Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume IV
Pro–Z

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Pseudo-Independent Models and Decision Theoretic Knowledge Discovery

Yang Xiang
University of Guelph, Canada

INTRODUCTION

Graphical models such as Bayesian networks (BNs) (Pearl, 1988; Jensen & Nielsen, 2007) and decomposable Markov networks (DMNs) (Xiang, Wong., & Cercone, 1997) have been widely applied to probabilistic reasoning in intelligent systems. Knowledge representation using such models for a simple problem domain is illustrated in Figure 1: Virus can damage computer files and so can a power glitch. Power glitch also causes a VCR to reset. Links and lack of them convey dependency and independency relations among these variables and the strength of each link is quantified by a probability distribution. The networks are useful for inferring whether the computer has virus after checking files and VCR. This chapter considers how to discover them from data.

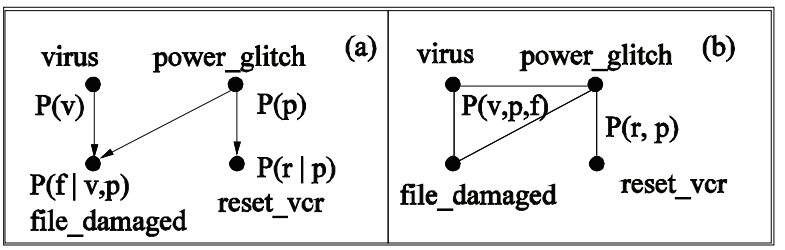
Discovery of graphical models (Neapolitan, 2004) by testing all alternatives is intractable. Hence, heuristic search are commonly applied (Cooper & Herskovits, 1992; Spirtes, Glymour, & Scheines, 1993; Lam & Bacchus, 1994; Heckerman, Geiger, & Chickering, 1995; Friedman, Geiger, & Goldszmidt, 1997; Xiang, Wong, & Cercone, 1997). All heuristics make simplifying assumptions about the unknown data-generating models. These assumptions preclude certain models to gain efficiency. Often assumptions and models they exclude are not explicitly stated. Users of such heuristics may suffer from such exclusion without even knowing. This chapter examines assumptions underlying common heuristics and their consequences to graphical model discovery. A decision theoretic strategy for choosing heuristics is introduced that can take into account a full range of consequences (including efficiency in discovery, efficiency in inference using the discovered model, and cost of inference with an incorrectly discovered model) and resolve the above issue.

BACKGROUND

A graphical model encodes probabilistic knowledge about a problem domain concisely (Pearl, 1988; Jensen & Nielsen, 2007). Figure 1 illustrates a BN in (a) and a DMN in (b). Each node corresponds to a binary variable. The graph encodes dependence assumptions among these variables, e.g., that *f* is directly dependent on *v* and *p*, but is independent of *r* once the value of *p* is observed. Each node in the BN is assigned a conditional probability distribution (CPD) conditioned on its parent nodes, e.g., $P(f \mid v, p)$ to quantify the uncertain dependency. The joint probability distribution (JPD) for the BN is uniquely defined by the product $P(v, p, f, r) = P(f \mid v, p)\, P(r \mid p)\, P(v)\, P(p)$. The DMN has two groups of nodes that are maximally pairwise connected, called *cliques*. Each is assigned a probability distribution, e.g., $\{v, p, f\}$ is assigned $P(v, p, f)$. The JPD for the DMN is $P(v, p, f)\, P(r, p) / P(p)$.

When discovering such models from data, it is important that the dependence and independence relations

Figure 1. (a) An example BN (b) A corresponding DMN



expressed by the graph approximate true relations of the unknown data-generating model. How accurately can a heuristics do so depends on its underlying assumptions.

