Representational Upper Bounds of Bayesian Networks

Huajie Zhang HZHANG@CSD.UWO.CA
Charles X. Ling LING@CSD.UWO.CA
Department of Computer Science,The University of Western Ontario, London, Ontario, Canada, N6A 5B7

Abstract

One of the fundamental issues of Bayesian networks is their representational power, reflecting what kind of functions they can or cannot represent. In this paper, we first prove an upper bound on the representational power of Augmented Naive Bayes. We then extend the result to general Bayesian networks. Roughly speaking, if a function contains an m-XOR, there exists no Bayesian networks with node having at most $m-1$ parents to represent it.

1. Introduction

Bayesian networks (BNs) are probabilistic models that combine the probability theory and the graph theory. A Bayesian network consists of a structural model and a set of conditional probabilities. The structural model is a directed acyclic graph in which nodes represent random variables and arcs represent informational or causal dependencies among the variables. The dependencies are quantified by conditional probabilities for each node given its parents in the network.

Bayesian networks have often been used in classification. Assume A_1, $A_2, \cdots$, A_n are n attributes. An example e is represented by a vector $(a_1, a_2, , \cdots, a_n)$, where a_i is the value of A_i. Let C represent the classification variable that corresponds to the class, and c represent the value that C takes.

From the Bayes Rule, the probability of an example $e = (a_1, a_2, \cdots, a_n)$ being class c is

$$p(c|e) = \frac{p(a_1, a_2, \cdots, a_n|c)p(c)}{p(a_1, a_2, \cdots, a_n)}.$$

e is classified into the most probable class c, i.e.,

$$G(e) = \max_c p(c|a_1, a_2, \cdots, a_n), \tag{1}$$

where $G(e)$ is called a Bayesian classifier.

The term $p(c|a_1, a_2, \cdots, a_n)$ is difficult to estimate. Assume all attributes are independent given the value of the class variable; that is,

$$p(a_1, a_2, \cdots, a_n|c) = \prod_{i=1}^{n} p(a_i|c),$$

the resulting $G(e)$ is then:

$$G(e) = \max_c p(c) \prod_{i=1}^{n} p(a_i|c). \tag{2}$$

$G(e)$ is called a Naive Bayesian classifier, or simply Naive Bayes (NB). Figure 1 shows an example of Naive Bayes. In Naive Bayes, each attribute node has the class node as its parent, but does not have any parent from attribute nodes.

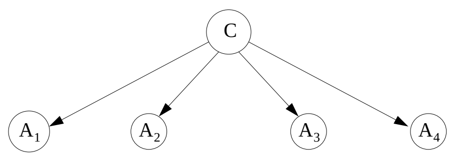

Figure 1. An example of Naive Bayes

Because values of $p(a_i|c)$ can be estimated more accurately from the training examples, Naive Bayes is easy to construct. It is also, however, surprisingly effective (Kononenko, 1990; Langley et al., 1992) in classification. Unfortunately, the independence assumption is rarely true in real-world applications. Indeed, Naive Bayes is found to work poorly for regression problems (Frank et al., 2000), and produces poor probability estimations (Bennett, 2000; Monti & Cooper, 1999). Therefore, researchers have extended the structure of Naive Bayes to represent dependencies among attributes to further improve the results. Tree Augmented Naive Bayes (TAN) is such an extension, in which the class node points directly to all attributes (as in Naive Bayes), but an attribute can have one parent from another attribute (Friedman et al., 1997). TAN is a specific case of general Augmented Naive Bayesian networks (or simply ANB), where the class node also points directly to all attributes, but where there is no limitation on arcs among attributes (except that they do not form any directed cycle).

One of the fundamental issues of ANB or general Bayesian networks is their representational power. The representational power denotes whether there is a Bayesian network that, given a target function, can classify exactly the same as the target function. Representational power is independent of learning algorithms and training data. Therefore, it is a fundamental issue about the capability and limitation of Bayesian networks.

2. Related Work

Bayesian networks have a variety of structures with different structural complexities. Obviously, the more parents a node is allowed to have, the more complex is the structure. It is natural to ask the question, what is the relation between the structural complexity of a Bayesian network and its representational power? Intuitively, the more complex structure a network has, the more representational power it owns.

