Large Data Sets, Conditional Entropy and the Cooper-Herskovitz Bayesian Score

Saaid Baraty and **Dan A. Simovici**
University of Massachusetts, Boston, MA 02125

Abstract

We examine the relationship between the Cooper-Herskovitz score of a Bayesian network and the conditional entropies of the nodes of the networks conditioned on the probability distributions of their parents. We show that minimizing the conditional entropy of each node of the BNS conditioned on its set of parents amounts to maximization of the CH score.

The main result is a lower bound on the size of the data set that ensures that the divergence of between conditional entropy and the Cooper-Herskovitz score is under a certain threshold.

1. Introduction

The construction of a Bayesian Network Structure (BNS) from a data set that captures the probabilistic dependencies among the attributes of the data set has been one of the prominent problems among community of uncertainty researchers since early 90s. The problem is particularly challenging due to enormity of number of possible structures for a given collection of data.

Formally, a *Bayesian Belief Network* is a pair $(\mathcal{B}_s, \mathcal{B}_p)$, where $\mathcal{B}_s$ is a directed acyclic graph, commonly referred to as a *Bayesian Network Structure* (BNS), and $\mathcal{B}_p$ is a collection of distributions which quantifies the probabilistic dependencies present in the structure, as we discuss in detail in the next paragraph. Each node of the BNS corresponds to a random variable; edges represent probabilistic dependencies among these random variables. A BNS captures the split of the joint probability of a set of random variables, presented by its nodes, into a product of probabilities of its nodes conditioned upon a set of other nodes, namely the set of its *predecessors* or *parents*.

The set of values (or states) of a random variable Z is referred to as the *domain of Z*, borrowing a term from relational databases. This set is denoted by $\mathrm{Dom}(Z)$.

If a random variable X is a node of $\mathcal{B}_s$ with $\mathrm{Dom}(X) = \{1, \cdots, R_X\}$ and set of random variables $Pa_X = \{Y_1, Y_2, \cdots, Y_k\}$ as its set of parents, and if we agree upon some enumeration of set $\mathrm{Dom}(Pa_X) = \prod_{i=1}^{k} \mathrm{Dom}(Y_i)$, then denote by θ_{lj}^X the conditional probability $P(X = l | Y_1 = y_1, Y_2 = y_2, \cdots, Y_k = y_k)$, where l is some state of X and $(y_1, \cdots, y_k)$ is the j^{th} element of the enumeration. Also, we denote with $\theta_{\cdot j}^X$, the probability distribution of X conditioned on its set of parents taking on the j^{th} assignment of its domain. $\mathcal{B}_p$ is collection of distributions $\theta_{\cdot j}^X$ for all nodes X of $\mathcal{B}_s$ and $1 \leq j \leq |\mathrm{Dom}(Pa_X)|$.

A Bayesian network structure for a data set is a BNS in which the nodes are labelled with its attributes treated like random variables. Such a BNS is said to be fit for the data set if the probabilistic dependencies (or independencies) captured by the structure closely reflects the dependencies (or independencies) among the attributes according to that data set. Several scoring solutions have been proposed for evaluating this fitness. All these scorings schemes are based on three major approaches: scores based on maximization of the posterior probability of the structure conditioned upon data, scores based on MDL (Minimum Description Length) principle and scores based on minimization of conditional entropy.

The first approach was initially introduced in (Cooper and Herskovits 1993), where the scoring formula was derived based on a number of assumptions such as assuming that the distribution of tuples $(\theta_{1j}^X, \cdots, \theta_{R_X j}^X)$ is uniform for all X and j, or is a Dirichlet distribution. We refer to the scoring criterion introduced in (Cooper and Herskovits 1993) as the *CH score*. In (Heckerman, Geiger, and Chickering 1995) Heckerman et al. replaced the Dirichlet distribution assumption by the *likelihood equivalence* assumption and argued that under this assumption tuples $(\theta_{1j}^X, \cdots, \theta_{R_X j}^X)$ obey a Dirichlet distribution.

The second approach is based on the minimum description length principle, introduced in (Rissanen 1978). Later, (Lam and Bacchus 1994) this principle was applied on learning a BNS from data. The application of same methods on learning the local structure in the conditional probability distributions with variable number of parameters that quantify these networks was suggested in (Friedman and Goldszmidt 1998).

Another approach was introduced in (Simovici and Baraty 2008) and extended in (Baraty and Simovici 2009) to be used in evaluating the edges of a Bayesian structure and pruning unimportant edges. In this approach one tries to minimize the entropy of each child node conditioned on its set of parents.

The first two methods and in particular the first one

are computationally expensive, while the third approach is cheaper to calculate and the resulting numbers are in a more manageable range. (Suzuki 1999) showed the close relationship between the MDL scheme and CH score.

Our main goal in this paper is to show the relationship between conditional entropy and CH score. In particular, we show that minimization of conditional entropy and maximization of the CH score yield the same result if the data set at hand is large and this is precisely the case when the CH score is computationally impractical. Thus, in case the data set is large, using entropy makes more sense. Also, we obtain a lower bound of the size of the data set necessary for substituting entropy measure with CH score for inferring a BNS with a good prediction capability.

In Section 2. we examine the relationship between CH score and conditional entropy. A lower bound on size of data set is obtained in Section 3. Experimental results are presented in Section 4.The final section contains the conclusion of this paper.