To analyze assumptions underlying common heuristics, we introduce key concepts for describing dependence relations among domain variables in this section. Let V be a set of discrete variables $\{x_1, \ldots, x_n\}$. Each x_i has a finite space $S_{xi} = \{x_{i,j} | 1 \le j \le D_i\}$. When there is no confusion, we write $x_{i,j}$ as x_{ij}. The space of a set $X \subseteq V$ of variables is the Cartesian product $S_X = \prod_{x_i \in X} S_i$. Each element in S_X is a configuration of X, denoted by $x = (x_1, \ldots, x_n)$. A probability distribution $P(X)$ specifies the probability $P(x) = P(x_1, \ldots, x_n)$ for each x. $P(V)$ is the *JPD* and $P(X)$ ($X \subset V$) is a *marginal distribution.* A probabilistic domain model (PDM) over V defines $P(X)$ for every $X \subseteq V$.

For disjoint subsets W, U and Z of V, W and U are *conditionally independent* given Z, if $P(w \mid u, z) = P(w \mid z)$ for all configurations such that $P(u, z) > 0$. The condition is also denoted $P(W \mid U, Z) = P(W \mid Z)$. It allows modeling of dependency within $W \bigcup U \bigcup Z$ through *overlapping* subsets $W \bigcup Z$ and $U \bigcup Z$.

W and U are *marginally independent* if $P(W \mid U) = P(W)$ holds whenever $P(U) > 0$. The condition allows dependency within $W \bigcup U$ to be modeled over *disjoint* subsets. If each variable x_i in a subset X is marginally independent of $X \backslash \{x_i\}$, then variables in X are *marginally independent*.

Variables in a subset X are *generally dependent* if $P(Y \mid X \setminus Y) \neq P(Y)$ for every $Y \subset X$. For instance, $X = \{x_1, x_2, x_3\}$ is *not* generally dependent if $P(x_1, x_2 | x_3) = P(x_1, x_2)$. It is generally dependent if $P(x_1, x_2 | x_3) \neq P(x_1, x_2)$, $P(x_2, x_3 | x_1) \neq P(x_2, x_3)$ and $P(x_3, x_1 | x_2) \neq P(x_3, x_1)$. Dependency within X cannot be modeled over disjoint subsets but may through overlapping subsets, due to conditional independence in X.

Variables in X are *collectively dependent* if, for each proper subset $Y \subset X$, there exists no proper subset $Z \subset X \setminus Y$ that satisfies $P(Y \mid X \setminus Y) = P(Y \mid Z)$. Collective dependence prevents modeling through overlapping subsets and is illustrated in the next section.

MAIN THRUST OF THE CHAPTER

Pseudo-Independent (PI) Models

A *pseudo-independent* (PI) model is a PDM where proper subsets of a set of collectively dependent variables display marginal independence (Xiang, Wong., & Cercone, 1997). Common heuristics often fail in learning a PI model (Xiang, Wong., & Cercone, 1996). Before analyzing how assumptions underlying common heuristics cause such failure, we introduce PI models below. PI models can be classified into three types: *full, partial, and embedded*. The basic PI model is a full PI model.

Definition 1. A PDM over a set V ($|V| \ge 3$) of variables is a *full PI model* if the following hold:

(S_I) Variables in each proper subset of V are marginally independent.

(S_{II}) Variables in V are collectively dependent.

Example 1 Patient of a chronicle disease changes the health state (denoted by variable s) daily between stable ($s = t$) and unstable ($s = u$). Patient suffers badly in an unstable day unless treated in the morning, at which time no indicator of the state is detectable. However, if treated at the onset of a stable day, the day is spoiled due to side effect. From historical data, patient's states in four consecutive days observe the estimated distribution in Table 1.

The state in each day is uniformly distributed, i.e., $P(s_i = t) = 0.5$ where $1 \le i \le 4$. The state of each day is marginally independent of that of the previous day, i.e., $P(s_i = t | s_{i-1}) = 0.5$ where $2 \le i \le 4$. It is marginally independent of that of the previous two days, i.e., $P(s_i = t | s_{i-1}, s_{i-2}) = 0.5$ where $3 \le i \le 4$. However, states of four days are collectively dependent, e.g., $P(s_4 = u | s_3 = u, s_2 = t, s_1 = t) = 1$, which allows the state of the last day to be predicted from states of previous three days. Hence, the patient's states form a full PI model.