In the binary domain in which each attribute has only two possible values, the representational power of Naive Bayes is well-known (Duda & Hart, 1973). That is, Naive Bayes represents only linear functions. For ANB, its representational power is also known: an ANB in which each node has at most k parents can represent the polynomial of degree k. This can be derived as follows:

For arbitrary ANB G, assume that each node A_i can have up to $k+1$ parents, since one of them is the class node, A_i can have up to k parents from attribute nodes. Let $pa_i = \{A_{i1}, \cdots, A_{ik}\}$ denote the parents of A_i from attribute nodes. Then

$$
\begin{aligned}
& p(A_i|c, pa_i) = \\
& \qquad \begin{cases}
\theta_1 & A_i = 0, A_{i1} = 0, A_{i2} = 0, \cdots, A_{ik} = 0 \\
1-\theta_1 & A_i = 1, A_{i1} = 0, A_{i2} = 0, \cdots, A_{ik} = 0 \\
\theta_2 & A_i = 0, A_{i1} = 1, A_{i2} = 0, \cdots, A_{ik} = 0 \\
1-\theta_2 & A_i = 1, A_{i1} = 1, A_{i2} = 0, \cdots, A_{ik} = 0 \\
\cdots & \\
\theta_m & A_i = 0, A_{i1} = 1, A_{i2} = 1, \cdots, A_{ik} = 1 \\
1-\theta_m & A_i = 1, A_{i1} = 1, A_{i2} = 1, \cdots, A_{ik} = 1
\end{cases} \\
& = \theta_1^{(1-A_i)(1-A_{i1})\cdots(1-A_{ik})}(1-\theta_1)^{A_i(1-A_{i1})\cdots(1-A_{ik})} \\
& \quad \cdots \theta_m^{(1-A_i)A_iA_{i1}\cdots A_{ik}}(1-\theta_m)^{A_iA_{i1}\cdots A_{ik}}
\end{aligned}
$$

and

$$G(e) = \frac{p(C=+)}{p(C=-)} \prod_{i=1}^{n} \frac{p(A_i|+, pa_i)}{p(A_i|-, pa_i)}. \tag{3}$$

When we apply a logarithm to it and break the product into a sum of logs, we get a set of terms of at most degree $k+1$: $\bar{A}_i\bar{A}_{i1}\cdots\bar{A}_{ik}$, where $\bar{A}_i$ is either A_i or $1-A_i$, $\bar{A}_{ij}$ is similar.

Thus, in binary domains, Naive Bayes represents linear functions, and TAN represents quadratic functions, and so on. When we consider the representational power of ANB in the nominal domain, however, it is much more complex than that in the binary domain. It cannot be extended from the above derivation. Attributes in the nominal domain can have more than two values. Assume A_1, A_2, $\cdots$, A_n are n nominal attributes, each attribute A_i may have m values a_{i1}, a_{i2}, $\cdots$, and a_{im} ($m \geq 2$). Domingos and Pazzani (Domingos & Pazzani, 1997) and Peot (Peot, 1996) introduced m new Boolean attributes B_{i1}, B_{i2}, $\cdots$, and B_{im} for each attribute A_i, and proved that Naive Bayes is linear over those new binary attributes. However, the linear separability on n original attributes is transformed to $m_1 \times m_2 \cdots \times m_n$ new attributes.

Zhang and Ling (2001b) investigated the representational power of Naive Bayes and ANB in the nominal domain. They proposed a measure which uses the maximum XOR contained in a function as its complexity measure. They proved that Naive Bayes cannot represent any nominal function containing a 2-XOR, and TAN cannot represent any nominal function containing a 3-XOR (k-XOR means an XOR function with k variables). They proposed a conjecture extending these results for the general case:

Conjecture 1 *If an function f defined on nominal attributes contains an m-XOR, $m \geq 2$, then no ANB with nodes having at most $m-1$ parents can represent f.*

In this paper, we prove that conjecture first, and then extend it from ANB to general Bayesian networks.

3. The Representational Upper Bound of ANB

3.1 The Complexity Measures for Functions and Bayesian Networks

Essentially, exploring the representational power of Bayesian networks means to establish an association between the structural complexity of a Bayesian network and the complexity of the functions representable to it. It is straightforward to use the maximum number of parents of a node in a Bayesian network as the measure of its structural complexity. Thus, we have the following definition.

Definition 1 *Given a BN G, the maximum number of parent of a node on G is called the structural order, denoted by $O_s(G)$.*

According to the above definition, Naive Bayes has an order of 1, and TAN has an order of 2.

We restrict our discussion to Boolean functions with nominal (discrete) attributes, as defined below.

Definition 2 *Given n nominal attributes A_1, A_2, $\cdots$, A_n, and two class labels $\{+,-\}$, a function f from $A_1 \times A_2 \cdots \times A_n$ to $\{+,-\}$ is called n-dimensional nominal function.*

We adopt the measure for the complexity of nominal functions proposed by Zhang and Ling (2001b). Let us first briefly review the related concepts. A function containing a k-XOR is defined below (Zhang & Ling, 2001b).

