Structure Inference of Bayesian Networks from Data: A New Approach Based on Generalized Conditional Entropy

Dan A. Simovici*, Saaid Baraty*

*Univ. of Massachusetts Boston, Massachusetts 02125, USA
{dsim,sbaraty}@cs.umb.edu

Abstract. We propose a novel algorithm for extracting the structure of a Bayesian network from a dataset. Our approach is based on generalized conditional entropies, a parametric family of entropies that extends the usual Shannon conditional entropy. Our results indicate that with an appropriate choice of a generalized conditional entropy we obtain Bayesian networks that have superior scores compared to similar structures obtained by classical inference methods.

1 Introduction

A Bayesian Belief Network (BBN) structure is a directed acyclic graph which represents probabilistic dependencies among a set of random variables.

Inducing a BBN structure for the set of attributes of a dataset is a well known problem and a challenging one due to enormity of the search space. The number of possible BBN structures grows super-exponentially with respect to the number of the nodes.

In Cooper and Herskovits (1993), where the K2 heuristic algorithm is introduced, a measure of the quality of the structure is derived based on its posterior probability in presence of a dataset. An alternative approach to compute a BBN structure is based on the Minimum Description Length principle (MDL) first introduced in Rissanen (1978). The algorithms of Lam and Bacchus (1994) and Suzuki (1999) are derived from this principle.

We propose a new approach to inducing BBN structures from datasets based on the notion of β-generalized entropy (β-GE) and its corresponding β-generalized conditional entropy (β-GCE) introduced in Havrda and Charvat (1967) and axiomatized in Simovici and Jaroszewicz (2002) as a one-parameter family of functions defined on partitions (or probability distributions). The flexibility that ensues allows us to generate BBNs with better scores than published results.

One important advantage of our approach is that, unlike Cooper and Herskovits (1993) it is not based on any distributional assumption for developing the formula.

2 Generalized Entropy and Structure Inference

The set of partitions of a set S is denoted by $\mathsf{PART}(S)$. The *trace of a partition* π on a subset T of S is the partition $\pi_T = \{T \cap B_i \mid i \in I \text{ and } T \cap B_i \neq \emptyset\}$ of T. The usual order between set partitions is denoted by "$\leq$". It is well-known that $(\mathsf{PART}(S), \leq)$ is a bounded lattice. The infimum of two partitions π and $\pi' = \{B_j | j \in J\}$ on S, denoted with $\pi \wedge \pi'$, is the partition $\{B_i \cap B_j | i \in I, j \in J, B_i \cap B_j \neq \emptyset\}$ on S . The least element of this lattice is the partition $\alpha_S = \{\{s\} \mid s \in S\}$; the largest is the partition $\omega_S = \{S\}$.

The notion of generalized entropy or β-entropy was introduced in Havrda and Charvat (1967) and axiomatized for partitions in Simovici and Jaroszewicz (2002). If S is a finite set and $\pi = \{B_1, \ldots, B_m\}$ is a partition of S, the β-entropy of π is the number $\mathcal{H}_\beta(\pi) = \frac{1}{1-2^{1-\beta}} \left(1 - \sum_{i=1}^{m} \left(\frac{|B_i|}{|S|}\right)^\beta\right)$ for $\beta > 1$. The Shannon entropy is obtained as $\lim_{\beta \to 1} \mathcal{H}_\beta(\pi)$.

For $\beta \geq 1$ the function $\mathcal{H}_\beta : \mathsf{PART}(S) \longrightarrow \mathbb{R}_{\geq 0}$ is anti-monotonic. Thus, $\mathcal{H}_\beta(\pi) \leq \mathcal{H}_\beta(\alpha_S) = \frac{1-n^{\beta-1}}{(2^{1-\beta}-1) \cdot n^{\beta-1}}$, where $n = |S|$.