By relaxing condition *(S_I)*, full PI models are generalized into partial PI models defined through *marginally independent partition* (Xiang, Hu, Cercone, & Hamilton, 2000):

Table 1. Estimated distribution of patient health state

(s_1,s_2,s_3,s_4)	P(.)	(s_1,s_2,s_3,s_4)	P(.)
(t, t, t, t)	1/8	(u, t, t, t)	0
(t, t, t, u)	0	(u, t, t, u)	1/8
(t, t, u, t)	0	(u, t, u, t)	1/8
(t, t, u, u)	1/8	(u, t, u, u)	0
(t, u, t, t)	0	(u, u, t, t)	1/8
(t, u, t, u)	1/8	(u, u, t, u)	0
(t, u, u, t)	1/8	(u, u, u, t)	0
(t, u, u, u)	0	(u, u, u, u)	1/8

Definition 2. Let V ($|V| \geq 3$) be a set of variables, and $B = \{ B^1, ..., B^m \}$ ($m \geq 2$) be a partition of V. B is a *marginally independent partition* if, for every subset $X = \{x_i^k \mid x_i^k \in B^k, k = 1,...,m\}$, variables in X are marginally independent. Each B^i is a *marginally independent block*.

A marginally independent partition groups variables into m blocks. If a subset X is formed by taking one element from each block, then variables in X are marginally independent. Partial PI models are defined by replacing marginally independent subsets with the marginally independent partition.

Definition 3. A PDM over a set V ($|V| \geq 3$) of variables is a *partial PI model* if the following hold:

(S_I') V can be partitioned into marginally independent blocks.

(S_{II}) Variables in V are collectively dependent.

Table 2 shows the JPD of a partial PI model over $V = \{x_1, x_2, x_3\}$ where x_1 and x_2 are ternary. The marginal probabilities are

$P(x_{1.0}) = 0.3$, $P(x_{1.1}) = 0.2$, $P(x_{1.2}) = 0.5$,
$P(x_{2.0}) = 0.3$, $P(x_{2.1}) = 0.4$, $P(x_{2.2}) = 0.3$,
$P(x_{3.0}) = 0.4$, $P(x_{3.1}) = 0.6$.

The marginally independent partition is $\{\{x_1\}, \{x_2, x_3\}\}$. Variable x_1 is marginally independent of each variable in the other block, e.g., $P(x_1, x_{2.0}) = P(x_{1.1})\,P(x_{2.0}) = 0.06$. However, variables in block $\{x_2, x_3\}$ are dependent, e.g., $P(x_{2.0}, x_{3.1}) = 0.1 \neq P(x_{2.0})\,P(x_{3.1}) = 0.18$. The three variables are collectively dependent, e.g., $P(x_{1.1} \mid x_{2.0}, x_{3.1}) = 0.1$ and $P(x_{1.1} \mid x_{2.0}) = P(x_{1.1} \mid x_{3.1}) = 0.2$.

A partial PI model may involve only a proper subset of V and remaining variables show normal dependency. The subset is an *embedded* PI submodel. A PDM can embed multiple submodels.

Definition 4. Let a PDM be over a set V of generally dependent variables. A proper subset $V' \subset V$ ($|V'| \geq 3$) of variables forms an *embedded* PI submodel if the following hold:

(S_{III}) V' forms a partial PI model.

(S_{IV}) The marginal independent partition $B = \{ B^1, ..., B^m \}$ of V' extends into V. That is, V partitions into $\{X^1,...,X^m\}$ such that $B^i \subseteq X^i, (i = 1,...,m)$ and, for each $x \in X_i$ and $y \in X_j$ $(i \neq j)$, x and y are marginally independent.