Definition 3 *An n-dimensional nominal function f, defined on nominal attributes A_1, A_2, $\cdots$, A_n, is said to contain a k-XOR, if there are k attributes A_{k1}, A_{k2}, $\cdots$, A_{kk}, and for each attribute A_{ki}, there are two different values a_{ki} and $\bar{a}_{ki}$, such that when all attributes other than A_{k1}, A_{k2}, $\cdots$, A_{kk}, have been set to a fixed value, the partial function f_p of f from $\{a_{k1}, \bar{a}_{k1}\} \times \cdots \times \{a_{kk}, \bar{a}_{kk}\}$ to $\{+,-\}$ is a k-XOR function.*

Definition 4 *An n-dimensional nominal function f is said to have an order of m, if the maximum XOR it contains is an m-XOR, denoted by $O_f(f)$.*

Now we already have complexity measures for both functions and Bayesian networks. An ideal relation between them is a one-to-one relation. That is, Bayesian networks of order m can and only can represent functions of order m. In this paper, we establish only a part of this one-to-one relation; i.e., Bayesian networks of order m cannot represent functions of order $m+1$.

3.2 The Proof of the Representational Upper Bound of ANB

We begin to prove the conjecture on the representational power of ANB, which establishes an representational upper bound of ANB. Before we prove it, we prove a lemma first, in which the number of attributes of the function (also the number of attribute nodes of its correspondent ANB) is limited equal to its order.

Lemma 1 *If an m-dimensional nominal function f has an order of m (contains an m-XOR), $m \geq 2$, then no ANB with order at most $m-1$ can represent f.*

Proof: Suppose that A_1, A_2,$\cdots$, A_m are the nominal attributes of f. If there were an ANB G with nodes of A_1, A_2,$\cdots$, A_m having at most $m-1$ parents (including the class node), which could represent f, then according to Definition 4 and 3, there would be two values a_i, $\bar{a}_i$ for each attribute A_i, such that the partial function f_G of the corresponding classifier of G on $\{a_1, \bar{a}_1\} \times \cdots \times \{a_m, \bar{a}_m\}$ is an m-XOR. That is, when the number of A_i taking value a_i is even, the classification is +; else -, where $i = 1, \cdots, m$. Then we have 2^m inequalities below:

$$f_G(A_1, A_2, \cdots, A_m) \begin{cases} \geq 1 & \text{if the number of } a_i \text{ is even} \\ < 1 & \text{if the number of } a_i \text{ is odd} \end{cases} \tag{4}$$

where $i = 1, \cdots, m$ and $f_G(A_1, A_2, \cdots, A_m)$ is specified below.

$$f_G(A_1, A_2, \cdots, A_m) = \frac{p_G(A_1, A_2, \cdots, A_m, +)}{p_G(A_1, A_2, \cdots, A_m, -)}, \tag{5}$$

where
$p_G(A_1, A_2, \cdots, A_m, +)$ and $p_G(A_1, A_2, \cdots, A_m, -)$ are the joint distribution of G in the positive class and the negative class respectively.

Since G is an ANB, we have:

$$p_G(A_1, A_2, \cdots, A_m, +) = p(+) \prod_{i=1}^{m} p(A_i | pa_i, +), \tag{6}$$

$$p_G(A_1, A_2, \cdots, A_m, -) = p(-) \prod_{i=1}^{m} p(A_i | pa_i, -), \tag{7}$$

where pa_i is an assignment of all of A_i's parents from attribute nodes.

Let $A_1 = a_1$. We have 2^{m-2} inequalities below:

$$\frac{p_G(a_1, A_2, \cdots, A_m, +)}{p_G(a_1, A_2, \cdots, A_m, -)} \geq 1, \tag{8}$$

where the number of A_i taking value a_i is odd ($i \neq 1$).

Multiply all these 2^{m-2} inequalities together, we have:

$$\prod^{odd} \frac{p_G(a_1, A_2, \cdots, A_m, +)}{p_G(a_1, A_2, \cdots, A_m, -)} \geq 1. \tag{9}$$

Similar to (8), we have 2^{m-2} inequalities below:

$$\frac{p_G(a_1, A_2, \cdots, A_m, +)}{p_G(a_1, A_2, \cdots, A_m, -)} < 1, \tag{10}$$

where the number of A_i taking value a_i is even ($i \neq 1$). Multiply all these 2^{m-2} inequalities together, we have:

$$\prod^{even} \frac{p_G(a_1, A_2, \cdots, A_m, +)}{p_G(a_1, A_2, \cdots, A_m, -)} < 1. \quad (11)$$

Divide (9) by (11), we have:

$$\prod^{odd} \frac{p_G(a_1, A_2, \cdots, A_m, +)}{p_G(a_1, A_2, \cdots, A_m, -)} \prod^{even} \frac{p_G(a_1, A_2, \cdots, A_m, -)}{p_G(a_1, A_2, \cdots, A_m, +)} > 1 \quad (12)$$

Denote the left side of the above inequality by L_1.

