The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Learning Diverse Bayesian Networks

Cong Chen, Changhe Yuan
Graduate Center and Queens College
City University of New York
{cong.chen, changhe.yuan}@qc.cuny.edu

Abstract

Much effort has been directed at developing algorithms for learning optimal Bayesian network structures from data. When given limited or noisy data, however, the optimal Bayesian network often fails to capture the true underlying network structure. One can potentially address the problem by finding multiple most likely Bayesian networks (K-Best) in the hope that one of them recovers the true model. However, it is often the case that some of the best models come from the same peak(s) and are very similar to each other; so they tend to fail together. Moreover, many of these models are not even optimal respective to any causal ordering, thus unlikely to be useful. This paper proposes a novel method for finding a set of *diverse* top Bayesian networks, called *modes*, such that each network is guaranteed to be optimal in a local neighborhood. Such mode networks are expected to provide a much better coverage of the true model. Based on a global-local theorem showing that a mode Bayesian network must be optimal in all local scopes, we introduce an A* search algorithm to efficiently find top M Bayesian networks which are highly probable and naturally diverse. Empirical evaluations show that our top mode models have much better diversity as well as accuracy in discovering true underlying models than those found by K-Best.

Introduction

Bayesian networks (BN) (Pearl 1988) are graphical models that represent probabilistic dependencies between random variables. While BNs have become one of the most popular and well-studied probabilistic models, a common bottleneck lies in deciding upon their structure. Exactly learning the network structures from the data is known to be NP-hard, even if we restrict each variable to have at most two parents (Chickering 1996). Despite the difficulty of structure learning, a variety of algorithms have been proposed to learn optimal Bayesian networks, including dynamic programming (Koivisto and Sood 2004; Silander and Myllymäki 2006), linear programming (Jaakkola et al. 2010; Cussens 2011), and admissible heuristic search (Yuan, Malone, and Wu 2011; Yuan and Malone 2013). They all take scores of candidate parent sets of all variables as input, and use various optimization techniques to find a structure that is a good predictor of the data.

However, even the optimal Bayesian network may fail to capture the true underlying network structure. Such discrepancy is called generalization error of the learning problem, which may come from the approximation error, due to the limitations of the models, or estimation error, due to the insufficiencies of the data (Bousquet and Bottou 2008). One can potentially address the limitation of the optimal model by finding multiple best Bayesian networks with the highest scores (K-Best) in hope that one of them recovers the true model. It is often the case, however, that some of these top models come from the same peak(s) and are very similar to each other. For example, one of the top solutions may be obtained from modifying the best Bayesian network a bit by selecting a slightly worse parent set for one of the variables. Such a solution should not be considered as a top solution for two reasons: (1) it is too similar to another better solution, and (2) it is not even optimal respective to any causal ordering.

In this paper, we propose a novel method for finding a set of *diverse* top Bayesian networks, called *modes*, such that each network is guaranteed to be optimal in a local neighborhood. Such a diverse set of mode models are expected to provide a much better coverage of true underlying model. This work is inspired by recent success in developing methods for finding multiple diverse predictions, including Diverse M-Best (Batra et al. 2012), joint Diverse M-Best (Kirillov et al. 2015), and M-Modes (Chen et al. 2013; 2018). Based on a global-local theorem showing that a mode Bayesian network must be optimal in all local scopes, we introduce an A* search algorithm to efficiently find the top M Bayesian networks which are highly probable and naturally diverse. Empirical evaluations show that our top mode models have much better accuracy in discovering the true causal model than the most likely models found by K-Best.

Bayesian Network Structure Learning

A *Bayesian network* is a directed acyclic graph (DAG) that represents the uncertain relations between a set of random variables $V = \{ v_1, \ldots, v_{|V|} \}$. The relations are quantified using a set of conditional probability distributions $\Pr\big(v_i \mid \mathrm{pa}(v_i) \big)$, where $\mathrm{pa}(v_i)$ represents the immediate predecessors of v_i in the model. The joint probability distribution over all the variables factorizes as the product of the condi-

tional probability distributions. We use the terms of variable and vertex interchangeably.

Bayesian network structure learning is the problem of learning a structure from a complete discrete dataset $D = \{ d_1, \ldots, d_{|D|} \}$, where d_i is a data point consisting of values of all variables; each variable takes a value from a finite domain. Our goal is to find a network structure G that is a good predictor for the data D. One of the most popular approaches to this task is score-based learning. Given a network structure G, there is a unique score $\mathrm{score}(\, G : D \,)$ which evaluates the goodness of fit of G to the data D. For convenience, we will omit D and just use $\mathrm{score}(G)$ instead. The scoring function is typically assumed to be decomposable, that is, the *global score* $\mathrm{score}(G)$ can be decomposed into a summation of *local scores* as in: $\mathrm{score}(G) = \sum_i \mathrm{score}_i\big(\, \mathrm{pa}(v_i) \,\big)$, where the local score $\mathrm{score}_i\big(\, \mathrm{pa}(v_i) \,\big)$ is the score for each variable v_i over its parents $\mathrm{pa}(v_i)$. Several scoring functions have been proposed; most common are the BIC/MDL score (Lam and Bacchus 1994) and the BDe score (Heckerman, Geiger, and Chickering 1995). Without loss of generality, we assume score is being minimized in the remainder of this paper.