Table 3 shows the JPD of a PDM with an embedded PI submodel over x_1, x_2 and x_3. The marginal probabilities are $P(x_{1.0}) = 0.3$, $P(x_{2.0}) = 0.6$, $P(x_{3.0}) = 0.3$, $P(x_{4.0}) = 0.34$, $P(x_{5.0}) = 0.59$.

The marginally independent partition of the submodel is $\{B^1 = \{x_1\}, B^2 = \{x_2, x_3\}\}$.

Table 2. A partial PI model where $v = (x_1, x_2, x_3)$

$\underline{v}$	P(.)	$\underline{v}$	P(.)	$\underline{v}$	P(.)	$\underline{v}$	P(.)	$\underline{v}$	P(.)	$\underline{v}$	P(.)
(0,0,0)	0.05	(0,1,1)	0.11	(1,0,0)	0.05	(1,1,1)	0.08	(2,0,0)	0.10	(2,1,1)	0.11
(0,0,1)	0.04	(0,2,0)	0.06	(1,0,1)	0.01	(1,2,0)	0.03	(2,0,1)	0.05	(2,2,0)	0.01
(0,1,0)	0.01	(0,2,1)	0.03	(1,1,0)	0	(1,2,1)	0.03	(2,1,0)	0.09	(2,2,1)	0.14

Table 3. A PDM with an embedded PI submodel where $v = \{x_1, x_2, x_3, x_4, x_5\}$

v	$P(.)$	v	$P(.)$	v	$P(.)$	v	$P(.)$
(0,0,0,0,0)	0	(0,1,0,0,0)	.0018	(1,0,0,0,0)	.0080	(1,1,0,0,0)	.0004
(0,0,0,0,1)	0	(0,1,0,0,1)	.0162	(1,0,0,0,1)	.0720	(1,1,0,0,1)	.0036
(0,0,0,1,0)	0	(0,1,0,1,0)	.0072	(1,0,0,1,0)	.0120	(1,1,0,1,0)	.0006
(0,0,0,1,1)	0	(0,1,0,1,1)	.0648	(1,0,0,1,1)	.1080	(1,1,0,1,1)	.0054
(0,0,1,0,0)	.0288	(0,1,1,0,0)	.0048	(1,0,1,0,0)	.0704	(1,1,1,0,0)	.0864
(0,0,1,0,1)	.0072	(0,1,1,0,1)	.0012	(1,0,1,0,1)	.0176	(1,1,1,0,1)	.0216
(0,0,1,1,0)	.1152	(0,1,1,1,0)	.0192	(1,0,1,1,0)	.1056	(1,1,1,1,0)	.1296
(0,0,1,1,0)	.0288	(0,1,1,1,1)	.0048	(1,0,1,1,1)	.0264	(1,1,1,1,1)	.0324

Outside the submodel, B^1 extends to include x_4 and B^2 extends to include x_5. Each variable in one block is marginally independent of each variable in the other block, e.g.,

$$P(x_{1,1}, x_{5,0}) = P(x_{1,1})\, P(x_{5,0}) = 0.413.$$

Variables in the same block are pairwise dependent, e.g.,

$$P(x_{2,1}, x_{3,0}) = 0.1 \neq P(x_{2,1})\, P(x_{3,0}) = 0.12.$$

Variables in the submodel are collectively dependent, e.g.,

$$P(x_{1,1} \mid x_{2,0}, x_{3,1}) = 0.55,\ P(x_{1,1} \mid x_{2,0}) = P(x_{1,1} \mid x_{3,1}) = 0.7.$$

However, x_5 is independent of other variables given x_3, displaying conditional independence, e.g.,

$$P(x_{5,1} \mid x_{2,0}, x_{3,0}, x_{4,0}) = P(x_{5,1} \mid x_{3,0}) = 0.9.$$

PDMs with embedded PI submodels are the most general PI models.

