Expert Systems with Applications 38 (2011) 7123–7134

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Constructing the Bayesian network structure from dependencies implied in multiple relational schemas ☆

Wei-Yi Liu, Kun Yue *, Wei-Hua Li

Department of Computer Science and Engineering, School of Information Science and Engineering, Yunnan University, Kunming 650091, PR China

ARTICLE INFO

Keywords:
Relational data model
Bayesian network
Acyclic database schema
Harmoniousness multi-valued dependency set
Join dependency

ABSTRACT

Relational models are the most common representation of structured data, and acyclic database theory is important in relational databases. In this paper, we propose the method for constructing the Bayesian network structure from dependencies implied in multiple relational schemas. Based on the acyclic database theory and its relationships with probabilistic networks, we are to construct the Bayesian network structure starting from implied independence information instead of mining database instances. We first give the method to find the maximum harmoniousness subset for the multi-valued dependencies on an acyclic schema, and thus the most information of conditional independencies can be retained. Further, aiming at multi-relational environments, we discuss the properties of join graphs of multiple 3NF database schemas, and thus the dependencies between separate relational schemas can be obtained. In addition, on the given cyclic join dependency, the transformation from cyclic to acyclic database schemas is proposed by virtue of finding a minimal acyclic augmentation. An applied example shows that our proposed methods are feasible.

1. Introduction

Relational models are the most common representation of structured data. Enterprise business data, medical records, and scientific data sets are most stored in relational databases. Knowledge representation and discovery on relational databases are important for real applications and widely studied in these years.

It is well known that Bayesian networks, as the graphical representation of probabilistic relationship between variables, are effective and widely used frameworks (Pearl, 1988). A Bayesian network is an acyclic directed graph for representing a joint probabilistic distribution, in which the nodes represent variables and edges signify causal relationships between directly-linked variables. With associated conditional probabilities, the quantitative causal relationships among given variables can be captured. The influences between variables are expressed by probabilistic dependencies. Actually, probabilistic dependencies, especially conditional independencies, are used to simplify the representation of a joint distribution. Based on the Bayesian network, probabilistic reasoning can be made effectively.

It has been concluded that constructing the structure is critical and challenging when a Bayesian network is to be learned. Several methods (Buntine, 1996; Cheng, Greiner, Kelly, Bell, & Liu, 2002; Cooper & Herskovits, 1992; Heckerman, 1995) have been proposed for constructing Bayesian networks from data. On the other hand, many researchers have noticed the similarities between probabilistic concepts and relational terminologies including conditional independence versus embedded multi-valued dependency, Markov network versus acyclic join dependency (Liao, Wang, Li, & Liu, 2006; Liu & Song, 2003; Wong, 1997; Wong & Butz, 2001; Wong, Butz, & Wu, 2000). Their research work provides some methods for constructing a probabilistic network from a new perspective. In this paper, we will right develop our methods from the above point of view.

Every instance r_i on a relation schema R_i conforms to data dependencies over R_i, so it is not always necessary to mine these independency information from instances, which will reduce the computation cost consequently. There have been some methods proposed pertinent to the above problem. Based on the multi-valued dependencies implied in a relational schema, Wong and Butz (2001) proposed the method for constructing a probabilistic network. Nevertheless in real-world applications, the 3rd Normal Form (3NF) or the Boyce–Codd Normal Form (BCNF) relational schemas are adopted frequently and they are obtained based on functional dependencies instead of multi-valued dependencies. Accordingly, Liao et al. (2006) gave the method for constructing the Bayesian network structure based on functional dependencies

☆ This work was supported by the National Natural Science Foundation of China (Nos. 61063009, 60933001), the Ph.D. Programs Foundation of Ministry of Education of China (No. 2010531120001), and the Natural Science Foundation of Yunnan Province (No. 2008CD083).

* Corresponding author. Tel.: +86 871 5033146; fax: +86 871 5031598.
E-mail address: kyue@ynu.edu.cn (K. Yue).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.12.053

implied in a relational schema. Both of these two methods are aiming at the case that there is only one relational schema in the database.

Actually, a relational database typically consists of several tables (i.e., relational schemas) and not just one table. Thus it is desired to construct the probabilistic network with regard to the database including multiple relational schemas. For such multi-relational situations, it is natural to first join the given schemas, and then construct the network structure based on the methods given in Liao et al. (2006), Wong and Butz (2001). However, the relationships between attributes in separate relational schemas have not been discussed in Liao et al. (2006), Wong and Butz (2001), while such relationships may critically determine the implied conditional independencies and then the ultimate network structure. Therefore, we will have to construct the Bayesian network structure followed by the mining of implied relationships between separate relational schemas. In addition, on the prerequisite of acyclic conditions, the methods in Liao et al. (2006), Wong and Butz (2001) have not concerned the case that the given relational schema is cyclic.

Exactly motivated by constructing the effective Bayesian network structure in multi-relational environments, in this paper, based on implied dependencies in relational schemas, we develop an approach for constructing the Bayesian network structure and meanwhile discuss relevant properties, regardless of acyclic or cyclic cases.