Let $\pi, \sigma \in \mathsf{PART}(S)$ be two partitions, where $\pi = \{B_1, \ldots, B_m\}$ and $\sigma = \{C_1, \ldots, C_n\}$. The β-conditional entropy of π and σ is $\mathcal{H}_\beta(\pi|\sigma) = \sum_{j=1}^{n} \left(\frac{|C_j|}{|S|}\right)^\beta \mathcal{H}_\beta(\pi_{C_j})$. It is immediate that $\mathcal{H}_\beta(\pi|\omega_S) = \mathcal{H}_\beta(\pi)$ and that $\mathcal{H}(\pi|\alpha_S) = 0$. Also, in Simovici and Jaroszewicz (2006) it is shown that $\mathcal{H}_\beta(\pi|\sigma) = \mathcal{H}_\beta(\pi \wedge \sigma) - \mathcal{H}_\beta(\sigma)$, a property that extends the similar property of Shannon entropy.

When $\beta \geq 1$ the β-GCE is dually anti-monotonic with respect to its first argument and is monotonic with respect to its second argument. Moreover, we have $\mathcal{H}_\beta(\pi|\sigma) \leq \mathcal{H}_\beta(\pi)$.

Let $\mathcal{D}$ be a dataset with set of attributes $\mathbf{Attr}(\mathcal{D})$. The domain of attribute $A_i \in \mathbf{Attr}(\mathcal{D})$ is $\mathrm{Dom}(A_i)$. The projection of a tuple $t \in \mathcal{D}$ on X is the restriction $t[X]$ of t to the set X. The set of attributes X defines a partition π^X on $\mathcal{D}$, which groups together the tuples that have the equal projections on X.

Let A be an attribute and let X be set of parents for A, where $\mathrm{Dom}(A) = \{v_1, v_2, ..., v_n\}$ and $\mathrm{Dom}(X) = \prod_{B \in X} \mathrm{Dom}(B) = \{u_1, u_2, ..., u_m\}$. Define $p_{ij} = P(t[A] = v_i | t[X] = u_j)$. We have $\frac{1}{n^{\beta-1}} \leq \sum_{i=1}^{n} p_{ij}^\beta \leq 1$ for $\beta \geq 1$. X is considered as a "good" parent set for A if knowing the its value enables us to predict the value of A with a high probability, that is, if $a_j = \sum_{i=1}^{n} p_{ij}^\beta$ is close to 1 for every j where $P(t[X] = u_j)$ is sufficiently large. Clearly, X is a "perfect" parent if $\sum_{j=1}^{m} a_j = m$. The β-GCE captures exactly this parenthood quality measure. Indeed, suppose that $\pi^A = \{B_i | 1 \leq i \leq n\}$ and $\pi^X = \{C_j | 1 \leq j \leq m\}$, where for $t \in B_i$ we have $t[A] = v_i$, and for $s \in C_j$ we have $s[X] = u_j$. Then, $p_{ij} = P(t[A] = v_i | t[X] = u_j) = P(t \in B_i | t \in C_j) = \frac{|B_i \cap C_j|}{|C_j|}$, which implies $\mathcal{H}_\beta(\pi^A|\pi^X) = \frac{1}{1-2^{1-\beta}} \sum_{j=1}^{m} P^\beta(C_j)\,(1 - a_j)$. Thus, minimizing $\mathcal{H}_\beta(\pi^A|\pi^X)$ amounts to reducing the values of $(1 - a_j)$ as much as possible for those j's where $|C_j|$ is large, that is, $P(C_j) = P(t[X] = u_j)$ is non-trivial. We refer to quantity $\mathcal{H}_\beta(\pi^A|\pi^X)$ as the *entropy of node A in presence of set X*. However, even if $X = argmin_X(\mathcal{H}_\beta(\pi^A|\pi^X))$, the value of the minimum itself may be too high to insure good predictability. An alternative is to measure the reduction of entropy of node A as a result of presence of set X as $\frac{\mathcal{H}_\beta(\pi^A|\pi^X)}{\mathcal{H}_\beta(\pi^A)}$. Since $0 \leq \mathcal{H}_\beta(\pi^A|\pi^X) \leq \mathcal{H}_\beta(\pi^A)$ we have $0 \leq \frac{\mathcal{H}_\beta(\pi^A|\pi^X)}{\mathcal{H}_\beta(\pi^A)} \leq 1$. If X is a perfect parent set for A, then $a_j = 1$ for $1 \leq j \leq m$, so $H_\beta(\pi^A|\pi^X) = 0$.