For each variable, the total number of potential parent sets is $2^{|V|-1}$, the number of subsets of all other variables. The parent sets of all variables are placed in a *table of local scores* L, with each row containing parent sets of one variable in sorted order. The size of local score tables can be significantly reduced by pruning parent sets that are provably non-optimal (Chen, Choi, and Darwiche 2016), that is, these parent sets are never used in any optimal network. Similar to existing research, we also assume that the table of local scores is given as input to our learning problem. How to calculate and prune local scores can be found in, e.g., (Campos and Ji 2011).

Given a table of local scores, Bayesian network structure learning becomes a combinatorial optimization problem of selecting one parent set for each variable such that all parent-child relations jointly form a DAG with the minimum score. In particular, if a *causal ordering* τ of variables is provided, there is an optimal Bayesian network that is consistent with this ordering. Each variable can independently choose the best parent set from preceding variables in the ordering.

The complexity of the selection is polynomial in $\mathcal{O}(\, |L| \,)$. Since each causal ordering maps to a unique optimal score, the goal of learning optimal Bayesian networks is thus to find a causal ordering with the minimum score. There are $|V|\,!$ different causal orderings. The complexity of a brute force search of the optimal network is $\mathcal{O}(\, |L| \cdot |V|\,! \,)$.

M-Modes Causal Orderings

We propose to *redefine* the problem of finding multiple best Bayesian networks as finding a set of networks that are not only optimal respective to the best causal orderings but also *diverse*. Causal ordering is an inherent property for pruning worse Bayesian networks; it is especially useful when many local scores are close to each other. Moreover, because of insufficiencies in the data, different causal orderings may lead to similar or even the same Bayesian networks. We propose to only consider top *diverse* causal orderings that are locally optimal, that is, they should be better than other causal orderings in its local neighborhood. This is because local swapping of the variables only results in marginally different models. Instead, we are interested in finding Bayesian network structures that are *qualitatively* different from each other.

In this section, we will start by reviewing an ordering-based distance measure called Kendall tau rank distance. We then define *mode* causal orderings and prove its global-local property.

Kendall Tau Rank Distance

The Kendall tau rank distance (Kendall 1938) is a metric that defines the distance between two variable orderings. The larger the distance, the more dissimilar the two orderings are. This distance is defined as counting the total number of discordant pairs for two same length orderings. For example, there are two same length orderings: $\tau_1 = \langle\, 1, 2, 3 \,\rangle$ and $\tau_2 = \langle\, 3, 1, 2 \,\rangle$. 3 pairs are in these two orderings: $\langle\, 1, 2 \,\rangle$, $\langle\, 1, 3 \,\rangle$ and $\langle\, 2, 3 \,\rangle$. The pair $\langle\, 1, 2 \,\rangle$'s orders are same in both orderings, but $\langle\, 1, 3 \,\rangle$ and $\langle\, 1, 2 \,\rangle$ are different. So, in all, the distance is 2. It is easy to see that, if two orderings are identical, the distance will be 0, and if one ordering is the reverse of the other, the distance will reach the maximum $\binom{|\tau|}{2}$, i.e. all pairs of the two orderings are reversed.

M-Modes Orderings

We first define precedence relations between causal orderings. An ordering τ_1 precedes another ordering τ_2, i.e. $\tau_1 \prec \tau_2$, if and only if either (1) the score of τ_1 is less than τ_2, or (2) they have the same score, but τ_1 is smaller than τ_2 in the lexicographical order. Lexicographical order is used only for breaking ties in a same local neighborhood. We say $\tau \prec T$ when an ordering τ has the highest precedence in the set T (τ precedes all the other orderings in set T).

We use the Kendall tau rank distance $\mathrm{kd}(\cdot, \cdot)$ as the distance metric. Given a non-negative integer δ, called the *scale*, the *δ-neighborhood* of τ is defined as $\mathcal{N}_\delta(\tau) \triangleq \{\, \tau' \mid \mathrm{kd}(\tau, \tau') \leqslant \delta \,\}$, i.e., a δ-neighborhood of an ordering τ is a set including all of the orderings which are within distance δ from τ. Once ordering precedence and δ-neighborhood are defined, the concepts of local neighborhood and local optima become clear. We define a δ-mode ordering as:

Definition 1 (δ-Mode)**.** *τ is a δ-mode* $\Leftrightarrow \tau \prec \mathcal{N}_\delta(\tau)$.

A δ-mode ordering precedes all the other orderings in its δ-neighborhood. This definition ensures that there is only one mode within each given δ-neighborhood. This also ensures the set of δ-mode orderings are diverse: any two modes are at least distance δ away. As δ increases, the δ-neighborhood of each τ grows, and the set of δ-modes monotonically shrink until only the global optimum is left: $\hat{\tau} \prec \mathcal{N}_\infty(\hat{\tau})$. The $\hat{\tau}$ is the global optimal ordering. Figure 1 illustrates what mode solutions and δ-neighborhoods are and how the scale δ affects the number of modes.