Let $A_1 = \bar{a}_1$, we have the similar inequality below.

$$\prod^{even} \frac{p_G(\bar{a}_1, A_2, \cdots, A_m, +)}{p_G(\bar{a}_1, A_2, \cdots, A_m, -)} \prod^{odd} \frac{p_G(\bar{a}_1, A_2, \cdots, A_m, -)}{p_G(\bar{a}_1, A_2, \cdots, A_m, +)} > 1 \quad (13)$$

Let us denote the left side by L_2. Notice that L_1 and $\frac{1}{L_2}$ are almost the same except a_1 and $\bar{a}_1$. Next we try to prove that $L_1 = \frac{1}{L_2}$ by showing that all items containing a_1 or $\bar{a}_1$ will be cancelled out in both L_1 and L_2.

Let c be $+$ or $-$. Note that $p_G(A_1, \cdots, A_n, c)$ can be decomposed in terms of (6) and (7), so all items in the inequalities are in the form of $p(A_i|pa_i, c)$, and none of them is zero (otherwise, it is not possible for A_1, $\cdots$, A_m to form an m-XOR). For the item $p(A_i|pa_i, c)$ in which both a_1 and $\bar{a}_1$ do not occur, it is obvious that if that item occurs in the numerator of L_1, it should occur in the denominator of L_2. Thus, we only need to consider the items containing a_1 or $\bar{a}_1$. There are two cases.

(1) The items in the form of $p(A_1|pa_1, c)$ occur, where A_1 is a_1 or $\bar{a}_1$. Since A_1 has at most $m-1$ parents and one of them is the class node, A_1 has at most $m-2$ parents from attribute nodes. Thus, there is at least one attribute which is not in pa_1. Let $\bar{pa}_1$ denote all the attributes that are not in pa_1, and t be the number of such attributes ($t \geq 1$). Then there are 2^{t-1} assignments of $\bar{pa}_1$ to make the number of attribute A_i taking a_i ($i \neq 1$) even, and 2^{t-1} assignments odd. Thus each $p(A_1|pa_1, c)$ occurs 2^{t-1} times in both the numerator and denominator of L_1, and then is cancelled out. It is similar in L_2.

(2) The items in the form $p(A_i|pa_i, c)$ occur, where $i \neq 1$ and A_1 is in pa_i. Similar to (1), since there is at least one attribute not in pa_i, each $p(A_i|pa_i, c)$ occurs the same times in the numerator and denominator of L_1 or L_2, and then is cancelled out.

Thus, $L_1 = \frac{1}{L_2}$. It is impossible to satisfy both inequalities (12) and (13). Therefore, we conclude that no such ANB can represent f.
$\square$

Now we are ready to prove the theorem for general ANB with no limitation on the number of attributes.

Theorem 1 *If an n-dimensional nominal function f has an order of m (contains an m-XOR), $m \geq 2$, then no ANB with order at most $m-1$ can represent f.*

Proof: Suppose that A_1, A_2,$\cdots$, A_n are the nominal attributes of f ($n \geq m$), and an m-XOR is defined on $\{A_1, \cdots, A_m\}$. If there were an ANB G with nodes of A_1, A_2,$\cdots$, A_n having at most $m-1$ parents (including the class node), which can represent f, then according to Definition 4 and 3, there are two values a_i, $\bar{a}_i$ for attribute A_i ($1 \leq i \leq m$), and a value a_k for attribute A_k ($m < k \leq n$), such that the partial function f_G of the corresponding classifier of G on $\{a_1, \bar{a}_1\} \times \cdots \times \{a_m, \bar{a}_m\}$ is an m-XOR, while $A_k = a_k$ ($m < k \leq n$). Here we have two types of attributes: unfixed attributes A_1, $\cdots$, A_m that compose an m-XOR, fixed attributes A_{m+1}, $\cdots$, A_n that have values a_{m+1}, $\cdots$, a_n respectively.