Let $\epsilon \in [0, 1]$ be a number referred to as *prediction threshold*. We regard X as a *ϵ-suitable* parent of A if $\frac{\mathcal{H}_\beta(\pi^A|\pi^X)}{\mathcal{H}_\beta(\pi^A)} \leq \epsilon$.

To avoid cycles in the network we start from a sequence of attributes $A_1, A_2, ..., A_p$ and we seek the set of parents for A_i in the set $\Phi(A_i) = \{A_1, \ldots, A_{i-1}\}$, a frequent assumption

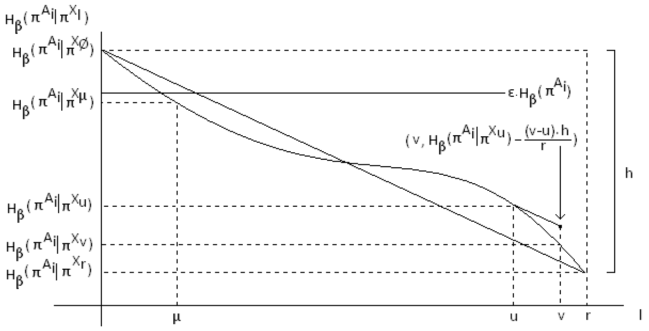


FIG. 1 – *Visualization of the Algorithm*

(see Cooper and Herskovits (1993); Suzuki (1999)). In addition, we set a bound r on the maximum number of parents. The set $\Phi(A_i)$ may contain many subsets that are ϵ-suitable. A possible solution is to choose an ϵ-suitable parent set $X \subseteq \Phi(A_i)$ with minimum β-GCE $H_\beta(\pi^A|\pi^X)$. By the monotonicity property of β-GCE with respect to second argument we have $\mathcal{H}_\beta(\pi^{A_i}|\pi^{\Phi(A_i)}) \leq \mathcal{H}_\beta(\pi^{A_i}|\pi^{\{A_1,A_2,\ldots A_{i-2}\}}) \leq \cdots \leq \mathcal{H}_\beta(\pi^{A_i}|\pi^{\{A_1\}}) \leq \mathcal{H}_\beta(\pi^{A_i})$. Then, for a given ϵ, if X has the minimum $\mathcal{H}_\beta(\pi^{A_i}|\pi^X)$ among all ϵ-suitable parents of A_i, then X has the maximum possible size. To simplify the structure, we trade some predictability for simplicity by adopting a heuristic approach which finds a minimal set of parents for a node with highest possible reduction of entropy of that child node on its presence.

Define $\Theta^\epsilon_l(A_i) = \{X \subseteq \Phi(A_i)|X$ is an ϵ-suitable parent of A_i and $|X| = l\}$ and $\mu = min\{n \in \mathbb{N}|\Theta^\epsilon_n(A_i) \neq \emptyset\}$. When $\mu \leq r$, we have the sequence of nonempty collections of sets of attributes $\Theta^\epsilon_\mu(A_i), \Theta^\epsilon_{\mu+1}(A_i), ..., \Theta^\epsilon_r(A_i)$ by the monotonicity property of β-GCE.

