discussion that it exists an optimal sequence of movements that would allow a model selection process to select the right model in a minimal amount of steps. It is also clear that to find this optimal sequence is an intractable problem. Nevertheless there might be procedures that help approaching this optimal sequence. Hence, one may ask whether is it possible to find a procedure of polynomial cost that outputs a sequence of variables that allows the model selection process to finish having taken into consideration an optimal (minimal) amount of models.

The combining function $g(w^{n_i}, w^{n_j})$ we have used, which just unifies the two input sets of edges, is very naive, but very little can be done given two models from two spaces $\Omega^{n_i}, \Omega^{n_j}$ defined over disjoint subsets of variables. We are currently investigating the possibility of using spaces that overlap, thus it might be possible then to find a combining function that outputs something more promising than just the union of edges of two Bayesian Networks.

It is also important to realize that the core idea of this work is rather general, so it may be applied to other kind of graphical models, having in mind then, that maybe a different measure of association might be more convenient.

The relation among different types of graphical models, and those results that allow to move from one type to another through models that lie on

overlapping regions of these classes, are a promising ground to address the two previous lines of work.

References

[1] Michael R. Anderberg. *Cluster Analysis for Applications.* Academic Press, 1973.

[2] Chavez R.M. Beinlich I.A., Suermodt H.J. and Cooper G.F. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine*, pages 247–256, London, 1989.

[3] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[4] Harald Cramér. *Mathematical Methods of Statistics.* Princeton University Press, Princeton, 1946.

[5] Leo A. Goodman and William H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:733–764, 1954.

[6] J. Pearl. *Probabilistic Reasoning in intelligent systems.* Morgan Kaufmann, 1988.

[7] R.W. Robinson. Counting labeled acyclic digraphs. In Frank Harary, editor, *New Directions in the Theory of Graphs*, pages 239–273. Academic Press, New York, 1973.