2. Equivalence of CH Score and Conditional Entropy

Let $\mathcal{D}$ be a data set with set of attributes $\mathbf{Attr}(\mathcal{D}) = \{A_1, \cdots, A_n\}$. If t is a tuple of $\mathcal{D}$ and L a subset of $\mathbf{Attr}(\mathcal{D})$, the restriction of the tuple t to L is denoted by $t[L]$; we refer to $t[L]$ as the *projection of* t *on* L. For each $A_i \in \mathbf{Attr}(\mathcal{D})$, define the *active domain* of attribute A in $\mathcal{D}$ to be

$$\mathrm{Adom}_{\mathcal{D}}(A_i) = \{a_i^1, \cdots, a_i^{v_i}\}.$$

The notion of active domain is extended to sets of attributes by defining $\mathrm{Adom}_{\mathcal{D}}(L) = \{t[L] \mid t \in \mathcal{D}\}$.

A *partition* of a finite set S is non-empty collection of pairwise disjoint, non-empty subsets of S, $\pi = \{B_i | i \in I\}$, such that $\bigcup_{i\in I} B_i = S$.

If $\pi = \{B_1, \ldots, B_m\}$ is a partition of S, its entropy is the number

$$\mathcal{H}_p(\pi) = -\sum_{i=1}^{m} \frac{|B_i|}{|S|} \log_2 \frac{|B_i|}{|S|},$$

which corresponds to the Shannon entropy of a probability distribution $(p_1, \ldots, p_m)$, where $p_i = \frac{|B_i|}{|S|}$ for $1 \le i \le m$. If $\sigma = \{C_1, \ldots, C_k\}$ is another partition on S, then, the *entropy of* π *conditioned on* σ is the number,

$$\mathcal{H}_p(\pi|\sigma) = -\sum_{i=1}^{m}\sum_{j=1}^{k} \frac{|B_i \cap C_j|}{|S|} \log \frac{|B_i \cap C_j|}{|C_j|}.$$

It is known that $0 \le \mathcal{H}_p(\pi|\sigma) \le \mathcal{H}_p(\pi)$.

A similar notion to partition entropy is the *entropy of a finite set of natural numbers*. If $U = \{n_1, \cdots, n_q\} \subseteq \mathbb{N}$, then the entropy of set U is defined as

$$\mathcal{H}_n(U) = \mathcal{H}_n(n_1, \cdots, n_q) = -\sum_{i=1}^{q} \frac{n_i}{\sum_{i=1}^{q} n_i} \log_2 \frac{n_i}{\sum_{i=1}^{q} n_i}$$

Definition 1 The equivalence relation "$\sim_{\mathbf{A}_I}$" defined by the sequence of attributes $\mathbf{A}$ on $\mathcal{D}$, consists of those pairs $(t, t') \in \mathcal{D}^2$ such that $t[\mathbf{A}] = t'[\mathbf{A}]$. The corresponding partition $\pi^{\mathbf{A}} \in \mathsf{PART}(\mathcal{D})$ is the *partition generated by* $\boldsymbol{A}$. □

Define the number $\mathsf{n}_{ijk}^{\mathcal{D}}$ to be the cardinality of the set, $\{t \in \mathcal{D} \mid t[A_i, \mathsf{Par}(A_i)] = (a_i^k, \mathbf{a}_i^j)\}$ where $a_i^k \in \mathrm{Adom}_{\mathcal{D}}(A_i)$ and $\mathbf{a}_i^j \in \mathrm{Adom}_{\mathcal{D}}(\mathsf{Par}(A_i)) = \{\mathbf{a}_i^1, \cdots, \mathbf{a}_i^{q_i}\}$. Also define $r_i = |\mathrm{Adom}_{\mathcal{D}}(\mathsf{Par}(A_i) \cup \{A_i\})|$ and v_i^j to be the cardinality of the set $\{1 \le k \le v_i \mid n_{ijk}^{\mathcal{D}} \ne 0\}$.

In general, a data set $\mathcal{D}$ can be regarded as a multiset of tuples. Let, $\mathsf{N}_{ij}^{\mathcal{D}} = \sum_{k=1}^{v_i} \mathsf{n}_{ijk}^{\mathcal{D}}$. Note that $\mathsf{N}_{ij}^{\mathcal{D}}$ is the number of tuples $t \in \mathcal{D}$ such that $t[\mathsf{Par}(A_i)] = \mathbf{a}_i^j$. When $\mathcal{D}$ is clear from context, we drop the $\mathcal{D}$ subscript or superscript from notations introduced above.

A BNS for data set $\mathcal{D}$ is a graph $\mathcal{B}_s$ with set of nodes $\mathbf{Attr}(\mathcal{D})$ and its set of edges a subset of $\mathbf{Attr}(\mathcal{D}) \times \mathbf{Attr}(\mathcal{D})$. The attributes of the data set are treated as random variables. The BNS represents probabilistic dependencies among these attributes. We denote by $\mathsf{Par}_{\mathcal{B}_s}(A_i)$, the set of parent nodes of A_i in $\mathcal{B}_s$. The subscript $\mathcal{B}_s$ is omitted whenever possible.

Define $\mathsf{BNS}(\mathcal{D})$ to be the set of all possible structures for $\mathcal{D}$. Also denote by $\theta_{ijk} = P(A_i = a_k | \mathsf{Par}(A_i) = \mathbf{a}_j)$.

