Information Sciences 179 (2009) 197–209

Contents lists available at ScienceDirect

Information Sciences

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

Degrees of conditional (in)dependence: A framework for approximate Bayesian networks and examples related to the rough set-based feature selection

Dominik Ślęzak

Infobright Inc., ul. Krzywickiego 34 pok. 219, 02-078 Warszawa, Polska, Poland

ARTICLE INFO

Article history:
Received 24 April 2008
Received in revised form 25 August 2008
Accepted 14 September 2008

Keywords:
Approximate independence
Mutual information
Bayesian networks
Multi-valued dependencies
Feature selection
Rough sets
Data discretization

ABSTRACT

Bayesian networks provide the means for representing probabilistic conditional independence. Conditional independence is widely considered also beyond the theory of probability, with linkages to, e.g. the database multi-valued dependencies, and at a higher abstraction level of semi-graphoid models. The rough set framework for data analysis is related to the topics of conditional independence via the notion of a decision reduct, to be considered within a wider domain of the feature selection. Given probabilistic version of decision reducts equivalent to the data-based Markov boundaries, the studies were also conducted for other criteria of the rough-set-based feature selection, e.g. those corresponding to the multi-valued dependencies. In this paper, we investigate the *degrees of approximate conditional dependence*, which could be a topic corresponding to the well-known notions such as conditional mutual information and polymatroid functions, however, with many practically useful approximate conditional independence models unmanageable within the information theoretic framework. The major paper's contribution lays in extending the means for understanding the degrees of approximate conditional dependence, with appropriately generalized semi-graphoid properties formulated and with the mathematical soundness of the Bayesian network-like representation of the approximate conditional independence statements thoroughly proved. As an additional contribution, we provide a case study of the approximate conditional independence model, which would not be manageable without the above-mentioned extensions.

1. Introduction

Conditional independence (CI) provides us, in its original and still most widely researched probabilistic version, with the means for expressing relationships among random variables [35]. In machine learning, e.g. the variables are interpreted as attributes (columns, features) in data tables and the joint probabilistic distributions are estimated by looking at combinations of the attributes' values observed as the data records [27]. In such cases, probabilistic CI is often renamed as statistical CI [15]. Many researchers find probabilistic CI useful while interpreting the tasks of the feature selection [24,25], which may be regarded as aiming at searching for attributes that provide (almost) the same information about some specified targets as the set of all available attributes. For example, given widely studied relationships between rough sets and probability [32,64,65,69], the concepts related to probabilistic CI were considered within the framework of probability-based attribute reduction [44,48].

E-mail address: slezak@infobright.com
URL: http://www.infobright.com

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.09.007

There are many ways of representing probabilistic CI. The most popular one refers to Bayesian networks (BNs) – directed acyclic graphs (DAGs) enriched with conditional probability distributions labeling their nodes [17,35]. BNs can be constructed basing on the expert knowledge or on the probabilistic CI statements discovered during the data-based learning processes [5,19]. Each BN is supposed to be an independence mapping (IM), which graphically encodes knowledge about the probabilistic CI statements by means of so called *d*-separation. The corresponding DAG structure serves as a probabilistic CI knowledge base. This way, BNs can be used also as the means for knowledge visualization, which is important for interaction between the experts and the decision support systems. For instance, researchers in bioinformatics often represent knowledge derivable from the gene expression data [2,26] using DAGs spanned over the gene-related attributes [14,30]. BNs and their extensions are widely applicable also to such areas as the new case classification, databases, information retrieval and data compression, where their ability to represent (in)dependencies among the sets of attributes plays the key role [1,8,20,52].

Having in mind real-life problems concerning various types of data and related reasoning strategies, one can ask whether probabilistic model of CI is the only one. It is also important in the above-mentioned domain of feature selection, where information about (in)dependencies between attributes does not need to be expressible in terms of probabilities. In [36,60], it is stated that the reason for BNs to be able to encode knowledge about probabilistic CI lays in its so called semi-graphoid properties. Such properties hold also for some other interpretations of CI. Consequently, BNs can be reconsidered for non-probabilistic approaches too. For instance, one of the known alternative CI models is based only on a part of probabilistic information, namely, whether the value-vectors have zero or non-zero probability. Such occurrence-based CI model was given as an example in [35] and analyzed in [45] as corresponding to the rough-set-based attribute reduction framework relying on so called generalized decisions [33,43]. On the other hand, it corresponds to the database-related framework for the embedded multi-valued dependencies (EMVDs) [12,39,42,63]. Hence, there is a direct linkage between the rough set approach to the feature selection and the principles of modeling dependencies in databases.