Fortunately, we can establish our discussion based on the existing results and conclusions that are well accepted and proved to be effective. The theory of acyclic database schemas is important in relational databases, and it is leaved alongside the theory of the relational database technology. In this paper, we apply the theory to constructing the Bayesian network structure. This application is both providing a theoretical basis for learning probabilistic models and finding a new application area for the traditional theory.

Generally speaking, a typical database schema $R = \{R_1, \ldots, R_n\}$ is dependency-preserving, lossless-join decomposition into 3NF. For each R_i, the set F_i of functional dependencies over R_i is knowable. From F_i, we can obtain a set M_i of multi-valued dependencies implied by F_i, written as $F_i \models M_i$. Sciore (1981) presents the concept of the conflict-free set of multi-valued dependencies, which is equivalent to an acyclic join dependency. Based on conflict-free set, Wong and Butz (2001) gave an algorithm to obtain an acyclic dependency structure.

For conflicting set M of multi-valued dependencies (MVDs), Wong and Butz (2001) pointed out "if conflicting generalized multi-valued dependencies (GMVDs) are detected, we have to rely on the domain experts to resolve these conflicts. Henceforth, we may assume that a conflict-free full minimal cover has been determined from the GMVDs supplied by the individual domain experts." We note that the concept of conflict-free set is too harsh in term of equivalence. In fact, a subset of multi-valued dependencies, which is equivalent to an acyclic join dependency, may not be a conflict-free set. In this paper, we give the concept of the harmoniousness set of multi-valued dependencies, which is equivalent to an acyclic join dependency. And we then give an algorithm to obtain a maximum harmoniousness subset M_i^* preserving information of M_i. Furthermore, a Markov network (MN) can be viewed as an acyclic join dependency (Wong, 1997). In this paper, the process to find a probabilistic model on R_i is summed up: $F_i \models M_i \supseteq M_i^* \equiv AJD = MN$. Based on the obtained MN, edge orientations can be done in line with a certain sequence of functional dependencies, so the BN structure can be constructed. Above interpretation shows the underlying theory and main idea of our proposed methods throughout this paper.

Generally, the main contributions of this paper can be summarized as follows:

- On an acyclic relational schema, we give the method for finding a best-preserved harmoniousness subset M^* of the multi-valued dependency set M. Thus, the most information of conditional independencies in M can be retained.
- For the condition of several tables, we develop the properties of join graphs of multiple 3NF database schemas and discuss the relationships between nonprime attributes in separate relational schemas. For acyclic conditions, we point out there are no any functional dependencies between the nonprime attributes. It is easy to obtain the ultimate network structure for multiple relational schemas, since the Markov network corresponding to an acyclic join dependency is decomposable.
- We give an algorithm to transform a cyclic schema into an acyclic schema by virtue of finding a minimal acyclic augmentation on the given cyclic join dependency. This algorithm guarantees the probabilistic relations will not be lost.
- We implement our methods and apply them into a real-world example. Accordingly, we make the comparison between the result obtained by our methods and that learned from data. The applied example shows that our methods are feasible.

The remainder of this paper is organized as follows: Section 2 introduces related work. Section 3 presents the preliminaries of this paper. Section 4 discusses the probabilistic model implied by a relational schema, and presents our method to obtain the maximum harmoniousness subset. Section 5 presents the properties of join graphs of multiple 3NF relational schemas. Section 6 gives the transformation of a cyclic schema to an acyclic schema. Then, Section 7 shows the applied example. At last, Section 8 concludes and points out the future work.

2. Related work

A variety of algorithms have been proposed to induct probabilistic networks from data. These algorithms generally fall into two categories: scoring-and-search-based and dependency-analysis-based ones (Heckerman, Mamdani, & Wellman, 1995). The scoring-and-search-based algorithms use a scoring metric to evaluate a candidate network, and try to search for network structure that best fits the data (Cooper & Herskovits, 1992; Lan & Bacchus, 1994; Wong & Leung, 2004; Wong, Lam, & Leung, 1999). The network learning problem can be viewed as a search problem, which is NP-hard (Chickering, Meek, & Heckerman, 2003). The dependency-analysis-based algorithms try to construct a Bayesian network using dependency information obtained from the data (Cheng et al., 2002). Generally speaking, it is difficult to know if two nodes are conditionally independent. In the worst case, all possible combinations of the conditioning set need to be examined, which would require an exponential number of conditional independence (CI) tests. When a high-order conditional independence relation is tested in a small data set, the test result may be unreliable (Wong & Leung, 2004; Wong et al., 1999).

Many researchers have noticed the similarities between probabilistic concepts and relational terminologies. Dechter (1990) stated that conditional independence parallels the notion of embedded multi-valued dependency. Wong (1997) pointed out a Markov network can be viewed as an acyclic join dependency. The relational data model and probabilistic dependencies are studied in Butz, Wong, and Yao (1999), Liao et al. (2006), Liu and Song (2003), Wong (1997), Wong and Butz (2001), Wong, Butz, and Xiang (1998), Wong et al. (2000), Yao, Butz, and Hamilton (2005).

Join dependencies can be divided into two classes: cyclic and acyclic join dependencies. It is shown that the latter class has a number of desirable properties, Beeri, Fagin, Maier, and Yannakakis (1983) proved the equivalence of the properties and characteriza-