Finally, we define the problem of finding M-Modes causal orderings. Hereafter, we omit δ in some notations if the context is clear.

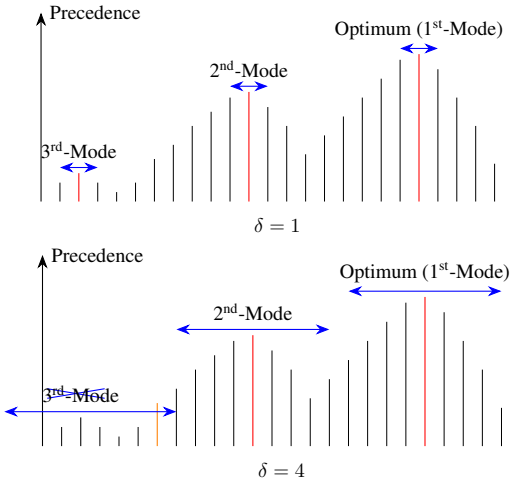

Figure 1: An illustration of modes under different δ. Each vertical bar corresponds to an ordering, and the height corresponds to its precedence. (Top) When $\delta = 1$, there are three modes (red). (Bottom) When $\delta = 4$, only two modes are left. The third mode is no longer locally optimal in its δ-neighborhood because of the orange solution.

Problem 1 (M-Modes Causal Orderings)**.** *Given a scale δ, compute top M mode causal orderings.*

Global-Local Theorem

We cannot rely on Definition 1 in verifying whether a causal ordering is a mode or not, as there are two many causal orderings in its δ-neighborhood. Inspired by Theorem 2 in (Chen et al. 2013) and Theorem 1 in (Chen, Yuan, and Chen 2016), which show a close connection between the mode labeling of a graph and the local MAPs of its subgraphs, we present some theoretical properties of a mode causal ordering and its local patterns.

Consider an example ordering $\tau = \langle 1, 2, 3, 4, 5, 6 \rangle$. We can create three different orderings with distance 3 via: (a) reverse adjacent variable pairs — $\langle 1, 2 \rangle$, $\langle 3, 4 \rangle$, and $\langle 5, 6 \rangle$ — resulting in $\langle \underline{2}, \underline{1}, \underline{4}, \underline{3}, \underline{6}, \underline{5} \rangle$; (b) reversing $\langle 1, 2, 3 \rangle$, resulting in $\langle \underline{3}, \underline{2}, \underline{1}, 4, 5, 6 \rangle$; or (c) moving 1 to between 4 and 5, resulting in $\langle \underline{2}, \underline{3}, \underline{4}, \underline{1}, 5, 6 \rangle$. The numbers in underline represent the *scope* of variables involved in the change. However, case (a) above can be considered three separate changes in smaller scopes independent from each other, namely, reversing $\langle 1, 2 \rangle$, reversing $\langle 3, 4 \rangle$, and reversing $\langle 5, 6 \rangle$. We call such a scope of change *divisible*. Both case (b) and case (c) are *indivisible* scopes of change. Since the Kendall tau rank distance is defined as counting reverse pairs, for a given distance δ, the *widest* indivisible scope of change has a size of at most $(\delta + 1)$, achieved by moving only one variable δ distance away. Case (c) is an example of that.

The concept of *indivisibility* allows us to exploit the decomposability of scoring functions. If we can divide a scope of variables into smaller pieces, we only need to verify whether each indivisible scope is optimal. Given an ordering τ and a scale δ, we define *local ordering*, $\tau_{[:]}$ as a connected part of τ, where [:] indicates slicing a part from τ. The number of variables in a local ordering is its *size*. For example, assuming $\tau = \langle 4, 3, 2, 1, 5, 6 \rangle$, $\tau_{[2:4]}$ represents the local ordering $\langle 3, 2, 1 \rangle$ with size 3. The example also makes it clear that the orderings with the same distance from an ordering τ can have different sizes of scope. Therefore, given a distance δ, we further define δ-*scope* of a local ordering $\tau_{[:]}$ as the set of permutations of the local ordering which are no more than distance δ from $\tau_{[:]}$. More formally, $\text{scope}_\delta(\tau_{[:]}) \triangleq \{ \tau'_{[:]} \mid \text{kd}(\tau_{[:]}, \tau'_{[:]}) \leqslant \delta \}$. The highest precedential local ordering in a δ-scope is called a *local optimum*, $\hat{\tau}_{[:]}$. With our previous definition of the precedence rule, a δ-scope of an ordering must have one and only one local optimum.

Finally, we have the following theorem about a mode ordering.

Theorem 1 (Global-Local)**.** *An ordering τ is a δ-Mode if and only if each of its local ordering $\tau_{[:]}$ with size $\delta + 1$ is optimum in its δ-scope.*

Proof. $\Rightarrow$ (Sufficiency): Suppose τ is a mode and $\exists \tau_{[:]}$ is not a local optimum $\hat{\tau}_{[:]}$. Both $\tau_{[:]}$ and $\hat{\tau}_{[:]}$ have size $\delta + 1$.