Let $\mathbf{A} = (A_1, \ldots, A_n)$ be a list of $\mathbf{Attr}(\mathcal{D})$ which represents expert's prior knowledge of the domain in the following way: attribute A_i is in the set of candidate parents for A_j, but not vice versa if $i < j$. We denote by $\mathsf{BNS}_{\mathbf{A}}(\mathcal{D})$ the set of all structures for $\mathcal{D}$ conforming to the ordering $\mathbf{A}$.

Definition 2 The *complete BNS for the list* $\boldsymbol{A}$ is the structure $\mathcal{B}_{cs}^{\mathbf{A}}$ in which $\mathsf{Par}(A_i) = \{A_1, \cdots, A_{i-1}\}$ for $1 \le i \le n$. □

(Cooper and Herskovits 1993) use the probability $P(\mathcal{B}_s, \mathcal{D})$ as a score of the fitness of the structure $\mathcal{B}_s$ in representing the probabilistic dependencies among attributes of $\mathcal{D}$. They assume the tuple $(\theta_{ij1}, \cdots, \theta_{ij(v_i-1)})$ has a *Dirichlet* distribution with parameters $((\mathsf{n}'_{ij1}+1), \cdots, (\mathsf{n}'_{ijv_i}+1))$ for all $1 \le i \le n$ and $1 \le j \le q_i$. Based on this, and a number of other assumptions they show that

$$P(\mathcal{B}_s, \mathcal{D}) = P(\mathcal{B}_s) \cdot \prod_{i=1}^{n} f_i, \tag{1}$$

where

$$f_i = \prod_{j=1}^{q_i} \frac{(\mathsf{N}'_{ij} + v_i - 1)!}{(\mathsf{N}_{ij} + \mathsf{N}'_{ij} + v_i - 1)!} \cdot \prod_{k=1}^{v_i} \frac{(\mathsf{n}_{ijk} + \mathsf{n}'_{ijk})!}{\mathsf{n}'_{ijk}!},$$

and $\mathsf{N}'_{ij} = \sum_{k=1}^{v_i} \mathsf{n}'_{ijk}$. Observe that $\sum_{j=1}^{q_i} \mathsf{N}_{ij} = |\mathcal{D}|$, $\mathsf{N}_{ij} \ge 1$ and $0 \le \mathsf{n}_{ijk} \le \mathsf{N}_{ij}$ for $1 \le i \le n$, $1 \le j \le q_i$ and $1 \le k \le v_i$.To find the fittest structure for $\mathcal{D}$, we seek a structure $\mathcal{B}_s$ over $\mathbf{Attr}(\mathcal{D})$ that maximizes $P(\mathcal{B}_s, \mathcal{D})$. That is, we need to find,

$$\mathrm{argmax}_{\mathsf{Par}(A_1),\cdots,\mathsf{Par}(A_n)} P(\mathcal{B}_s, \mathcal{D}).$$

Since $\ln(x)$ is a strictly increasing function we have

$$\begin{aligned} &\mathrm{argmax}_{\mathsf{Par}(A_1),\cdots,\mathsf{Par}(A_n)}(P(\mathcal{B}_s, \mathcal{D})) \\ &\quad = \quad \mathrm{argmax}_{\mathsf{Par}(A_1),\cdots,\mathsf{Par}(A_n)}(\ln P(\mathcal{B}_s, \mathcal{D})) \end{aligned}$$

for $1 \le i \le n$.

Define $g_i = \ln(f_i)$. By Equation (1), we have

$$\ln(P(\mathcal{B}_s, \mathcal{D})) = \ln P(\mathcal{B}_s) + \sum_{i=1}^{n} g_i. \quad (2)$$

Next we establish lower and upper bounds for g_i.

Theorem 3 *Let $\mathsf{LO}(g_i)$ and $\mathsf{UP}(g_i)$ be the numbers defined by*

$$\mathsf{LO}(g_i) = \alpha_i - \Phi_1 + \Phi_6 - \Phi_2 - \Phi_3 - \Phi_7 - \Phi_8$$

and

$$\mathsf{UP}(g_i) = \alpha'_i + \Phi_1 + \Phi_6 + \Phi_3 - \Phi_7 + \Phi_4 - \Phi_8$$

where $\Phi_1, \cdots, \Phi_8$ are defined in Table 1 and α_i and α'_i are given by

$$\begin{aligned}
\alpha_i &= 4q_i + r_i - 2\sum_{j=1}^{q_i} \ln(a_i^j) - \sum_{j=1}^{q_i} \ln(\mathsf{N}'_{ij} + 1) \\
&\quad - \sum_{j=1}^{q_i} v_i^j \ln(b_i^j) - \sum_{j=1}^{q_i}\sum_{k=1}^{v_i^j} \ln(\mathsf{n}'_{ijk} + 1) - \sum_{j=1}^{q_i} \ln(c_i^j) \\
&\quad + \sum_{j=1}^{q_i} \mathsf{N}'_{ij} \cdot \mathcal{H}_n(\mathsf{n}'_{ij1}, \cdots, \mathsf{n}'_{ijv_i^j}) - \sum_{j=1}^{q_i} \ln(d_i^j) \quad (3)
\end{aligned}$$