Let $X_\ell = argmin_{X\in\Theta^\epsilon_\ell(A_i)}(\mathcal{H}_\beta(\pi^{A_i}|\pi^X))$ be the first set of size ℓ (in lexicographical order) that minimizes $\mathcal{H}_\beta(\pi^{A_i}|\pi^X)$. We limit our parent search to the sequence of sets $\mathcal{S} = (X_\mu, X_{\mu+1}, \ldots, X_r)$, where the sets are listed in increasing order of size. For the sequence $\mathcal{S} = (X_\mu, X_{\mu+1}, \ldots, X_r)$ defined above we have $\mathcal{H}_\beta(\pi^{A_i}|\pi^\mu) \geq \mathcal{H}_\beta(\pi^{A_i}|\pi^{X_{\mu+1}}) \geq \cdots \geq \mathcal{H}_\beta(\pi^{A_i}|\pi^{X_r})$. The set of points $\{(0, \mathcal{H}_\beta(\pi^{A_i}))\} \cup \{(p, \mathcal{H}_\beta(\pi^{A_i}|\pi^{X_p})) \mid \mu \leq p \leq r\}$ in $\mathbb{R}^2$ can be placed on a non-increasing curve with height $h = \mathcal{H}_\beta(\pi^{A_i}) - \mathcal{H}_\beta(\pi^{A_i}|\pi^{X_r})$ as shown in Figure 1. We initialize the current parent set X_u to $\emptyset$ and iterate over members of $\mathcal{S}$ in increasing order of their size. The member $X_v \in \mathcal{S}$ leads to a nontrivial improvement in predictability over X_u if $\frac{\mathcal{H}_\beta(\pi^{A_i}|\pi^{X_u})-\mathcal{H}_\beta(\pi^{A_i}|\pi^{X_v})}{\mathcal{H}_\beta(\pi^{A_i})-\mathcal{H}_\beta(\pi^{A_i}|\pi^{X_r})} \geq \frac{v-u}{r}$. This happens if the decrease in $\mathcal{H}_\beta(\pi^{A_i}|\pi^{X_\ell})$ when the parent set of A_i is changed from X_u to X_v is greater than or equal to linear decrease with respect to the two end points of the corresponding non-increasing curve as shown in Figure 1. The end points of the curve are $(0, \mathcal{H}_\beta(\pi^{A_i}))$ and $(r, \mathcal{H}_\beta(\pi^{A_i}|\pi^{X_r}))$ and the linear decrease with respect to two end points of the curve when we move from u to v on x-axis which correspond to parent sets X_u and X_v is $\frac{h\cdot(v-u)}{r} = \frac{(\mathcal{H}_\beta(\pi^{A_i})-\mathcal{H}_\beta(\pi^{A_i}|\pi^{X_r}))\cdot(v-u)}{r}$. Note that $v = u + w$ where $1 \leq w \leq r - u$. This suggests that we do not stop the process if X_{u+1} does not satisfy the above inequality since there may be a parent set $X_v \in \mathcal{S}$ where $v > u + 1$ with non-trivial improvement in predictability with respect to current parent set X_u.

Algorithm 1: BuildBayesNet

```
input  : Dataset $\mathcal{D}$, Real $\beta, \epsilon, r$
// $\epsilon \in [0, 1]$ is the prediction threshold.
// $\beta \geq 1$ is the parameter for $\beta$-entropy.
// $r$ is the maximum number of parents.
// $\mathbf{Attr}(\mathcal{D})$ is a list of attributes of $\mathcal{D}$ where if
// $1 \leq i < j \leq |\mathbf{Attr}(\mathcal{D})|$ the $i$th element of the list can
// be a parent of $j$th element, but not vice versa.
output : A Network Structure for $\mathcal{D}$
NetworkStructure $\mathcal{N}$
for $i \leftarrow |\mathbf{Attr}(\mathcal{D})|$ to $1$ do
    Node $A_i \leftarrow \mathbf{Attr}(\mathcal{D})[i]$;
    Integer $\mu \leftarrow 0, m \leftarrow \min(r, i-1)$
    Real $H[m+1]$
    Set $\mathcal{S}[m+1]$
    $H[0] \leftarrow \mathcal{H}_\beta(\pi^{A_i})$
    for $j \leftarrow m$ to $1$ do
        Compute $\Theta_j^\epsilon(A_i)$
        if $\Theta_j^\epsilon(A_i) = \emptyset$ then
            break
        else
            $\mathcal{S}[j] \leftarrow argmin_{x \in \Theta_j^\epsilon(A_i)}(\mathcal{H}_\beta(\pi^{A_i}|\pi^x))$
            $H[j] \leftarrow \mathcal{H}_\beta(\pi^{A_i}|\pi^{\mathcal{S}[j]})$
            $\mu \leftarrow j$
    $\mathcal{N}.addNode(A_i)$
    if $\mu \neq 0$ then
        Integer $u \leftarrow 0$
        for $v \leftarrow \mu$ to $m$ do
            if $\frac{H[u]-H[v]}{v-u} \geq \frac{H[0]-H[m]}{m}$ then
                $u \leftarrow v$
        forall $x \in \mathcal{S}[u]$ do
            $\mathcal{N}.addEdge(x \rightarrow A_i)$
return $\mathcal{N}$;  //end of algorithm
```