Consider τ' is the same as τ except that ${\tau'}_{[:]} = \hat{\tau}_{[:]}$, which means τ' goes along $\hat{\tau}_{[:]}$, so that $\hat{\tau}_{[:]} \subseteq \tau'$

$\because \text{scope}_\delta(\tau'_{[:]}) = \text{scope}_\delta(\tau_{[:]}) \therefore \tau' \in \mathcal{N}_\delta(\tau)$.

$\because \hat{\tau}_{[:]} \prec \tau_{[:]}$, $\hat{\tau}_{[:]} \subseteq \tau'$ and $\tau' \in \mathcal{N}_\delta(\tau) \therefore \tau' \prec \tau$.

This contradicts the fact that τ is a mode.

$\Leftarrow$ (Necessity): Suppose τ is not a mode but $\forall \tau_{[:]} \subseteq \tau$, $\tau_{[:]} = \hat{\tau}_{[:]}$, which means all of its local orderings are optima, then $\exists \bar{\tau} \in \mathcal{N}_\delta(\tau)$, such that $\bar{\tau}$ is a mode.

Consider τ_d is the maximal difference between τ and $\bar{\tau}$ of a scope (means connected). So that, $\text{scope}_\delta(\tau_d) = \text{scope}_\delta(\bar{\tau}_d)$. Let $\bar{\tau}'$ is the same as $\bar{\tau}$ except for τ_d, such that $\hat{\tau}_d \subseteq \bar{\tau}'$.

$\because \hat{\tau}_d \prec \tau_d$, $\hat{\tau}_d \subseteq \bar{\tau}'$ and $\bar{\tau}' \in \mathcal{N}_\delta(\bar{\tau}) \therefore \bar{\tau}' \prec \bar{\tau}$. So, $\bar{\tau}$ is not a mode.

This contradicts the fact that $\bar{\tau}$ is a mode. So, τ must be a mode.

□

M-Modes Bayesian Networks

We define *mode Bayesian networks* to be *unique* networks generated from the best mode causal orderings. Different causal orderings may produce the same network structure because there may not be sufficient data to distinguish between these orderings. If these orderings are δ distance away from each other, they are all considered valid mode causal orderings. In practice, we can alleviate, but not solve completely, the problem by increasing δ to promote diversity between the orderings, as shown in our empirical results.

We now introduce an A* search algorithm [1] for finding M-Modes Bayesian networks based on Theorem 1. We for-

[1] Although depth-first search can in principle be applied, it is not easy to find a good initial upper bound for our task of finding multiple top mode solutions

Algorithm 1 Finding M-Modes Bayesian Networks

Input: a local scores table L; δ; M
  function A*-MODE(L, δ, M)
    Initialize $Open$ list with an empty ordering
    while $Open$ not empty **do**
      Pops out the best state cur from $Open$
      if cur fails local-optimum test **then** Continue
      else
        if cur is a complete ordering **then**
          Extract a mode Bayesian network bn from cur
          if bn is unique, and M mode Bayesian networks are found **then** Exit
        else Generate successors of cur and add to $Open$

mulate our search space as a tree of all causal orderings. Each node of the search space is called a state which stores an incomplete ordering with a quality score. The root node is the empty ordering. Each node at layer l has l variables in its partial ordering, and can branch to maximally $|V| - l$ successor nodes. The A* algorithm uses an open list, usually a priority queue, to keep track of frontier states that have not been expanded. At each step, the A* algorithm pops out the *best* state from the open list. We subject the current state to a *local-optimum test*. If the test passes, and if the state is a complete ordering, a mode causal ordering as well as the corresponding mode Bayesian network are found. If the state is not a complete ordering, its successors are generated and added to the open list, called expanding a state. Otherwise if the test fails, the current state will be discarded. Note that because the search space is a tree with no loops, it is unnecessary to have a closed list for duplicate detection. The first few solutions found by A* are guaranteed to be the best mode causal orderings, from which a set of mode Bayesian networks can be extracted. The search continues until we find M unique mode Bayesian networks.

Two aspects of the A* algorithm need further explanation. One is the the local-optimum test. The test checks whether all (δ+1)-local orderings of the current state are optimal in their respective δ-scopes. Since the test is done at each search step, only one new (δ+1)-local ordering needs to be checked at any step. Let the best state popped up from the open list represents a partial ordering τ and is at layer l. We only need to check whether it has a local optimum on the *tail*. We slice the (δ+1)-local ordering $\tau_{[l-\delta:l]}$ from τ and enumerate its δ-scope containing all permutations of the same variables which are no more than distance δ from $\tau_{[l-\delta:l]}$. If any of the permutation precedes $\tau_{[l-\delta:l]}$, τ fails the local-optimum test and is discarded. Otherwise, τ passes the test.

For example, if $\delta = 2$, the open list pops out a state representing $\langle 1, 2, 3, 4 \rangle$. We check whether the local ordering $\langle 2, 3, 4 \rangle$ is a local maximum. We enumerate all the members in its δ-scope, which contains all permutations except $\langle 4, 3, 2 \rangle$ whose distance is 3. In total we get five permutations to compare with. If $\langle 2, 3, 4 \rangle$ precedes all of the five orderings, it passes the test and is eligible to generate successor states. Otherwise, state $\langle 1, 2, 3, 4 \rangle$ will be discarded.