$$\begin{aligned}
\alpha'_i &= -4q_i - r_i + \sum_{j=1}^{q_i} \ln(u_i^j) + \sum_{j=1}^{q_i} \ln(v_i^j) \quad (4) \\
&\quad + \sum_{j=1}^{q_i} \ln(\mathsf{N}'_{ij} + 1) + \sum_{j=1}^{q_i} \mathsf{N}'_{ij} \cdot \mathcal{H}_n(\mathsf{n}'_{ij1}, \cdots, \mathsf{n}'_{ijv_i^j}) \\
&\quad + 2\sum_{j=1}^{q_i} \ln(w_i^j) + \sum_{j=1}^{q_i} v_i^j \ln(z_i^j) + \sum_{j=1}^{q_i} \ln(\mathsf{N}'_{ij} + v_i^j).
\end{aligned}$$

where $a_i^j, b_i^j, c_i^j, d_i^j, u_i^j, v_i^j, w_i^j$ and z_i^j are constants in the range $[2, 3)$ for all i and j. Then, we have,

$$\mathsf{LO}(g_i) \le g_i \le \mathsf{UP}(g_i).$$

Note that neither α_i nor α'_i depend on $|\mathcal{D}|$. Also, since $a_i^j, b_i^j, c_i^j, d_i^j, u_i^j, v_i^j, w_i^j$ and z_i^j are approximately equal to e, we have,

$$\begin{aligned}
\delta_i &= \alpha'_i - \alpha_i \\
&\approx 2\sum_{j=1}^{q_i} \ln(\mathsf{N}'_{ij} + 1) + \sum_{j=1}^{q_i} \ln(\mathsf{N}'_{ij} + v_i^j) \\
&\quad + \sum_{j=1}^{q_i}\sum_{k=1}^{v_i^j} \ln(\mathsf{n}'_{ijk} + 1) \\
&\le 2q_i \ln(\frac{\mathsf{N}'_{i\cdot}}{q_i} + 1) + q_i \ln(\frac{\mathsf{N}'_{i\cdot}}{q_i} + v_i) \\
&\quad + r_i \ln(\frac{\mathsf{N}'_{i\cdot}}{r_i} + 1) = \mathsf{UP}(\delta_i),
\end{aligned}$$

where $\mathsf{N}'_{i\cdot} = \sum_{j=1}^{q_i} \mathsf{N}'_{ij}$.

Table 1: Table of Notations

Symbol	Formula
Φ_1	$\sum_{j=1}^{q_i} \ln(\mathsf{N}_{ij} + 1)$
Φ_2	$\sum_{j=1}^{q_i} \ln(\mathsf{N}_{ij} + \mathsf{N}'_{ij} + v_i^j)$
Φ_3	$\sum_{j=1}^{q_i} \ln(\mathsf{N}_{ij} + \mathsf{N}'_{ij} + 1)$
Φ_4	$\sum_{j=1}^{q_i} \sum_{k=1}^{v_i^j} \ln(\mathsf{n}_{ijk} + \mathsf{n}'_{ijk} + 1)$
Φ_5	$\sum_{j=1}^{q_i} \ln(\mathsf{N}_{ij} + \mathsf{N}'_{ij} + v_i^j - 1)$
Φ_6	$\sum_{j=1}^{q_i} (\mathsf{N}_{ij} + \mathsf{N}'_{ij}) \cdot \mathcal{H}_n(\mathsf{N}_{ij}, \mathsf{N}'_{ij})$
Φ_7	$\sum_{j=1}^{q_i} [\ (\mathsf{N}_{ij} + \mathsf{N}'_{ij} + v_i^j - 1) \cdot \mathcal{H}_n(\mathsf{N}_{ij}, \mathsf{N}'_{ij} + v_i^j - 1)\ ]$
Φ_8	$\sum_{j=1}^{q_i} [\ (\mathsf{N}_{ij} + \mathsf{N}'_{ij}) \cdot \mathcal{H}_n((\mathsf{n}_{ij1} + \mathsf{n}'_{ij1}), \cdots, (\mathsf{n}_{ijv_i^j} + \mathsf{n}'_{ijv_i^j}))\ ]$

Theorem 4 *We have*

$$\lim_{|\mathcal{D}|\to\infty} \frac{\mathsf{LO}(g_i)}{\mathsf{UP}(g_i)} = 1.$$

Corollary 5 *Let $\mathcal{H}_p\left(\pi^{A_i} | \pi^{\mathsf{Par}(A_i)}\right)$ be the conditional entropy of partition generated by A_i given the partition generated by its set of parents. We have:*

$$\begin{aligned}
&\lim_{|\mathcal{D}|\to\infty} \ln(P(\mathcal{B}_s, \mathcal{D})) \\
&= \ln P(\mathcal{B}_s) - \lim_{|\mathcal{D}|\to\infty} \sum_{i=1}^{n} |\mathcal{D}| \cdot \mathcal{H}_p\left(\pi^{A_i} | \pi^{\mathsf{Par}(A_i)}\right).
\end{aligned}$$

We conclude that when we have a large data set, minimizing the conditional entropy of each node of the BNS conditioned on its set of parents amounts to maximization of the CH score. We refer to this modified optimization process as *CH to entropy substitution*.

3. Estimating the Size of the Data Set for CH to Entropy Substitution

In the previous section we proved that as the size of the data set tends to infinity the CH score and the conditional entropy are equivalent in the sense that a BNS that minimizes the conditional entropies maximizes the CH score by Corollary 5.

