Theoretical Computer Science 603 (2015) 111–131

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Frontiers for propositional reasoning about fragments of probabilistic conditional independence and hierarchical database decompositions ☆

Sebastian Link

Department of Computer Science, The University of Auckland, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 31 January 2013
Accepted 10 August 2014
Available online 14 July 2015

Keywords:
Algorithm
Axiomatization
Complexity
Conditional independence
Discrete probability measure
Hierarchical database decomposition
Implication
Propositional logic

ABSTRACT

Conditional independence provides an essential framework to deal with knowledge and uncertainty in Artificial Intelligence, and is fundamental in probability and multivariate statistics. Its associated implication problem is paramount for building Bayesian networks. Unfortunately, the problem does not enjoy an axiomatization, by a finite set of Horn rules, and is already coNP-complete to decide for some fragments of conditional independencies. Saturated conditional independencies form an important fragment whose implication problem is decidable in almost linear time. We establish an axiomatization, by a finite set of Horn rules, for the fragment of generalized saturated conditional independencies. These state the conditional independence between finitely many sets of random variables. The special case of two sets captures Geiger and Pearl's finite axiomatization for saturated conditional independencies. Even for this special case, our completeness proof is new. The proof argument utilizes special probability models where two events have probability one half. Special probability models allow us to establish an equivalence between the implication of generalized saturated conditional independencies and that of formulae in a Boolean propositional fragment. This duality is then extended to a trinity including the implication problem of Delobel's full first-order hierarchical database decompositions. Already for the independence between two sets of random variables, the dualities cannot be extended to cover conditional independencies in general, or first-order hierarchical decompositions.

1. Introduction

The concept of conditional independence is important for capturing structural aspects of probability distributions, for dealing with knowledge and uncertainty in Artificial Intelligence, and for learning and reasoning in intelligent systems [43]. Application areas include natural language processing, speech processing, computer vision, robotics, computational biology, and error-control coding [20]. A conditional independence (CI) statement $I(Y, Z \mid X)$ represents the independence of two sets of random variables relative to a third: given three mutually disjoint subsets X, Y, and Z of a set S of random variables, if we have knowledge about the state of X, then knowledge about the state of Y does not provide additional evidence for the state of Z and vice versa. An important problem is the implication problem, which is to decide for an arbitrary finite set

☆ Preliminary results have been announced at 19th International Workshop on Logic, Language, Information and Computation (WoLLIC 2012).

E-mail address: s.link@auckland.ac.nz.

http://dx.doi.org/10.1016/j.tcs.2015.07.017

S, and an arbitrary set $\Sigma \cup \{\varphi\}$ of CI statements over S, whether every probability model that satisfies every CI statement in Σ also satisfies φ. The significance of this problem is due to its relevance for building Bayesian networks [43]. Indeed, if an important CI statement φ is not implied by Σ, then adding φ to Σ results in new opportunities to construct complex probability models with polynomially many parameters and to efficiently organize distributed probability computations [16]. The implication problem for CI statements is not axiomatizable by a finite set of Horn rules [49]. However, it is possible to express CI statements using polynomial likelihood formulae, and reasoning about polynomial inequalities is axiomatizable [20,25]. Recently, the implication problem of stable CI statements has been shown to be finitely axiomatizable and *coNP*-complete to decide [41,42]. Here, stability means that the validity of $I(Y, Z \mid X)$ over S implies the validity of every $I(Y, Z \mid X')$ where $X \subseteq X' \subseteq S - YZ$. An important subclass of CI statements are saturated conditional independence (SCI) statements. These are CI statements $I(Y, Z \mid X)$ over S that satisfy $XYZ = S$, that is, the set union XYZ of X, Y, and Z is S. Geiger and Pearl have established an axiomatization for the implication problem of SCI statements by a finite set of Horn rules [17].