The other is the *quality score* of a state s, $f(s)$. The $f(s)$ is calculated as the sum of a current score, $g(s)$, and a future score, $h(s)$. The $g(s)$ is the total score from the start state to s. The $h(s)$ lower bounds the score from s to a goal and is estimated from a heuristic function. For Bayesian network learning, $g(s)$ corresponds to the score of the subnetwork over the partial ordering, and $h(s)$ estimates a lower bound score of the remaining variables. We used the static pattern database-based heuristic function with two random equal-sized groups presented in (Yuan and Maone 2012). The basic idea of the heuristic is to enforce the acyclicity within the predefined groups of variables but relax acyclicity between groups.

The full search tree has up to $|V|!$ leaves and $1 + \sum_{i=1}^{|V|} \prod_{j=0}^{i-1}(|V| - j)$ nodes. However, because of the local-optimum test, only mode causal orderings will ever lead to a leave. Also, only one local ordering in the δ-scope of each (δ+1)-local ordering passes the test; all others will end a search branch immediately. Finally, the best-first search strategy of A* only explores the most promising search space in finding the top solutions. The practical search space explored is thus much smaller. Finally, a pseudo code of the A* algorithm is presented in Algorithm 1.

Experiments

We implemented and tested our proposed method (named M-Mode-BNs for short) on top of URLearning [2]. As commonly done in the multiple prediction literature (Batra et al. 2012), we use K-Best as our baseline. In particular, we used the *K-Best Software* [3] described in (Tian, He, and Ram 2012). K-Best takes input a complete table of local scores (without pruning) and finds top k best-scoring networks via a dynamic programming algorithm. K-Best uses the BDe score with a uniform structure prior and an equivalent sample size of 1. The same local score tables are used in our M-Mode-BNs method for consistency. Another implicit baseline is M-Mode-BNs method with $\delta = 0$, in which it returns the M best Bayesian networks that are optimal respective to any causal ordering. In our experiments, we focus on evaluating *exact* methods for solving K-Best or M-Mode-BNs. Comparing to approximate methods for learning diverse Bayesian networks is interesting but left as future

[2] www.urlearning.org

[3] web.cs.iastate.edu/jtian/Software/UAI-10/KBest.htm

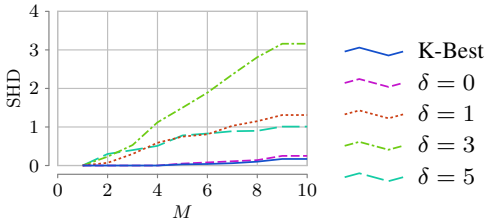

Figure 2: The average pairwise SHDs for K-Best and M-Mode-BNs on 10 data sets (100 data points each) sampled from Sachs.

work.

We selected several discrete benchmark models from *bnlearn Bayesian Network Repository* [4], including Survey (Scutari and Denis 2014), Asia (Lauritzen and Spiegelhalter 1988), Sachs (Sachs et al. 2005) and Child (Spiegelhalter et al. 1993). Our goal is to test the collective capability of a set of top Bayesian networks in discovering the true underlying causal model, while the goal in (Tian, He, and Ram 2012) was to use top models for Bayesian model averaging. We can also extend our method to perform approximate model averaging, although the extension is left as future work as well.

Our experiments were performed on an IBM System with 32 core 2.67GHz Intel Xeon Processors and 512G RAM. The program was written in C++ using the GNU compiler G++ on a Linux system. We also used functions from an R package called *bnlearn* in some data postprocessing tasks.

Diversity of Top Models

In the first experiment, we compared the diversity of the top Bayesian networks found by M-Mode-BNs and K-Best. We define the *diversity* of a set of Bayesian networks as the average pairwise structural Hamming distance (SHD) between the networks. A larger average SHD means higher diversity. We use the dataset Sachs in this experiment. We randomly sampled 10 data sets with 100 data points each from the network. Then we use our method with different δ values to learn different numbers of top mode Bayesian networks from each data set. We compared the average diversity of our models against the same numbers of models found by K-Best. Figure 2 shows the results. We can see that mode Bayesian networks found by M-Mode-BNs with $\delta > 0$ are much more diverse than those of K-Best and M-Mode-BNs with $\delta = 0$. In general, larger δ results in higher diversity of learned models, although it is not monotonic. The reason is that δ measures the distance between causal orderings, but SHD measures distance between equivalence classes.

Accuracy in Structure Learning

We also compared the ability to recover true models of M-Mode-BNs and K-Best. We use the *minimum* structural Hamming distance (SHD) between a set of candidate Bayesian networks and the ground truth network to measure the collective capability of the candidate set in discovering the true underlying structure. This is called *oracle accuracy*, i.e., the best one among the top results, which is commonly used in the literature on multiple diverse predictions (Batra et al. 2012; Kirillov et al. 2015; Chen et al. 2013).