The increase in size of the parent set is penalized by making the condition stricter for larger parent sets. Also, if none of the parent sets in $\mathcal{S}$ of size μ to $r-1$ satisfy the inequality, then X_r will.

3 Experimental Results

We compared the generated results with well-known Bayesian structures in literature using two scoring schemes, MDL used by Lam and Bacchus (1994) and Suzuki (1999) and the scoring method of Cooper and Herskovits (1993). Experiments involved the Brain Tumor dataset (Cooper (1984)), the Breast Cancer (Blake et al. (1998a)), ALARM (Beinlich et al. (1989)), and IRIS (Blake et al. (1998b)). The experimental results are presented in Table 1. The last row of each table contains the two scores for published structures (according to Williams and Williamson (2006) and Beinlich et al. (1989)). We assume that the distribution on priors of the structures for a given dataset is uniform Cooper and Herskovits (1993). Experiments were performed on a machine with 64-bit Intel Xeon processor.

The scores for generated network structures depends on β and ϵ and in many cases is better than the scores for established structures (C-H scores are higher and MDL scores are lower). Figure 2 represents four different structures for Brain Tumor dataset. Structure A is the one

TAB. 1 – *Experimental Results*

Generated Structures			10000 rows		
β	ϵ	r	log(C-H Score)	MDL Score	Time(ms)
1.0	1.0	3	-7483	13631.52	57
1.0	0.8	2	-7506	13474.37	51
1.6	0.7	2	-7588	13680.31	45
2.1	0.5	3	-7588	13693.21	55
Original Structure			-8115	14410.10	-

Brain Cancer Results

Generated Structures			286 rows		
β	ϵ	r	log(C-H Score)	MDL Score	Time(ms)
1.1	0.5	2	-1172	3210.22	144
1.0	0.6	3	-1197	8640.41	202
1.7	0.3	2	-1207	3669.88	121
1.8	0.7	3	-1214	3859.67	196
1.0	0.5	3	-1215	3511.35	202
1.2	0.4	2	-1224	4968.50	133
1.0	0.7	3	-1256	13667.40	202
Original Structure			-1201	4142.03	-

Breast Cancer Results

Generated Structures			20002 rows		
β	ϵ	r	log(C-H Score)	MDL Score	Time(s)
1.2	0.5	3	-114931	270298.25	542
1.2	0.5	4	-114981	271590.92	12801
1.2	0.6	4	-116081	272665.06	12802
1.1	0.7	3	-116914	271469.89	546
Original Structure			-159306	378518.37	-

Alarm Results

Generated Structures			150 rows		
β	ϵ	r	log(C-H Score)	MDL Score	Time(ms)
1.0	0.4	2	-902	127543.87	109
1.8	0.7	3	-905	13279.40	173
Original Structure			-932	261481.02	-

Iris Results

introduced by G. F. Cooper. Structures B $(\beta = 1.0, \alpha = 1.0, r = 3)$, C$(\beta = 1.0, \alpha = 0.8, r = 2)$ and D$(\beta = 2.1, \alpha = 0.5, r = 3)$ are the ones generated by our approach.

4 Conclusions

We developed an approach for generating a Bayesian network structure from data based on notion of generalized entropy.

The best parent-child relationships among attributes is obtained at values of β that are highly dependent on the data set, a fact that suggests that the GCE approach is preferable to using the Shannon entropy.