Example 1. Let $\textsc{Blu-ray} = \{m(ovie), a(ctor), r(ole), c(rew), f(eature), l(anguage), s(ubtitle)\}$ denote a set of random variables that model information about Blu-rays of movies, including which actors played which role, what crew members were involved, which features are available in which language, and what subtitles can be accessed. Let Σ consist of $I(s, arc, fl \mid m)$ and $I(s, ar, cfl \mid m)$, expressing that independent of each other are the set of subtitles, the set of actors, roles, and crew, and the set of features and languages, given a movie, and that independent are the set of subtitles, the set of actors and roles, and the set of crew, features, and languages, given a movie, respectively. Then Σ implies $\varphi_1 = I(s, c, ar, fl \mid m)$ expressing the independence between the set of subtitles, the set of a crew, the set of actors and roles, and the set of features and languages, given a movie. However, Σ does not imply $\varphi_2 = I(s, a, crfl \mid m)$ which states the independence of the set of subtitles, the set of actors, and the set of crew, role, feature, and language, given a movie. Indeed, we can define two events that match on every variable in $\{m, c, f, l, s\}$, differ on a and on r; and assign the probability one half to both events. This probability model satisfies both elements of Σ but violates φ_2. $\square$

The notion of saturated conditional independence $I(Y, Z \mid X)$ over S is very closely related to that of a multivalued dependency (MVD) $X \twoheadrightarrow Y \mid Z$ over S, studied in the framework of relational databases [4,5,13,31,33,45]. Here, a set X of attributes is used to denote the X-value of a tuple over S, i.e., those tuple components that appear in the columns associated with X. Indeed, $X \twoheadrightarrow Y \mid Z$ expresses the fact that an X-value uniquely determines the set of associated Y-values independently of joint associations with Z-values where $Z = S - XY$. Thus, given a specific occurrence of an X-value within a tuple, so far not knowing the specific association with an Y-value and Z-value within this tuple, and then learning about the specific associated Y-value does not provide any information about the specific associated Z-value. Previous research has established an equivalence between the implication problem for SCI statements and that for MVDs [52]. The equivalence proof between the implication of MVDs and that of SCI statements was only syntactic in the sense that they both enjoy a common axiomatization. In addition it is known that the implication problem of MVDs can be decided in almost linear time [13], and that it is equivalent to that of formulae in a Boolean propositional fragment $\mathcal{F}'$ [45], even in nested databases with finite list, and record constructors [21]. Indeed, Sagiv et al. showed that it suffices to consider two-tuple relations in order to decide the implication problem of MVDs [45]. This enabled them to define truth assignments from two-tuple relations, and vice versa, in such a way that the two-tuple relation satisfies an MVD if and only if the truth assignment is a model for the $\mathcal{F}'$-formula that corresponds to the MVD. It follows from these results that the implication of SCI statements is equivalent to that of $\mathcal{F}'$-formulae. In this article, we will extend the trinity of implication problems between MVDs, the fragment $\mathcal{F}'$, and saturated conditional independence statements, to a trinity of implication problems between Delobel's class of full first-order hierarchical decompositions, the fragment $\mathcal{F}$, and generalized saturated conditional independence statements. The following example illustrates the equivalence of implication problems between the generalized SCI statements from Example 1 and that of the fragment $\mathcal{F}$.

Example 2. Let $L = \{m', a', r', c', f', l', s'\}$ denote a set of propositional variables. Let Σ' consist of the formulae

$$\neg m' \vee (a' \wedge r' \wedge c' \wedge f' \wedge l') \vee (s' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r' \wedge c')$$

and

$$\neg m' \vee (a' \wedge r' \wedge c' \wedge f' \wedge l') \vee (s' \wedge c' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r'),$$

let φ'_1 be the formulae

$$\neg m' \vee (c' \wedge a' \wedge r' \wedge f' \wedge l') \vee (s' \wedge a' \wedge r' \wedge f' \wedge l') \vee (s' \wedge c' \wedge f' \wedge l') \vee (s' \wedge c' \wedge a' \wedge r'),$$

and let φ'_2 denote the formula

$$\neg m' \vee (a' \wedge c' \wedge r' \wedge f' \wedge l') \vee (s' \wedge c' \wedge r' \wedge f' \wedge l') \vee (s' \wedge a').$$

Indeed, Σ' logically implies φ'_1, but Σ' does not logically imply φ'_2. In fact, the truth assignment that assigns true to m', c', f', l', s' and false to a' and r' is a model for the formulae in Σ' but not a model for φ'_2. $\square$