For each benchmark network, we generated three kinds of data sets, with 10, 100 and 1000 data points respectively; 10 random data sets were sampled for each size and were enough to show the trends. We allow both K-Best and M-Mode-BNs to find varying numbers of top solutions; M-Mode-BNs with different δ values were tested. The average SHD distances over 10 random data sets for all settings are shown in Figure 3. K-Best cannot scale to Child network with 20 variables, because computing complete local score tables is extremely expensive for larger data sets. Therefore, we did two experiments on the network. In the first, we dropped 5 leaf nodes so that we can compare it with K-Best. In the second, we only ran M-Mode-BNs on the complete Child network with all 20 variables. M-Mode-BNs is more scalable than K-Best because it can alternatively take pruned tables of local scores as input.

The results show that M-Mode-BNs with $\delta > 0$ in general has much better oracle accuracies in discovering the true network structures than K-Best. When the data size is small, many network structures receive non-negligible probabilities. Many of the top Bayesian networks are structurally quite different from the true underlying structure, because there is simply not enough data to distinguish the models. The average SHD is generally large. When the data size increases, there is a dramatic decrease in the average SHD. This means the top Bayesian networks become more similar to the true network. The probability mass also concentrates more and more on the likely models. K-Best worked better when there are more data. Still, M-Mode-BNs can help to achieve better oracle accuracies in discovering true networks. It is not always predictable which δ works best for M-Mode-BNs. Our general observation is that the larger the δ, the sparser the top mode Bayesian networks. If there are many mode Bayesian networks, δ can be set higher to achiever better diversity. But if δ is set too high, only very few mode Bayesian networks are left; the accuracy results will suffer as a result. M-Mode-BNs with $\delta = 0$ is all over the map. Sometimes it has the better accuracy results, such as on Survey (10, 1000) and Asia (10, 1000), but some other times it is as bad as K-Best, such as on Sachs (100) and Child* (100). It means that simply finding Bayesian networks corresponding to best causal orderings cannot reliably address the diversity issue of top solutions. We should explicitly promote diversity by using a positive δ. Note that δ is a hyper-parameter whose optimal value depends on specific problems and should be tuned. An optimal δ should reach a balance between both diversity and scores of modes; both of which are necessary for high quality solutions. Figure 3 indicates that the optimal δs are at least correlated with the data set sizes and the network sizes.

There are a few exceptions in which K-Best outperformed some of the M-Mode-BNs methods, including Survey (10),