Consider $A_1 = a_1$ and $A_1 = \bar{a}_1$, we have the inequalities below similar to (12) and (13):

$$\prod^{odd} \frac{p_G(a_1, A_2, \cdots, A_m, a_{m+1} \cdots, a_n, +)}{p_G(a_1, A_2, \cdots, A_m, a_{m+1}, \cdots, a_n, -)} \times \prod^{even} \frac{p_G(a_1, A_2, \cdots, A_m, a_{m+1}, \cdots, a_n, -)}{p_G(a_1, A_2, \cdots, A_m, a_{m+1}, \cdots, a_n, +)} > 1 \quad (14)$$

$$\prod^{even} \frac{p_G(\bar{a}_1, A_2, \cdots, A_m, a_{m+1}, \cdots, a_n, +)}{p_G(\bar{a}_1, A_2, \cdots, A_m, a_{m+1}, \cdots, a_n, -)} \times \prod^{odd} \frac{p_G(\bar{a}_1, A_2, \cdots, A_m, a_{m+1}, \cdots, a_n, -)}{p_G(\bar{a}_1, A_2, \cdots, A_m, a_{m+1}, \cdots, a_n, +)} > 1. \quad (15)$$

where *odd*/*even* specifies that the number of A_i taking a_i is odd or even, $2 \leq i \leq m$. Let us denote the left sides of the above two inequalities by L_3 and L_4 respectively. Similarly, we try to prove that $L_3 = \frac{1}{L_4}$, and we also only need to consider the items containing a_1 or $\bar{a}_1$. (14) and (15) are different from (12) and (13) only in that there exist fixed attributes.

Let $A_1 = a_1$. There are two cases that a fixed attribute occurs in an item together with a_1 in L_3 or L_4.

(1) A fixed attribute has A_1 and a few other unfixed attributes as its parents. That is, the items in the form

of $p(a_k|pa_k^u, pa_k^f, c)$ occur, where a_k is the value of a fixed attribute A_k ($k > m$), and pa_k^u are the parents of A_k that are unfixed attributes and $A_1 \in pa_k^u$, and pa_k^f are the parents of A_k that are fixed attributes. Since A_k has at most $m-1$ parents and one of them is the class node, A_k has at most $m-2$ parents from attribute nodes. Thus, there are at least two attributes in $\{A_2, \cdots, A_m\}$ not in pa_k. Based on the same reason in proving Lemma 1, these items occur the same times in the numerator and denominator of L_3, and then are cancelled out. It is similar in L_4.

(2) The fixed attributes are the parents of A_1. That is, the items in the form of $p(a_1|pa_1^u, pa_1^f, c)$ occur. Since A_1 has at most $m-1$ parents and one of them is the class node, A_1 has at most $m-2$ parents from attribute nodes. Thus, there is at least one attribute in $\{A_2, \cdots, A_m\}$ not in pa_1. Similarly, those items will be cancelled out in both L_3 and L_4.

For $A_1 = \bar{a}_1$, we have a similar result. Thus, $L_3 = \frac{1}{L_4}$. It is impossible to satisfy both inequalities (14) and (15). Therefore, we conclude that no such ANB can represent f.
□

The above theorem establishes us the upper bound on the representational power of general ANB in the nominal domain. Intuitively, nominal functions containing a higher order XOR present more difficulty to Augmented Naive Bayes, regardless of learning algorithms.

4. Representational Upper Bounds on General Bayesian Networks

In the previous section, we proved that any ANB with nodes having at most m parents cannot represent any functions containing $(m+1)$-XOR. However, ANB is a particular form of general Bayesian networks (BNs). BNs have more general and flexible structures, and it will be interesting if we can extend our results to general Bayesian networks.

One important feature of Bayesian networks is their Markov property. That is, a node in a Bayesian network is affected only by the nodes in its Markov Blanket (Pearl, 1988), defined below.

Definition 5 *The Markov blanket of a node A in BN G is a set of nodes that make up of A's parents and children, and the parents of A's children, denoted by $MB(A)$.*

There are two main differences between general BN and ANB.

(1) From the viewpoint of classification, an ANB represents a classification function, in which the target variable (class variable) is fixed, whereas a general BN represents maybe a set of classification functions, in which each variable may be viewed as a target variable.

(2) In ANB, all attribute nodes are connected to the class node. This means that all attribute nodes are in the Markov blanket of the class node. In a general BN, no such a central node exists, and each node is affected only by the nodes in its Markov blanket.

Again, we use the maximum parent number of a node on a network as its structural complexity measure. If we extend the representational upper bound of ANB to general BN, the correspondent assertion would be: *If an n-dimensional nominal function f has an order of m, $m \geq 2$, then no BN with order at most $m-1$ can represent f.*

In this section, we will prove this assertion.

