International Journal of Approximate Reasoning 50 (2009) 1036–1045

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

VC dimension and inner product space induced by Bayesian networks ☆

Youlong Yang *, Yan Wu

School of Science, Xidian University, Xi'an 710071, PR China

ARTICLE INFO

Article history:
Received 16 August 2008
Received in revised form 2 April 2009
Accepted 3 April 2009
Available online 17 April 2009

Keywords:
Bayesian networks
VC dimension
Inner product space
Concept classes

ABSTRACT

Bayesian networks are graphical tools used to represent a high-dimensional probability distribution. They are used frequently in machine learning and many applications such as medical science. This paper studies whether the concept classes induced by a Bayesian network can be embedded into a low-dimensional inner product space. We focus on two-label classification tasks over the Boolean domain. For full Bayesian networks and almost full Bayesian networks with n variables, we show that VC dimension and the minimum dimension of the inner product space induced by them are $2^n - 1$. Also, for each Bayesian network $\mathcal{N}$ we show that $VCdim(\mathcal{N}) = Edim(\mathcal{N}) = 2^{n-1} + 2^i$ if the network $\mathcal{N}'$ constructed from $\mathcal{N}$ by removing X_n satisfies either (i) $\mathcal{N}'$ is a full Bayesian network with $n-1$ variables, i is the number of parents of X_n, and $i < n-1$ or (ii) $\mathcal{N}'$ is an almost full Bayesian network, the set of all parents of X_n $PA_n = \{X_1, X_2, X_{n3}, \ldots, X_{ni}\}$ and $2 \leqslant i < n-1$. Our results in the paper are useful in evaluating the VC dimension and the minimum dimension of the inner product space of concept classes induced by other Bayesian networks.

1. Introduction

The Bayesian network (BN) is a graphical structure for representing probabilistic relationships among a large number of variables. Over the last decade, it has become a popular tool for modeling various kinds of statistical problems [1–4]. Bayesian networks are particularly useful for dealing with high dimensional statistical problems. They reduce the complexity of the phenomenon being studied by representing joint relationships among its variables as conditional relationships among subsets of the variables. The parameters of a Bayesian network specify the conditional distribution of each variable given its parents in the graph, and the joint distribution for the variables in the domain is defined by the product of these conditional distributions. For more information see, for example, [5,6].

In recent years, we have seen a tremendous growth of interests in learning based on kernel methods. Kernel-based analysis is based on relations among observed data and offers a new viewpoint. Inner products of vectors are central to kernel-based methods. Quite often in kernel-based learning the inner product operation is not carried out explicitly, but reduced to the evaluation of the so-called kernel function that operates on instances of the original data space. A major advantage of this technique is that it enables efficient storage of high-dimensional feature spaces. Various kernels were suggested and extensively studied, for instance, [7–11].

We have also seen a growth of interests in the use of BNs for classification [3,12]. Some methods that combine BNs with kernel methods or probabilistic methods have been proposed [7,13–15]. Altun et al. [16] proposed a kernel for the Hidden Markov Model, which is a special case of a Bayesian network. Nakamura et al. [17] established upper and lower bounds on the dimension of the inner product space for Bayesian networks.

☆ This work was supported by National Natural Science Foundation of China (60574075).
* Corresponding author.
E-mail addresses: ylyang@mail.xidian.edu.cn (Y. Yang), yanwu@mail.xidian.edu.cn (Y. Wu).

0888-613X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2009.04.001

Let $\mathscr{N}$ be a Bayesian network. Let $VCdim(\mathscr{N})$ denote the Vapnik–Chervonenkis (VC) dimension of the concept class $\mathscr{C}_{\mathscr{N}}$ by $\mathscr{N}$, and let $Edim(\mathscr{N})$ denote the smallest d such that $\mathscr{C}_{\mathscr{N}}$ can be embedded into the d-dimensional Euclidean space $\mathbb{R}^d$. In this paper we study of the inner product space and the concept class that are induced by Bayesian networks. For each full Bayesian network and for each almost full Bayesian, we show that the VC dimension as well as the minimum dimension of the inner product space of the concept class induced by them are $2^n - 1$, where n is the number of variables of the network. Also, for each Bayesian network $\mathscr{N}$ we show that $VCdim(\mathscr{N}) = Edim(\mathscr{N}) = 2^{n-1} + 2^i$ if there is a variable X_n such that the network $\mathscr{N}'$ constructed form $\mathscr{N}$ by removing X_n satisfies either (i) $\mathscr{N}'$ is a full Bayesian network with $n-1$ variables, i is the number of parents of X_n, and $i < n-1$ or (ii) $\mathscr{N}'$ is an almost full Bayesian network, the set of all parents of X_n $PA_n = \{X_1, X_2, X_{n3}, \ldots, X_{ni}\}$ and $2 \leqslant i < n-1$.