However, in practice, to optimize the process of finding a fit BNS for data by CH to entropy substitution, we need to find out how large the data set needs to be in order to make this substitution feasible. Note that a combination of Equation (2), Theorems 3 and 4 and Corollary 5 suggest that the maximum divergence of two measures, conditional entropy and CH score, for each i is:

$$\mathsf{UP}(g_i) - \mathsf{LO}(g_i).$$

To account for the magnitudes of the data set and the entropies of various nodes in order to have a meaningful magnitude for divergence, we define the divergence for node A_i as

$$\frac{\mathsf{UP}(g_i) - \mathsf{LO}(g_i)}{|\mathcal{D}| \cdot \mathcal{H}_p\left(\pi^{A_i} | \pi^{\mathsf{Par}(A_i)}\right)}.$$

Next, we add up the divergence for each node A_i in $\mathcal{B}_s$ to get the total divergence and we denote it with $\mathsf{DIV}(\mathcal{D}, \mathcal{B}_s)$:

$$\begin{aligned}
&\mathsf{DIV}(\mathcal{D}, \mathcal{B}_s) \\
= &\sum_{i=1}^{n} \frac{\mathsf{UP}(g_i) - \mathsf{LO}(g_i)}{|\mathcal{D}| \cdot \mathcal{H}_p\left(\pi^{A_i}|\pi^{\mathsf{Par}(A_i)}\right)} \\
= &\sum_{i=1}^{n} \frac{2\Phi_1 + 2\Phi_3 + \Phi_2 + \Phi_4 + \delta_i}{|\mathcal{D}| \cdot \mathcal{H}_p\left(\pi^{A_i}|\pi^{\mathsf{Par}(A_i)}\right)} \qquad (5) \\
\leq &\frac{1}{|\mathcal{D}|} \cdot \sum_{i=1}^{n} \frac{5\Phi_2 + \Phi_4 + \mathsf{UP}(\delta_i)}{\mathcal{H}_p\left(\pi^{A_i}|\pi^{\mathsf{Par}(A_i)}\right)} \quad (\Phi_1 \leq \Phi_3 \leq \Phi_2) \\
\leq &\frac{1}{|\mathcal{D}|} \cdot \sum_{i=1}^{n} \Bigg(\frac{5q_i \ln(\frac{|\mathcal{D}|+\mathsf{N}'_{i\cdot}}{q_i} + v_i)}{\mathcal{H}_p\left(\pi^{A_i}|\pi^{\mathsf{Par}(A_i)}\right)} \qquad (6) \\
&+ \frac{r_i \ln(\frac{|\mathcal{D}|+\mathsf{N}'_{i\cdot}}{r_i} + 1) + \mathsf{UP}(\delta_i)}{\mathcal{H}_p\left(\pi^{A_i}|\pi^{\mathsf{Par}(A_i)}\right)} \Bigg).
\end{aligned}$$

We refer to quantity 6 as *an upper bound on divergence of* $\mathcal{B}_s$ and denote it with $\mathsf{UP}(\mathsf{DIV}(\mathcal{D}, \mathcal{B}_s))$. We can determine the size of data set $\mathcal{D}$ such that $\mathsf{UP}(\mathsf{DIV}(\mathcal{D}, \mathcal{B}_s)) \leq \epsilon$ for some user given threshold $\epsilon > 0$. Note that this measure is dependent on the BNS $\mathcal{B}_s$ for which we want to compute the CH score or conditional entropy. But measures are being used to find a fit BNS. To avoid this circular dependency, we need to find the cardinality of $\mathcal{D}$ in such that $\mathsf{UP}(\mathsf{DIV}(\mathcal{D}, \mathcal{B}_s)) \leq \epsilon$ for all $\mathcal{B}_s$ in our set of candidate structures.

If $\mathbf{A}$, a list of $\mathbf{Attr}(\mathcal{D})$, reflects the prior knowledge of experts, as discussed in previous section, then $\mathsf{BNS}_{\mathbf{A}}(\mathcal{D})$ is our candidate space of structures and if we assume that $\mathsf{N}'_{i\cdot}$ is the same in all $\mathcal{B}_s \in \mathsf{BNS}_{\mathbf{A}}(\mathcal{D})$ for $1 \leq i \leq n$ we have the following theorem.

Theorem 6 *If* $\mathsf{DIV}(\mathcal{D}, \mathcal{B}^{\mathbf{A}}_{cs}) \leq \epsilon$*, then* $\mathsf{DIV}(\mathcal{D}, \mathcal{B}_s) \leq \epsilon$ *for all* $\mathcal{B}_s \in \mathsf{BNS}_{\mathbf{A}}(\mathcal{D})$.

The above theorem enables us to use $\mathsf{UP}(\mathsf{DIV}(\mathcal{D}, \mathcal{B}^{\mathbf{A}}_{cs}))$ as an upper bound of divergence within space $\mathsf{BNS}_{\mathbf{A}}(\mathcal{D})$ not dependent on structures.