[4] www.bnlearn.com

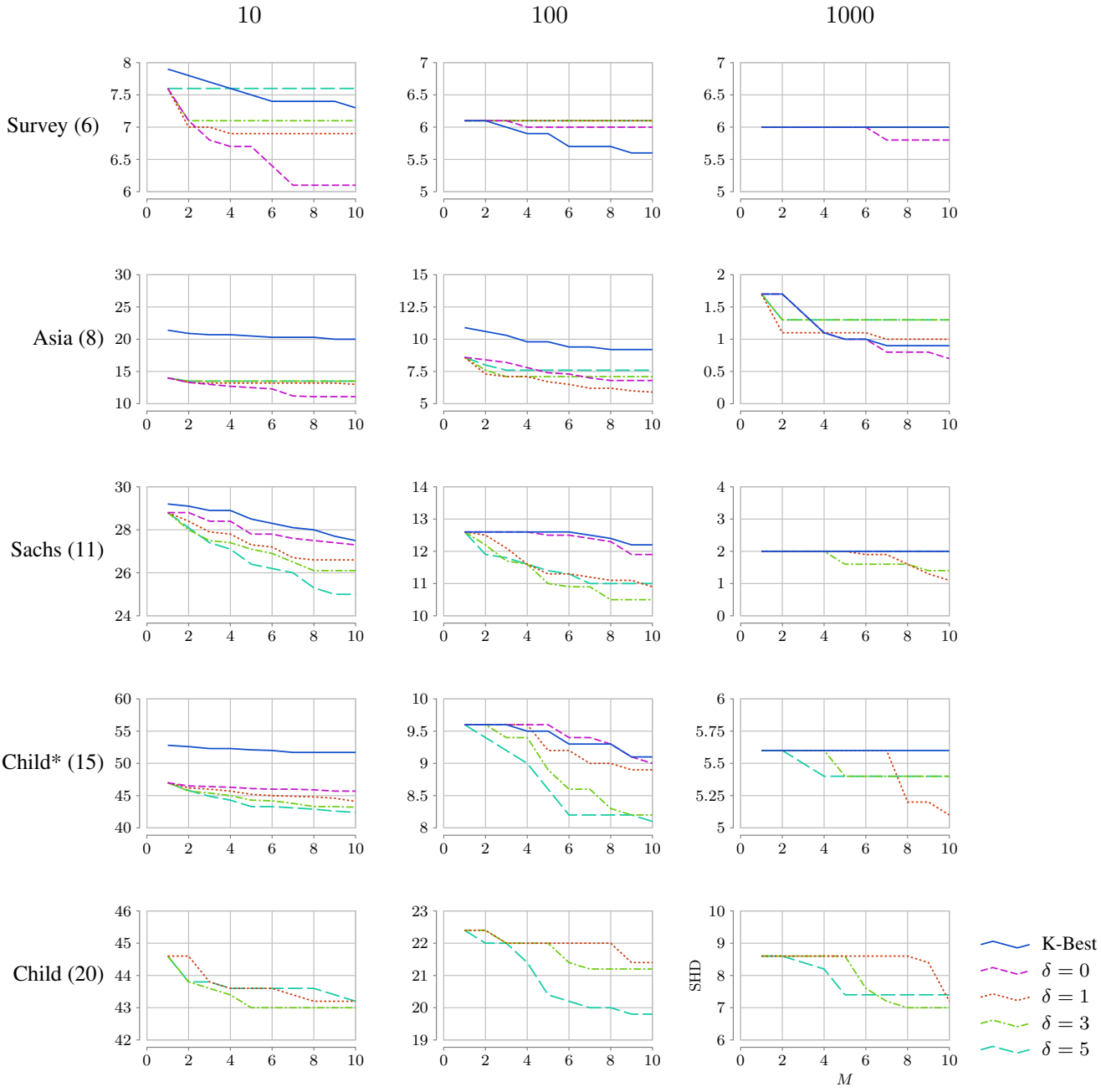

Figure 3: The average SHD accuracy results for K-Best and M-Mode-BNs on data sets with different sizes (columns) for different benchmarks (rows). The number after each network shows the network size.

	10	100	1000
K-Best	6.0/10.0	1.5/10.0	1.0/10.0
$\delta = 0$	6.7/10.0	1.4/10.0	1.0/10.0
$\delta = 1$	8.2/10.0	3.3/10.0	1.5/10.0
$\delta = 3$	9.4/10.0	6.1/9.9	1.8/9.5
$\delta = 5$	8.8/9.3	2.5/4.7	2.2/5.3

Table 1: The average numbers of unique equivalence classes/ top models found by K-Best and M-Mode-BNs for data sets with different sizes on Sachs.

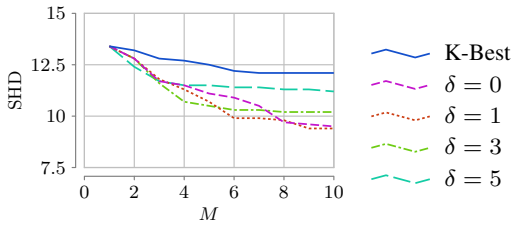

Figure 4: The average SHD accuracy for unique equivalence classes found by K-Best and M-Mode-BNs on data sets (100 data points each) sampled from Sachs.

Survey (100) and Asia (1000). On Survey (10), M-Mode-BNs with $\delta = 5$ only found one mode Bayesian network; all other likely Bayesian networks are suppressed because of the large δ. Similarly on Survey (100) and Asia (1000), the number of data points is large relative to the network size. Again M-Mode-BNs only found very few mode Bayesian networks. In comparison, K-Best often has many more top models to use.

One observation of the results may seem puzzling. Intuitively, the very best solution should be the same for both K-Best and M-Mode-BNs. Therefore, the accuracy curves of all methods should have the same starting point (when $M = 1$). However, in some of the graphs, K-Best has worse starting accuracies. Upon further investigations into the results, we found that, although K-Best and M-Mode-BNs always found top models with the same best score, the models may come from different equivalence classes, especially when the amount of data is limited. Therefore, the SHD of the top models found by K-Best and M-Mode-BNs may be different. For some unknown reason, M-Mode-BNs found top models with smaller SHD than K-Best in most cases, but not always.

As mentioned earlier, M-Mode-BNs is much more scalable than K-Best because it can alternatively take pruned local score tables as input. However, we can see that there are no results for M-Mode-BNs with $\delta = 0$ on Child* (15, 1000) or on Child (20). It is because, when $\delta = 0$, we are essentially performing exhaustive search in the search tree. When δ is too large, M-Mode-BNs can become less efficient because the local-mode test at each search step is very expensive. Otherwise, M-Mode-BNs with medium or small δ values tend to have similar efficiency as K-Best. As a typical example, K-Best took 9.5s on average on 10 random datasets of Child*(15, 100) to find the top 10 networks, while M-Mode-BNs took 10.1s when $\delta = 1$, 5.2s when $\delta = 3$, and 177s when $\delta = 5$.

Effect of Equivalence Classes

In Figure 2, the initial diversity stays at 0.0 until $M = 4$. Also in Figure 3, the SHD curves of K-Best often stay flat. This is because many top Bayesian networks found by K-Best are from the same equivalence class and lack diversity. As an example, we computed the number of different equivalent classes out of the top solutions for each setting on Sachs and present the results in Table 1. Again, we observe that a larger δ often leads to fewer mode Bayesian networks, although they tend to be from different equivalence classes.

Addressing such diversity issue was the main motivation for us to develop the M-Mode-BNs method. Nevertheless, it is a fair question to ask whether finding *top equivalence classes* (Chen and Tian 2014) instead of top Bayesian networks can similarly address the diversity issue. In order to obtain some initial insights, we allowed both K-Best and M-Mode-BNs to find as many top models as possible, during which we filtered out models that belong to the same equivalence classes of existing models. In other words, we only keep top models that belong to unique equivalence classes. We then compare the oracle accuracies of those filtered models. The results are shown in Figure 4. The results show that the equivalence classes found by M-Mode-BNs still have better oracle accuracies than those of K-Best. The results indicate that diversity in equivalence classes is also desirable.