Some investigation has been also conducted to extend the existing approaches towards approximate CI, that would better fit the real-life data. It was repeatedly noted in the literature that requirement for the precise equality between probability distributions while defining probabilistic CI is impractical or even self-contradictory, given that the probability theory is supposed to deal with imprecision [16,41,44,67]. The notion of approximate CI may have also an impact on knowledge discovery, where the most interesting patterns or dependencies usually turn out to hold in data only to some reasonably high degree. For example, within the rough set framework for feature selection [18,58], the following three principles of the *approximate attribute reduction* were considered [44]: (1) it is worth reducing irrelevant attributes and simplifying the corresponding decision system; (2) reduction (simplification) should not decrease the overall system's ability to approximate the target concepts; (3) in real-world situations, however, we should agree to slightly decrease the system's quality, if it leads to significantly simpler underlying dependencies. In other words, given previously-mentioned correspondence between CI and the feature selection, one may consider the decision systems based on the CI statements, which are simpler but only approximately satisfied in data.

Analogous ideas, referrable to the Occam's razor and the minimum-description-length principles [37,38], were considered in other areas related to CI. As an example, most of algorithms extracting BNs from data focus only on those out of edges that provide significant amount of information about inter-variable correlations [5,19]. The resulting DAGs may represent probabilistic CI statements that are only roughly true, which is often the only solution because of no exact probabilistic CI statements in the real-life data. However, before our later-discussed publications [47,49], there was no theoretical background for analyzing whether, and to what *degree*, the probabilistic CI statements represented by DAGs pre-learnt from the data in such an inexact fashion are actually valid against the same data. In other words, although there were some previous attempts to formalize the notion of consistency of DAGs with respect to the data [6], there was no analogous attempt to find correspondence between data-related consistency of DAGs and data-related degrees of satisfaction of the CI statements derivable from those DAGs using *d*-separation.

Approximate CI has a natural counterpart in the information theory [10,23]. Given the probability distribution equalities $p(xyz)p(z) = p(xz)p(yz)$ equivalent to $H(XYZ) + H(Z) = H(XZ) + H(YZ)$,[1] for the information entropy $H : \mathscr{P}(A) \to [0, +\infty)$,[2] one can regard *conditional mutual information* $I(X;Y|Z) = H(XZ) + H(YZ) - H(XYZ) - H(Z)$ as the degree of approximate conditional dependence (CD-degree) of X and Y subject to Z. H can be actually replaced by an arbitrary polymatroid function $F : \mathscr{P}(A) \to [0, +\infty)$ [57,66]. Different polymatroid functions let us look differently at the concept of approximate CI. On the other hand, one expects all the polymatroid-based interpretations of approximate CI to have similar properties.

In [47,49], we investigated the BN-like networks graphically representing the *F*-based approximate CI statements, i.e. the statements of independence of X and Y subject to Z, where the corresponding *F*-based CD-degree $F(X;Y|Z) = F(XZ) + F(YZ) - F(XYZ) - F(Z)$ does not exceed a presumed level. We showed in our previous research that for an arbitrary polymatroid $F : \mathscr{P}(A) \to [0, +\infty)$ one can reason about such approximate *F*-based CI statements using *d*-separation, whenever the underlying DAG satisfies the relevant bounds for the *F*-based consistency with data. The major contribution of this paper (Theorem 3) extends our previous framework (Theorem 2) onto the approximate CI models, which are not definable using polymatroids. Given A as the set of attributes of interest, it is indeed not said that the only way to express the CD-degrees

[1] We write xy and XY instead of x, y and $X \cup Y$, respectively; x, y, z denote the value-vectors over $X, Y, Z \subseteq A$; A is the set of all the attributes/variables of interest.

[2] By $\mathscr{P}(A)$ we denote the family of all subsets of A.