Heuristics for Model Discovery

Given a data set over n variables, the number of possible network structures is super-exponential (Cooper & Herskovits, 1992). To make discovery tractable, a number of heuristics are commonly applied. The most common is the *Naive Bayes* heuristic (Zhang, 2004). It restricts potential models to Naive Bayes models whose graph consists of a single root (the *hypothesis*) and its observable child nodes (the *attributes*). Since the hypothesis is given, discovery focuses on finding the CPD at each node and is very efficient.

Another heuristic is the *TAN* heuristic, that restricts potential models to *tree augmented Naive Bayes* models (Friedman, Geiger, & Goldszmidt, 1997). Its graph also has a single root (the hypothesis). However, attributes themselves form a tree (see Figure 2). Each attribute has the hypothesis and at most one other attribute as its parent nodes.

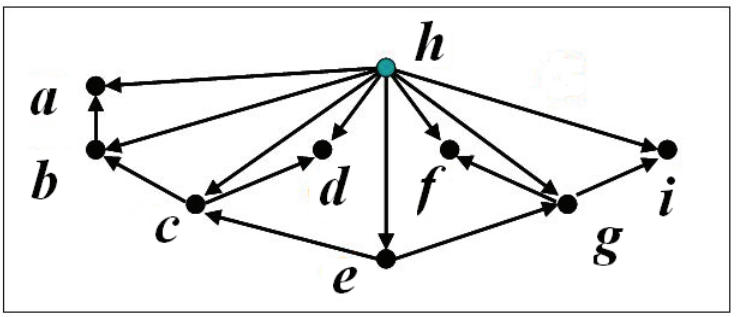
The above heuristics limit the model space. Heuristics below limit the search procedure. One common heuristic is the *single-link lookahead* (Cooper & Herskovits, 1992; Heckerman, Geiger & Chickering, 1995; Lam & Bacchus, 1994). Learning starts with an initial graph. Successive graphical structures, representing different sets of independence assumptions, are adopted. Each adopted structure differs from its predecessor by a single link and improves a score metric optimally.

An alternative is the *bounded multi-link lookahead* (Hu & Xiang, 1997) where an adopted structure differs from its predecessor by up to $k > 1$ links. The algorithm applies single-link lookahead and low-order (small k) multi-link lookahead as much as possible, and uses high-order (large k) multi-link lookahead only when necessary.

Underlying Assumptions and their Implications

Knowledge discovery starts with a dataset generated by an unknown PDM M. The goal is to uncover a graphical model that approximates M. The best outcome is often known as the *minimal I-map* (Pearl, 1988). It is a graph G whose nodes correspond to variables in

Figure 2. Graph structure of a TAN model where h is the hypothesis



M and whose links are as fewer as possible such that graphical separation among nodes in *G* implies conditional independence in *M*. The assumption underlying a heuristic determines its ability to discover minimal I-maps for various PDMs. The following are assumptions underlying Naïve Bayes and TAN.

Proposition 1. In a Naïve Bayes model, every two attributes are conditionally independent given the hypothesis.

Proposition 2. In a TAN model, every two non-adjacent attributes are conditionally independent given the parent of one of them and the hypothesis.

The general assumption underlying the single-link lookahead is unclear. Known results are based on particular algorithms using the heuristic and are centered around *faithfulness*. A PDM *M* is *faithful* if there exists some graph *G* such that conditional independence among variables in *M* implies graphical separation among corresponding nodes in *G*, and vice versa. Spirtes, Glymour and Scheines (1993) present a sufficient condition: If *M* is faithful, the algorithm in question can discover a minimal I-map of *M.* Xiang, Wong and Cercone (1996) present a necessary condition: If *M* is unfaithful, the output of the algorithm in question will not be an I-map. Hence, faithfulness will be regarded as the primary assumption underlying the single-link lookahead heuristic.