References

Beinlich, I., H. Suermondt, R. Chavez, and G. F. Cooper (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. Technical Report KSL-88-84, Stanford University, Knowledge System Laboratory.

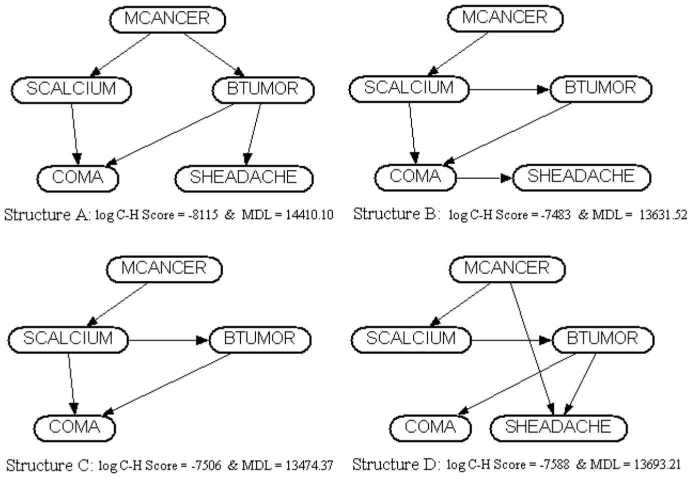
Blake, C., D. Newman, S. Hettich, and C. Merz (1998a). UCI repository of machine learning databases. A dataset from Ljubljana Oncology Institute provided by UCI, available at http://www.ics.uci.edu/ mlearn/MLRepository.html.

Blake, C., D. Newman, S. Hettich, and C. Merz (1998b). UCI repository of machine learning databases. A dataset created by R.A. Fisher and donated by Michael Marshall, available at http://www.ics.uci.edu/ mlearn/MLRepository.html.

Cooper, G. F. (1984). *NESTOR: A computer-based medical diagnosis aid that integrates casual and probabilistic knowledge*. Ph. D. thesis, Stanford University.

Cooper, G. F. and E. Herskovits (1993). A Bayesian method for the induction of probabilistic networks from data. Technical Report KSL-91-02, Stanford University, Knowledge System Laboratory.

Havrda, J. H. and F. Charvat (1967). Quantification methods of classification processes: Concepts of structural α entropy. *Kybernetica 3*, 30–35.



Structure A: log C-H Score = -8115 & MDL = 14410.10

Structure B: log C-H Score = -7483 & MDL = 13631.52

Structure C: log C-H Score = -7506 & MDL = 13474.37

Structure D: log C-H Score = -7588 & MDL = 13693.21

FIG. 2 – *Brain Tumor Structures*

Lam, W. and F. Bacchus (1994). Learning Bayesian belief networks: An approach based on the MDL principle. *Computational Intelligence 10*, 269–293.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica 14*, 456–471.

Simovici, D. A. and S. Jaroszewicz (2002). An axiomatization of partition entropy. *Transactions on Information Theory 48*, 2138–2142.

Simovici, D. A. and S. Jaroszewicz (2006). A new metric splitting criterion for decision trees. *International Journal of Parallel, Emergent and Distributed Systems 21*, 239–256.

Suzuki, J. (1999). Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique. *IEICE Trans. Information and Systems*, 356–367.

Williams, M. and J. Williamson (2006). Combining argumentation and Bayesian nets for breast cancer prognosis. *Journal of Logic, Language and Information 15*, 155–178.

Résumé

Nous proposons un nouvel algorithme pour extraire la structure d'un réseau Bayésien d'un ensemble de données. Notre approche est basée sur les entropies conditionnelles généralisées, une famille conditionnelle d'entropies qui étend l'entropie conditionnelle de Shannon.Nos résultats indiquent que, avec un choix approprié d'une entropie conditionnelle généralisée, nous obtenons des réseaux Bayésiens qui ont des scores supérieurs aux structures similaires obtenues par des méthodes classiques d'inférence.