Concluding Remarks

In this paper, we introduce a novel method called M-Mode-BNs for finding a diverse set of top mode Bayesian networks. Our results show that the top mode Bayesian networks found by M-Mode-BNs have much better oracle accuracies in discovering the true underlying network structures in comparison to K-Best, which simply finds the top models with the best scores. Preliminary results also show that such diversity cannot be achieved by learning top equivalence classes. Also, we only used oracle accuracy to evaluate the quality of mode solutions. In practice, we can ask an expert to choose a final solution (Flerova, Marinescu, and Dechter 2016), rank and combine a very large pool (Li, Carreira, and Sminchisescu 2010), or even further improve the solutions in a human-in-the-loop environment.

As future work, we plan to generalize our method to find top diverse equivalence classes. Open questions include how to define mode equivalence classes and how to efficiently search for them. We also want to extend our method to perform approximate model averaging and compare to approximate methods for finding diverse Bayesian networks, such as sampling and local search.

Acknowledgment

This research was supported by the NSF grant IIS-1829560 and the PSC-CUNY award 61542-00 49.

References

Batra, D.; Yadollahpour, P.; Guzman-Rivera, A.; and Shakhnarovich, G. 2012. Diverse M-best solutions in markov random fields. *Computer Vision–ECCV 2012* 1–16.

Bousquet, O., and Bottou, L. 2008. The tradeoffs of large scale learning. In Platt, J. C.; Koller, D.; Singer, Y.; and Roweis, S. T., eds., *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc. 161–168.

Campos, C. P. d., and Ji, Q. 2011. Efficient structure learning of bayesian networks using constraints. *Journal of Machine Learning Research* 12(Mar):663–689.

Chen, Y., and Tian, J. 2014. Finding the k-best equivalence classes of bayesian network structures for model averaging. In *AAAI*, 2431–2438.

Chen, C.; Kolmogorov, V.; Zhu, Y.; Metaxas, D.; and Lampert, C. H. 2013. Computing the M most probable modes of a graphical model. In *International Conf. on Artificial Intelligence and Statistics (AISTATS)*.

Chen, C.; Yuan, C.; Ye, Z.; and Chen, C. 2018. Solving m-modes in loopy graphs using tree decompositions. In *International Conference on Probabilistic Graphical Models*, 145–156.

Chen, E. Y.-J.; Choi, A.; and Darwiche, A. 2016. On pruning with the mdl score. In *Conference on Probabilistic Graphical Models*, 98–109.

Chen, C.; Yuan, C.; and Chen, C. 2016. Solving m-modes using heuristic search. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16)*.

Chickering, D. M. 1996. Learning bayesian networks is np-complete. In *Learning from data*. Springer. 121–130.

Cussens, J. 2011. Bayesian network learning with cutting planes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 153–160. AUAI Press.

Flerova, N.; Marinescu, R.; and Dechter, R. 2016. Searching for the m best solutions in graphical models. *Journal of Artificial Intelligence Research* 55:889–952.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20(3):197–243.

Jaakkola, T.; Sontag, D.; Globerson, A.; and Meila, M. 2010. Learning bayesian network structure using lp relaxations. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 358–365.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30(1/2):81–93.

Kirillov, A.; Savchynskyy, B.; Schlesinger, D.; Vetrov, D.; and Rother, C. 2015. Inferring M-Best diverse labelings in a single one. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Koivisto, M., and Sood, K. 2004. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research* 5(May):549–573.

Lam, W., and Bacchus, F. 1994. Learning bayesian belief networks: An approach based on the mdl principle. *Computational intelligence* 10(3):269–293.

Lauritzen, S. L., and Spiegelhalter, D. J. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 157–224.

Li, F.; Carreira, J.; and Sminchisescu, C. 2010. Object recognition as ranking holistic figure-ground hypotheses. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 1712–1719. IEEE.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721):523–529.

Scutari, M., and Denis, J.-B. 2014. *Bayesian networks: with examples in R*. CRC press.

Silander, T., and Myllymäki, P. 2006. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 445–452. AUAI Press.

Spiegelhalter, D. J.; Dawid, A. P.; Lauritzen, S. L.; and Cowell, R. G. 1993. Bayesian analysis in expert systems. *Statistical science* 219–247.

Tian, J.; He, R.; and Ram, L. 2012. Bayesian model averaging using the k-best bayesian network structures. *arXiv preprint arXiv:1203.3520*.

Yuan, C., and Malone, B. 2013. Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research (JAIR)* 48:23–65.

Yuan, C., and Maone, B. 2012. An improved admissible heuristic for learning optimal bayesian networks. In *Proceedings of The 28th Conference on Uncertainty in Artificial Intelligence (UAI-12)*.

Yuan, C.; Malone, B.; and Wu, X. 2011. Learning optimal bayesian networks using A* search. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, 2186.