Observe that $\ln x \leq x - 1$ for $x > 0$. It follows that

$$\ln(ax + b) \leq ax + b - 1,$$

so

$$\frac{\ln(ax+b)}{cx} \leq \frac{a}{c} + \frac{b-1}{cx}.$$

Therefore, if $\mathsf{H}(i) = \mathcal{H}_p\left(\pi^{A_i}|\pi^{\mathsf{Par}_{\mathcal{B}^{\mathbf{A}}_{cs}}(A_i)}\right)$,

$$\begin{aligned}
&\mathsf{UP}(\mathsf{DIV}(\mathcal{D}, \mathcal{B}^{\mathbf{A}}_{cs})) \\
= &\sum_{i=1}^{n} \Bigg(\frac{5q_i \ln(\frac{|\mathcal{D}|}{q_i} + \frac{\mathsf{N}'_{i\cdot}}{q_i} + v_i)}{|\mathcal{D}| \cdot \mathsf{H}(i)} \\
&+ \frac{r_i \ln(\frac{|\mathcal{D}|}{r_i} + \frac{\mathsf{N}'_{i\cdot}}{r_i} + 1) + \mathsf{UP}(\delta_i)}{|\mathcal{D}| \cdot \mathsf{H}(i)} \Bigg) \\
\leq &\sum_{i=1}^{n} \Bigg(\frac{5}{\mathsf{H}(i)} + \frac{5\mathsf{N}'_{i\cdot} + 5q_i(v_i - 1)}{|\mathcal{D}| \cdot \mathsf{H}(i)} \\
&+ \frac{1}{\mathsf{H}(i)} + \frac{\mathsf{N}'_{i\cdot}}{|\mathcal{D}| \cdot \mathsf{H}(i)} + \frac{\mathsf{UP}(\delta_i)}{|\mathcal{D}| \cdot \mathsf{H}(i)} \Bigg) \\
= &\sum_{i=1}^{n} \frac{6}{\mathsf{H}(i)} + \frac{1}{|\mathcal{D}|} \sum_{i=1}^{n} \frac{6\mathsf{N}'_{i\cdot} + 5q_i(v_i - 1) + \mathsf{UP}(\delta_i)}{\mathsf{H}(i)}.
\end{aligned}$$

The above number is an upper bound for $\mathsf{UP}(\mathsf{DIV}(\mathcal{D}, \mathcal{B}^{\mathbf{A}}_{cs}))$ which we denote with $\mathsf{UP}^2(\mathsf{DIV}(\mathcal{D}, \mathcal{B}^{\mathbf{A}}_{cs}))$ and to have it less than or equal to ϵ it suffices to have

$$|\mathcal{D}| \geq \frac{\sum_{i=1}^{n} \frac{6\mathsf{N}'_{i\cdot} + 5q_i(v_i-1) + \mathsf{UP}(\delta_i)}{\mathsf{H}(i)}}{\epsilon - \sum_{i=1}^{n} \frac{6}{\mathsf{H}(i)}}.$$

This establishes an explicit lower bound for $|\mathcal{D}|$.

Note that the above formula is obtained from several phases of amplifying the bound. So it may require a large data set in order to satisfy the inequality for some small threshold ϵ. Thus, if the inequality is not satisfied for a data set at hand, it does not necessarily mean that substitution of entropy for CH is not feasible. But we may need to resort to some randomized approaches in order to approximate a more realistic maximum divergence as we will see in the next section.

4. Experimental Results

We conducted two different experiments on three data sets, *Alarm*, *Neapolitan Cancer*, and *Breast Cancer* with 20002, 7565 and 277 rows with no missing values and 5, 37 and 10 attributes respectively. The attributes of the Neapolitan data set are all binary.

In the first experiment we computed the $\mathsf{UP}(\mathsf{DIV}(\mathcal{D}, \mathcal{B}^{\mathbf{A}}_{cs}))$ for the three data sets with different values for $|\mathcal{D}|$. We used the ordering of attributes of the three data sets, $\mathbf{A}_{AM}$, $\mathbf{A}_{NC}$ and $\mathbf{A}_{BC}$, which represent the prior knowledge of the domain from (Cooper 1984; Cooper and Herskovits 1993; Williams and Williamson 2006) respectively. To be able to compute the $\mathsf{UP}(\mathsf{DIV})$ for data set cardinalities greater than the actual size of the data set at hand , we make the simplifying assumption that the conditional entropy $\mathcal{H}_p\left(\pi^{A_i}|\pi^{\mathsf{Par}_{\mathcal{B}_s}(A_i)}\right)$ is relatively independent on the size of the database. This assumption is supported by experiments. Indeed, we show in Figure 1 the variation of several values of conditional entropy with respect to the size of the data set (obtained by random

extraction from the Neapolitan data set). Clearly, it is the case, that beyond a certain number of tuples this entropy is almost constant.

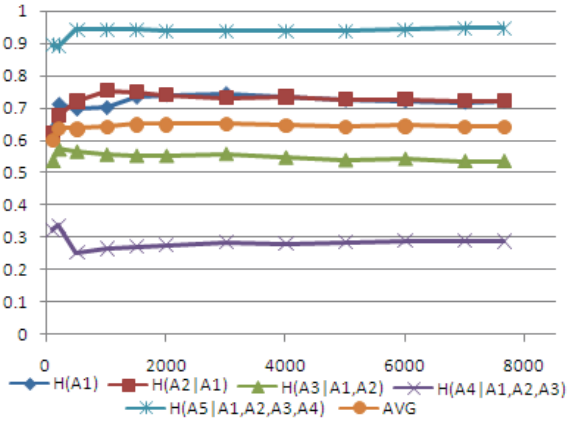

Figure 1: Variation of some conditional entropies with the size of the data set.