The paper is organized as follows. In Section 2 we discuss previous relevant work and describe our notations. In Section 3, we prove our results regarding the VC dimension and the minimum dimension of the inner product space induced by a full Bayesian network. In Section 4, we obtain the results regarding almost full Bayesian networks. Finally, in Section 5, we make a conclusion and a discussion about three problems on the concept classes induced by other Bayesian networks.

2. The basic concepts and terms

We begin by defining the notations related to Bayesian networks, concept classes and VC dimension.

2.1. Bayesian networks

A Bayesian network $\mathscr{N}$ on a set of variables $V = \{X_1, X_2, \ldots, X_n\}$ represents a joint probability distribution among those variables. The network consists of two components: (1) a network structure $\mathscr{G}$ and (2) a set P of conditional probabilities. $\mathscr{G}$ encodes conditional dependencies among the variables. It is an n-node acyclic directed graph (DAG for short) in which the nodes one-to-one correspond with the variables. We thus naturally identify each node with the variable it represents. For all i and j, $1 \leqslant i,\ j \leqslant n$, there exists an arc in $\mathscr{G}$ from X_i to X_j if and only if X_j is conditionally dependent on X_i. The set P describes the conditional dependencies among the set of variables that are directly connected with arcs in the structure $\mathscr{G}$. If a node X_i has arcs coming in from the set of nodes $X_{i1}, \ldots, X_{ik}$, then P contains the probability distribution of X_i conditioned on the variables $X_{i1}, \ldots, X_{ik}$.

Since the Bayesian network structure $\mathscr{G}$ is acyclic, the nodes of $\mathscr{G}$ can be *topologically sorted* in such a way that if an arc exists from a node X to another node Y, then X must precede Y in the ordering. The acyclicity of $\mathscr{G}$ also induces ancestral relations. If there is an arc from X to Y, then Y is a *child* of X and X is *parent* of Y. For each i, PA_i denotes the set of all parents of X_i, $m_i = \mid PA_i \mid$ be the number of parents and CH_i denotes the set of all children of X_i. The structure $\mathscr{G}$ is *fully connected* if for all $i \geqslant 2$, $m_i = i - 1$. In other words, a fully connected structure $\mathscr{G}$ has an arc between every pair of nodes. Since $\mathscr{G}$ is acyclic, a fully connected $\mathscr{G}$ has a unique topological ordering and in that ordering, the ith node has exactly $i-1$ parents (that is, all the nodes that precede it in the ordering). Fig. 1 shows two fully connected Bayesian networks $\mathscr{N}_{3F}$ and $\mathscr{N}_{4F}$ with 3 and 4 variables, respectively. We assume that each variable of our Bayesian networks is boolean. Thus the number of possible assignments to the variables in an n-variable network is 2^n. The class of distributions induced by $\mathscr{N}$, denoted as $\mathscr{P}_{\mathscr{N}}$, is a set of the distribution on $\{0,1\}^n$. $P \in \mathscr{P}_{\mathscr{N}}$ is given as follows: For each $X = (X_1, \ldots, X_N) \in \{0,1\}^n$,

$$P(X) = \prod_{i=1}^{n} p(X_i | PA_i),$$

where $P(X)$ is the probability of X and $p(X_i|PA_i)$ represents the conditional probability of X_i given the assignments to PA_i.

For each distribution $P \in \mathscr{P}_{\mathscr{N}}$, the total number of parameters, that is, the number of independent variables that express $P(X)$ is $\sum_{i=1}^{n} 2^{m_i}$. Thus we can use real values set

$$\bigcup_{i=1}^{n} \{p_{i,\alpha} = p(X_i = 1 | PA_i = \alpha) | \alpha \in \{0,1\}^{m_i}\}$$

to denote the distribution P, that is,

$$P = \bigcup_{i=1}^{n} \bigcup_{\alpha \in \{0,1\}^{m_i}} \{p_{i,\alpha}\}$$

and $P = \{p_{i,\alpha}\}$ for short.

Fig. 1. Two fully connected Bayesian networks with 3 and 4 variables, $\mathscr{N}_{3F}$ and $\mathscr{N}_{4F}$.