The bounded multi-link lookahead heuristic does not make any of the above assumptions and is the most general among heuristics mentioned above. Implications of these assumptions to discovery of PI models are summarized below (Xiang, 2007).

Theorem 1. Let Λ be the set of all Naive Bayes models and Λ' be the set of all PI models over *V*. Then $\Lambda \cap \Lambda' = \varnothing$.

Theorem 2. Let Λ be the set of all TAN models over *V*. Let Λ' be the set of all PI models over *V* such that each PI model in Λ' contains at least one embedded PI submodel over 4 or more variables. Then $\Lambda \cap \Lambda' = \varnothing$.

Theorem 3. A PI model is unfaithful.

Theorems 1 and 3 say that Naive Bayes and single-link lookahead heuristics cannot discover a minimal I-map if the unknown PDM is PI. Theorem 2 says that if the unknown PDM is beyond the simplest PI model (with exactly 3 variables), then the TAN heuristic cannot discover a minimal I-map.

Suppose that these three heuristics (coupled with known algorithms) are applied to the data in Example 1 in order to find the best strategy for patient treatment. They will be mislead by the marginal independence and return an empty graph (four nodes without links). This is equivalent to say that there is no way that the patient can be helped (either untreated and possibly suffering from the disease, or treated and possibly suffering from the side effect).

On the other hand, if a bounded 6-link lookahead heuristic is applied, the correct minimal I-map will be returned. This is due to the ability of multi-link lookahead to identify collective dependence. Although in this small example, the minimal I-map is a complete graph, the bounded 6-link lookahead can still discover a minimal I-map when the PI model is embedded in a much large PDM. From this discovered model, a targeted treatment strategy can be developed by predict-

ing the patient's state from states of the last three days. Discovery of a PI model from social survey data and experimental result on its performance can be found in (Xiang, Hu, Cercone, & Hamilton, 2000).

FUTURE TRENDS

Decision Theoretic Strategy

Heuristics such as Naïve Bayes, TAN and single-link lookahead are attractive to data mining practitioners due to mostly two reasons: First, they are more efficient. Second, PDMs that violate their underlying assumptions are less likely. For instance, unfaithful PDMs are considered much less likely than faithful ones (Spirtes, Glymour and Scheines, 1993). Although efficiency in discovery and prior probability of potential model are important factors, an additional factor, the cost of suboptimal decision (such as that according to the discovered empty graph for Example 1) has not been paid sufficient attention. A decision theoretic strategy (Xiang, 2007) that integrates all these factors is outlined below, where faithfulness is used as an example assumption.

Let A and A' be alternative discovery algorithms, where A assumes faithfulness and A' does not. Costs of discovery computation are $C_{disc}(A) = d$ and $C_{disc}(A') = d'$, where $d < d'$. The unknown PDM M has a small probability ε to be unfaithful. Choosing A, if M is faithful, the discovered model supports optimal actions. If M is unfaithful, the discovered model causes suboptimal actions. Choosing A', no matter M is faithful or not, the discovered model supports optimal actions. Let the action cost of a correct model (a minimal I-map) be $C_{opt} = 0$ and that of an incorrect model be $C_{sub} = \omega > 0$. The expected cost of choosing A is

$$C_{disc}(A) + (1-\varepsilon)\, C_{opt} + \varepsilon\, C_{sub} = d + \varepsilon\, \omega$$

and that of choosing A' is $C_{disc}(A') + C_{opt} = d'$. According to decision theory, A' is a better choice if and only if

$$\omega > (d' - d)/\varepsilon .$$

In other words, for mission critical applications, where the above inequation often holds, the less efficient but more open-minded algorithm A' should be preferred.