4.1 General Bayesian Networks as Classifiers

Naive Bayes and ANB represent a classification function with the root as the class variable. What does a general BN represent? One way to view a BN from a classification viewpoint is that each node could be a class variable. Thus, a BN represents a set of classification functions. Formally, we have the following definition.

Definition 6 *Given a BN G on A_1, $\cdots$, A_n and an example $e = (a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_n)$, the classification function f_i corresponding to node A_i is defined as:*

$$f_i(a_1, \cdots, a_n) = \max_{a_i} P_G(a_i|a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_n).$$

The set of classification functions represented by G is $\{f_1, \cdots, f_n\}$, denoted by $R(G)$.

The maximal order of the classification functions represented by a BN is then defined as its order:

Definition 7 *A BN G's representational order $O_r(G)$ is defined as:*

$$O_r(G) = \max\{O_f(f_i), f_i \in R(G)\}$$

As we mentioned above, for a given node A_i, all the nodes not in its Markov blanket $MB(A_i)$ are independent of A_i. Further, from the classification point of view, not all arcs related to the nodes in $MB(A_i)$ affect the classification function f_i corresponding to A_i. More precisely, for each node $A_k \in MB(A_i)$, if A_i is not one of its parents, all the arcs pointing to A_k do not affect f_i, then they can be removed. So we have the following definition.

Definition 8 *Given a BN G and its node A_i, a sub network of G, which makes up of the nodes in $MB(A_i)$ and the arcs that affect the classification function f_i corresponding to A_i, is called A_i's Markov classifier, denoted by $MC(A_i)$.*

Figure 2 shows a BN, and Figure 3 shows the Markov classifier of A_5. From Definition 8, when we discuss the classification functions of a node, we can reduce the original network to the Markov classifier of that node.

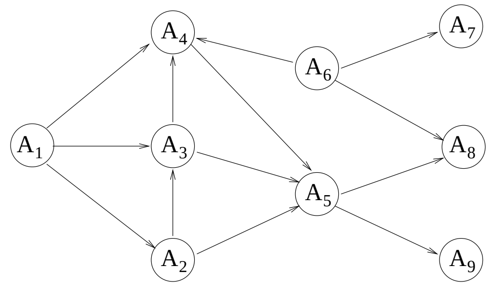

Figure 2. Bayesian network G

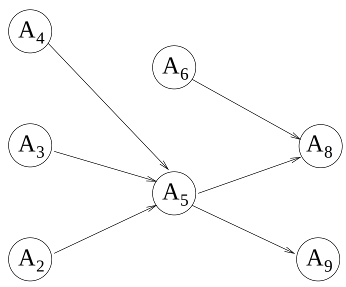

Figure 3. The Markov classifier of node A_5

Since each attribute is nominal in a BN, a Markov classifier for a node may have more than two classes. However, our discussion in Section 3 is restricted to nominal functions with only two classes (Definition 2). Actually our result can be extended to the nominal functions with more than two classes. Correspondingly, we have a more general definition of a nominal function containing a k-XOR, provided below.

Definition 9 *An n-dimensional nominal function f defined on nominal attributes A_1, A_2, $\cdots$, A_n, is said to contain a k-XOR, if there are k attributes A_{k1}, A_{k2}, $\cdots$, A_{kk}, and for each attribute A_{ki}, there are two different values, a_{ki}, $\bar{a}_{ki}$, and there are two different values $+$ and $-$ of class variable C, such that when all attributes other than A_{k1}, A_{k2}, $\cdots$, A_{kk}, have been set to a fixed value, the partial function f_p of f from $\{a_{k1}, \bar{a}_{k1}\} \times \cdots \times \{a_{kk}, \bar{a}_{kk}\}$ to $\{+, -\}$ is a k-XOR function.*

It is easy to get the similar results of Theorem 1 with this definition.

4.2 The Representational Power of 1-order BN

For a BN of order 1, each node has at most one parent. Naive Bayes belongs to this class (but not vice versa). It has been shown that Naive Bayes can represent linear functions and some nonlinear functions, but cannot represent any function containing 2-XOR (Zhang & Ling, 2001a). The following lemma extends the result on Naive Bayes to the BNs with order of 1, thus establishing an upper bound of BNs of order 1.

Lemma 2 *For any BN G, if $O_s(G) = 1$, then $O_r(G) \le 1$.*

Proof: Suppose G is a BN on $A_1, \cdots, A_n$ and $O_s(G) = 1$. if $O_r(G)$ were 2 or above, then there would be a function f of order 2, i.e., f contains a 2-XOR, G can represent it.