Also, we assume that $\mathsf{N}'_{ijk} = 0$ for all i, j and k. That is, the distribution of $(\theta_{ij1}, \cdots, \theta_{ij(v_i-1)})$ is uniform for all i and j. Table 2 represents the result of this experiment.

Table 2: DIV as a Function of Cardinality of Data Sets

$\lvert\mathcal{D}\rvert$	UPDIV$(\mathcal{D}, \mathcal{B}_{cs}^{\mathbf{A}_{AM}})$	UPDIV$(\mathcal{D}, \mathcal{B}_{cs}^{\mathbf{A}_{NC}})$	UPDIV$(\mathcal{D}, \mathcal{B}_{cs}^{\mathbf{A}_{BC}})$
277	3516294.597	5.000846603	437.554884
5000	215637.6066	0.497802902	63.79188858
7565	148993.292	0.355157238	47.15089133
10000	116996.2833	0.273901451	38.26594429
20002	65964.77078	0.15195053	22.41791235
30000	48007.49444	0.111300484	16.2496481
50000	32621.53866	0.07078029	10.74018886
75000	24172.33578	0.047520193	7.686459238
1.00E+05	19570.13933	0.036790145	6.045594429
2.00E+05	11735.87967	0.019765073	3.361797214
5.00E+05	5856.573866	0.008578029	1.523918886
1.00E+06	3395.911933	0.004539015	0.829959443
5.00E+06	903.9083866	0.001007803	0.197391889
1.00E+07	500.9891933	0.000549301	0.105195944
1.00E+08	66.45191933	6.39901E-05	0.012819594
5.00E+08	15.58238387	1.407E-05	0.002923919
1.00E+09	8.281191933	7.30791E-06	0.001511959
5.00E+09	1.882238387	1.5878E-06	0.000333062
1.00E+10	0.996119193	8.21501E-07	0.000173296
5.00E+10	0.223223839	1.7698E-07	3.78232E-05
1.00E+11	0.111611919	9.12231E-08	1.95916E-05
5.00E+11	0.025322384	1.9513E-08	4.23402E-06
1.00E+12	0.013161192	1.00295E-08	2.18496E-06

The first column is the cardinality of the data set which may be greater than the actual size of data. The second column is UP(DIV) for the Alarm data set based on sequence of attributes, $\mathbf{A}_{AM}$. The third and forth columns are UP(DIV)'s for Neapolitan Cancer and Breast Cancer data sets respectively. Note that the UP(DIV) for Neapolitan data set is much smaller than the UP(DIV) for Alarm for the same data set size. This deviation is due to the cardinality of the domain of the tuples of Alarm data set being much larger than that of Neapolitan. For the Neapolitan data set the upper bound on divergence is small for moderate data sizes even though this measure is very pessimistic and this guarantees a sound substitution of entropy for CH measure.

As we discussed in previous section, for some data sets to approximate a more realistic divergence we may need to resort to some randomized approach which is the motivation for our second experiment. Let us denote the Expression (5) with $\Delta(|\mathcal{D}|, \mathcal{B}_s)$ and denote by $P_{\mathcal{D}}$, the frequency extracted from $\mathcal{D}$. We can substitute N_{ij} by $|\mathcal{D}| \cdot \hat{P}(\mathsf{Par}_{\mathcal{B}_s}(A_i) = \mathbf{a}_j)$ and n_{ijk} by $|\mathcal{D}| \cdot \hat{P}(\mathsf{Par}_{\mathcal{B}_s}(A_i) = \mathbf{a}_j, A_i = a_k)$ in $\Delta(|\mathcal{D}|, \mathcal{B}_s)$. Then, as in previous experiment, we assume extracted frequencies and the conditional entropies are not dependent on the size of data. So we can go over the actual size of data in hand in our experiment.

In this experiment, given a data set $\mathcal{D}$, we randomly select n structures from the set

$$\mathcal{S}_m = \left\{\mathcal{B}_s \in \mathsf{BNS}(\mathcal{D}) \middle| \ |\mathsf{Par}_{\mathcal{B}_s}(A_i)| \leq m \text{ for } 1 \leq i \leq n\right\}$$

and compute the $\Delta(|\mathcal{D}|, \mathcal{B}_s)$ for each randomly selected structure $\mathcal{B}_s$ and for different values of $|\mathcal{D}|$ ranging from a couple of hundreds to $50,000,000$. Then, we take their average which we denote by $\mathsf{Avg}_{\mathcal{D}}(n, |\mathcal{D}|)$. This number represents the approximate divergence of substitution for instances of the data set at hand if the data set size is $|\mathcal{D}|$. Clearly, as the number of trials n gets larger, we have a more precise approximation.

We limit the complexity of the structures we are evaluating by limiting the number of parents of a node. This is necessary because the number of unconstrained structures is super-exponential in the number of nodes which renders any algorithm with no constrains impractical. The upper bound on the number of parents of a node is denoted by m.

Other restrictions can be applied in the random selection phase of this approximation if the algorithm under consideration in which we want to substitute conditional entropy with CH score imposes other types of candidate space-limiting constraints. We have plotted the results of this experiment in Figures 2,3 and 4 for Alarm, Breast Cancer and Neapolitan data sets respectively. The number of trials n is 50 in the first columns of all the three figures and $n = 130$ for the second columns. The space of candidate structures for random selection is extended by adding more complex structures which is achieved by increasing m, the upper bound on the number of parents for a single node.