CONCLUSION

Heuristics must be used to render discovery of graphical models computationally tractable. They gain efficiency through underlying assumptions. Naive Bayes makes the strongest assumption, followed by TAN, followed by single-link lookahead, followed by bounded multi-link lookahead, while their complexities are reversely ordered. These assumptions are not subject to verification in the discovery process. The stronger the assumption made, the more likely that the discovered model is not the minimal I-map and, as a result, the model does not support the optimal decision. A decision-theoretic strategy chooses heuristic based on discovery efficiency, likelihood of discovering an incorrect model, as well as consequence in applying an incorrectly discovered model in decision making. For mission critical applications, a more open-minded heuristic should be preferred even though the computational cost of discovery may be higher.

REFERENCES

Cooper, G.F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309-347.

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

Heckerman, D., Geiger, D., & Chickering, D.M. (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning, 20*, 197-243.

Hu, J., & Xiang, Y. (1997). Learning belief networks in domains with recursively embedded pseudo independent submodels, In *Proc. 13th Conf. on Uncertainty in Artificial Intelligence*, (pp. 258-265).

Jensen, F.V., & Nielsen, T.D. (2007). *Bayesian networks and decision graphs* (2nd Ed.). Springer.

Lam, W., & Bacchus, F. (1994). Learning Bayesian networks: an approach based on the MDL principle. Computational Intelligence, 10(3):269--293.

Neapolitan, R.E. (2004). *Learning Bayesian networks*. Prentice Hall.

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Springer-Verlag.

Wong, S.K.M., & Xiang, Y. (1994). Construction of a Markov network from data for probabilistic inference. *Proc. 3rd Inter. Workshop on Rough Sets and Soft Computing*, 562-569.

Xiang, Y. (2007). A Decision theoretic view on choosing heuristics for discovery of graphical models. In *Proc. 20th Inter. Florida Artificial Intelligence Research Society Conf*., (pp. 170-175).

Xiang, Y., Hu, J., Cercone, N., & Hamilton, H. (2000). Learning pseudo-independent models: Analytical and experimental results. In H. Hamilton, (Ed.), *Advances in Artificial Intelligence*, (pp. 227-239).

Xiang, Y., Lee, J., & Cercone, N. (2003). Parameterization of pseudo-independent models. In *Proc. 16th Inter. Florida Artificial Intelligence Research Society Conf.*, (pp. 521-525).

Xiang, Y., Wong, S.K.M., & Cercone, N. (1996). Critical remarks on single link search in learning belief networks. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, (pp. 564-571).

Xiang, Y., Wong, S.K.M., & Cercone. N. (1997). A `microscopic' study of minimum entropy search in learning decomposable Markov networks. *Machine Learning, 26*(1), 65-92.

Zhang, H. (2004). The optimality of naive Bayes. In *Proc. of 17th International FLAIRS conference (FLAIRS 2004)* (pp. 562-567).

KEY TERMS

Bounded Multi-Link Lookahead Heuristic: It differs from the single-link lookahead in that each adopted graph is selected from candidates that differ from its successor by up to $k>1$ links. It requires higher but bounded computational cost, makes the weakest assumption, and can discover PDMs that are not discoverable by the single-link lookahead such as PI models.

Embedded PI Submodel: An embedded PI submodel is a full or partial PI model over a proper subset of domain variables. The most general PI models are those that embed PI submodels in large problem domains.

Full PI Model: A full PI model is a PDM where every proper subset of variables is marginally independent but the entire set is collectively dependent. They are the most basic PI models.

Naïve Bayes Heuristic: It assumes that the model graph consists of a single root (the *hypothesis*) and its observable child nodes (the *attributes*). It makes the strongest independence assumption and is the most efficient.

Partial PI Model: A partial PI model is otherwise the same as a full PI model except that some subsets of variables may not be marginally independent. A full PI model is also a partial PI model. Hence, partial PI models are more general.

Single-Link Lookahead Heuristic: The discovery process using this heuristic consists of a sequence of adopted graphs such that each is selected from candidates that differ from its successor by exactly one link. Models discoverable with this heuristic are usually faithful PDMs.

TAN Heuristic: It assumes the same as Naïve Bayes plus that each attribute may have at most one other attribute as the additional parent.