Suppose f were from A_1, $\cdots$, A_{k-1}, A_{k+1}, $\cdots$, A_n to A_k. Since f contains a 2-XOR, according to Definition 3, there are two attributes A_i and A_j, i and $j \ne k$, and each of them has two different values, a_i, $\bar{a}_i$, and a_j, $\bar{a}_j$, respectively, and two different values a_k and $\bar{a}_k$ of A_k such that:

$$f(a_1, \cdots, a_i, \cdots, a_j, \cdots, a_n) = a_k \quad (16)$$
$$f(a_1, \cdots, a_i, \cdots, \bar{a}_j, \cdots, a_n) = \bar{a}_k \quad (17)$$
$$f(a_1, \cdots, \bar{a}_i, \cdots, a_j, \cdots, a_n) = \bar{a}_k \quad (18)$$
$$f(a_1, \cdots, \bar{a}_i, \cdots, \bar{a}_j, \cdots, a_n) = a_k \quad (19)$$

where a_l is a value of A_l, $l \ne i$, j and k.

If one of A_i and A_j is out of the Markov blanket of A_k, it does not affect the value of A_k while other attributes are assigned values. Therefore, it is not possible that Equation (16), (17), (18) and (19) hold true simultaneously.

Assume that both A_i and A_j are in the Markov blanket of A_k. Since each node has at most one parent, A_i and

A_j are connected directly with A_k. There are two cases for the connection patterns.

(1) A_k points to both A_i and A_j. In this case, it is a structure of Naive Bayes with A_k as the class variable. It is known that Naive Bayes cannot represent 2-XOR (Zhang & Ling, 2001b).

(2) A_i points to A_k and A_k points to A_j. In this case, the joint distribution P_G can be represented as below.

$$\begin{aligned} P_G(A_i, A_j, A_k) &= P(A_k)P(A_i|A_k)P(Aj|A_i, A_k) \\ &= P(A_k)P(A_i|A_k)P(Aj|A_k) \end{aligned}$$

This means that P_G can be represented by a Naive Bayes with A_k as the class variable. Therefore, if G could represent 2-XOR, the correspondent Naive Bayes could too.

Therefore, G cannot represent any function of order 2. □

4.3 The Representational Power of General BN

Let us consider the representational power of general BNs. We will prove that a BN with nodes having at most m parents cannot represent any function of $m+1$ order. The basic idea to prove it is that we convert the Markov classifier of a node in a BN to an equivalent ANB with the same order, then apply Theorem 1. Thus, we need to prove that the Markov classifier of any node in a BN can be represented by an ANB with the same order. This is proved by the following lemma.

Lemma 3 *Given a BN G with $O_s(G) = m$, and its node A, there is an ANB G_A with A as the class node and $O_s(G_A) \leq m$, that represents A's classification function f.*

Proof: According to Definition 5, there are three types of nodes in $MB(A)$: A's parents, A's sons and the parents of A's sons. Let us consider them one by one.

Suppose A_1, $\cdots$, A_k are the all parents of A, and $k \leq m$. According to Definition 8, all arcs among A_1, $\cdots$, A_k should not occur in $MC(A)$, as shown in Figure 3. The local distribution $p(A|A_1, \cdots, A_k)$ can be represented as below.

$$p(A|A_1, \cdots, A_k) = \frac{p(A)p(A_1|A)p(A_2|A, A_1) \cdots p(A_k|A, A_1, \cdots, A_{k-1})}{p(A_1, \cdots, A_k)}$$

Since $p(A_1, \cdots, A_k)$ is independent of A, we can drop it and consider only the numerator. Thus, $p(A|A_1, \cdots, A_k)$ can be represented by a structure in which A points to each A_i, $1 \leq i \leq k$, and A_i has A, A_1, $\cdots$, and A_{i-1} be its parents. Notice that the maximum parents number of a node in the structure is equal to k. We use the arcs on this structure to replace the arcs from A_1, $\cdots$, A_k to A, shown in Figure 4.

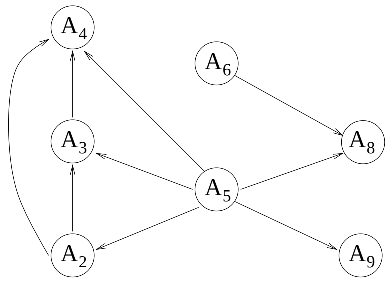

Figure 4. An equivalent structure with ANB-like structure

In the second case, the nodes are A's sons, which is already an ANB-like structure.