Experimental results show that the upper bound of divergence of the substitution given an ordering $\mathbf{A}$ of attributes, UP(DIV$(\mathcal{D}, \mathcal{B}_{cs}^{\mathbf{A}})$), is overstated. That is, DIV$(\mathcal{D}, \mathcal{B}_s)$ for an average structure in terms of complexity, $\mathcal{B}_s$ converges to zero much faster than UP(DIV$(\mathcal{D}, \mathcal{B}_{cs}^{\mathbf{A}})$) as the data set $\mathcal{D}$ gets larger. As we increase m, the divergence gets slightly larger. But, the rate of decrease in divergence as data set gets larger is almost constant. The plots are very stable with respect to increasing the number of trials n.

5. Conclusion and Future Works

Our main result shows that for Bayesian structures inferred from large data sets the CH score, that is difficult to compute (but is the standard evaluation tool for Bayesian Networks) can be replaced with the total conditional entropy of the network obtained by summing the conditional entropy of each node conditioned on its parents. We obtained a lower bound

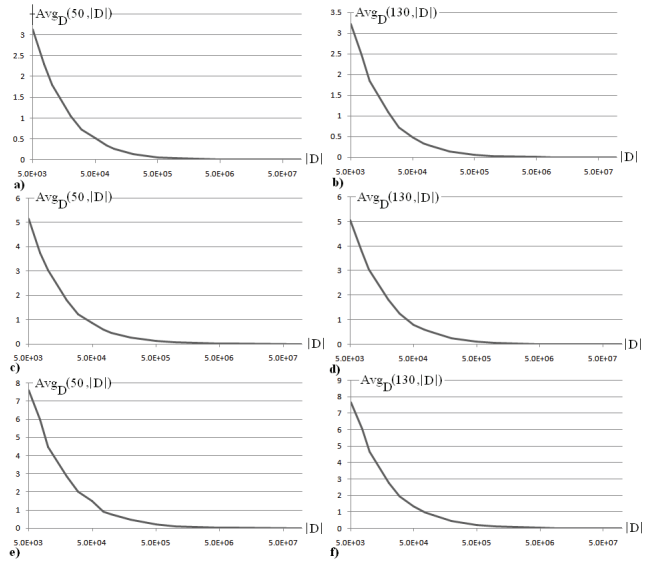

Figure 2: The Plots for Alarm data set. The (n, m) pairs for different plots are as follows a.$(50, 3)$, b.$(130, 3)$, c.$(50, 4)$, d.$(130, 4)$, e.$(50, 5)$, f.$(130, 5)$.

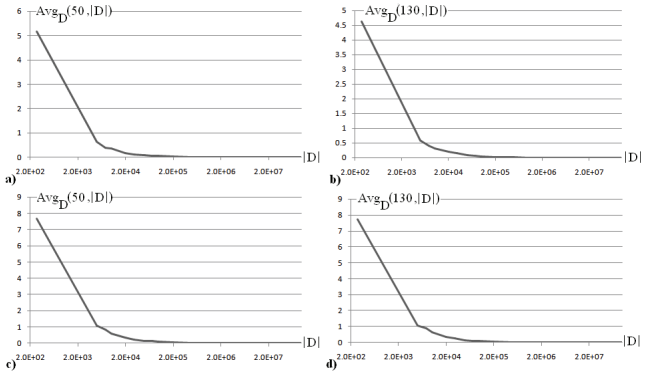

Figure 3: The Plots for Breast Cancer data set. The (n, m) pairs for different plots are as follows a.$(50, 3)$, b.$(130, 3)$, c.$(50, 4)$, d.$(130, 4)$.

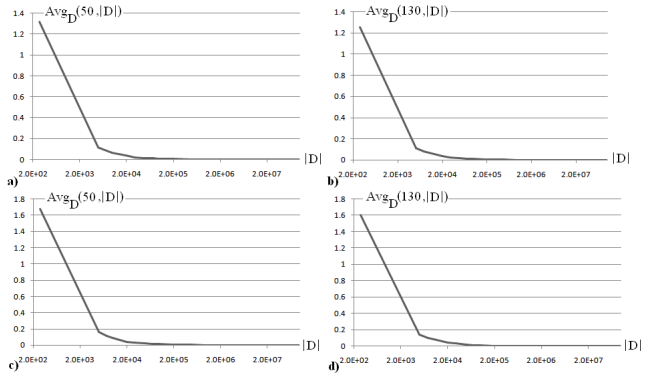

Figure 4: The Plots for Neapolitan data set. The (n, m) pairs for different plots are as follows a.$(50, 2)$, b.$(130, 2)$, c.$(50, 3)$, d.$(130, 3)$.

of the size of data sets that allow a safe replacement of the CH score and provided experimental evidence that this replacement is feasible. We intend to work on improving this lower bound.

References

Baraty, S., and Simovici, D. A. 2009. Edge evaluation in bayesian network structures. In *8th Australian Data Mining Conference - AusDM 2009*.

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

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

Friedman, N., and Goldszmidt, M. 1998. *Learning in Graphical Models*. Cambridge, MA, USA: MIT Press. 421–459.

Heckerman, D.; Geiger, D.; and Chickering, D. M. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, 197–243.

Lam, W., and Bacchus, F. 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 Baraty, S. 2008. Structure inference of Bayesian networks from data: A new approach based on generalized conditional entropy. In *EGC*, 337–342.

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