In the third case, the nodes are the parents of A's sons. Assume that A_j is such a node. If A is not one of A_j's parents, the local distribution $p(A_j|pa_j)$ is independent of f, then all arcs pointing to A_j do not occur on $MC(A)$. Thus we can add an arc from A to A_j, and for each possible value of A_j, set

$$p(A_j = a_j|A = a_1) = \cdots = p(A_j = a_j|A = a_n),$$

where a_i, $i = 1, \cdots, n$, are all possible values of A. Obviously, the resulting network still represent f.

Now we has a resulting network G_A with ANB structure and $O_s(G_A) \leq m$, that is, node A points to each other nodes and the maximum parent number of a node is at most equal to m. For example, Figure 3 is converted to an equivalent ANB (A_5 is the class node), shown in Figure 5.

From the discussion above, G_A also represents f.

Therefore, any classification function of a BN can be represented by an ANB with the same structural order. □

Now we now ready to prove the main theorem about the representational upper bound of BNs.

Theorem 2 *For any BN G, $O_r(G) \leq O_s(G)$.*

Proof: Suppose G is a BN on nominal attributes A_1, $\cdots$, A_n and $O_s(G) = m - 1$, i.e., each node has at

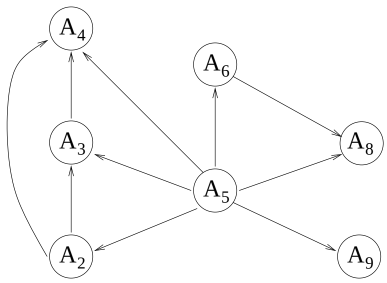

Figure 5. The resulting ANB equivalent to Figure 3

most $m-1$ parents. If there were a function f of order m, i.e., f contains an m-XOR, G could represent it. Assume the m-XOR contained in f is from A_1, $\cdots$, A_m to A_{m+1}.

We construct the Markov classifier $MC(A_{m+1})$ from G. It is obvious that $O_s(MC(A_{m+1})) \leq m-1$. According to Lemma 3, there is an equivalent ANB G_A of structural order at most $m-1$ to represent the m-XOR. However, Theorem 1 shows that such an ANB does not exist.

Therefore, G could not represent a function containing m-XOR.
$\square$

Theorem 2 presents a representational upper bound for general BNs, thus establishing an explicit association between the structural complexity of a BN and its representational capacity.

5. Conclusion

In this paper, we proved a representational upper bound for ANB, that is, an ANB with nodes having at most m parents cannot represents any function containing $m+1$-XOR. Then we extended this result to general Bayesian networks, and proved a similar result. That is, any BNs of order m cannot represent a target function of order $m+1$.

We give only a representational upper bound of BNs in this paper. A natural question is: what is the representational lower bound of BNs? An interesting and intuitive conjecture is that any function of order m can be represented by some BN of order m. This conjecture is correct for Naive Bayes and TAN (Zhang & Ling, 2001b), but the general case has not been proved, and it is one of our future research work.

References

Bennett, P. N. (2000). Assessing the calibration of Naive Bayes' posterior estimates. In *Technical report no. cmu-cs00-155*.

Domingos, P., & Pazzani, M. (1997). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning*, *29*, 103–130.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. A Wiley-Interscience Publication.

Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Naive Bayes for regression. *Machine Learning*, *41(1)*, 5–15.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163.

Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In B. Wielinga (Ed.), *Current trends in knowledge acquisition*. IOS Press.

Langley, P., Iba, W., & Thomas, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the tenth national conference of artificial intelligence*, 223–228. AAAI Press.

Monti, S., & Cooper, G. F. (1999). A Bayesian network classifier that combines a finite mixture model and a Naive Bayes model. In *Proceedings of the 15th conference on uncertainty in artificial intelligence*, 447–456. Morgan Kaufmann.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems:networks of plausible inference*. Morgan Kauhmann.

Peot, M. A. (1996). Geometric implications of the Naive Bayes assumption. In *Proceedings of the 12th conference on uncertainty in artificial intelligence*. Morgan Kaufmann.

Zhang, H., & Ling, C. X. (2001a). Geometric properties of Naive Bayes in nominal domains. In L. D. Raedt and P. Flach (Eds.), *Proceedings of 12th european conference on machine learning*, 588–599. Springer.

Zhang, H., & Ling, C. X. (2001b). Learnability of augmented Naive Bayes in nominal domains. In C. E. Brodley and A. P. Danyluk (Eds.), *Proceedings of the eighteenth international conference on machine learning*, 617–623. Morgan Kaufmann.