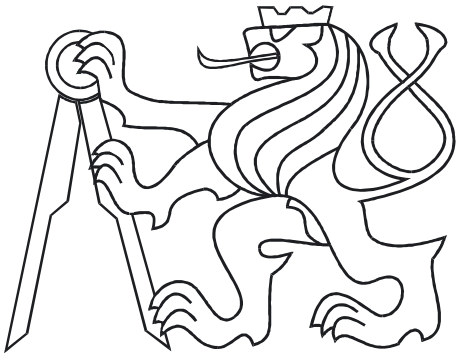
Czech Technical University in Prague



Doctoral Thesis

Czech Technical University in Prague

Faculty of Nuclear Sciences and Physical Engineering

Department of Mathematics

Ing. Václav Kratochvíl

Probabilistic Compositional Models: solution of an equivalence problem

Ph.D. Programme: Mathematical Engineering
Branch of study: Software Engineering

Doctoral thesis statement for obtaining the academic title of "Doctor",
abbreviated to "Ph.D."

Prague, September 2011

The doctoral thesis was produced in full-time manner PhD study at the Department of Mathematics of the Faculty of Nuclear Sciences and Physical Engineering of the CTU in Prague.

Candidate: Ing. Václav Kratochvíl
Department of Mathematics
Trojanova 13
120 00 Praha 2

Supervisor: Prof. Radim Jiroušek DrSc.
ÚTIA AV ČR, v.v.i.
Pod Vodárenskou věží 4
182 08 Praha 8

Opponents: .....

.....

.....

The doctoral thesis statement was distributed on .....

The defence of the doctoral thesis will be held on ..... at ..... before the Board for the Defence of the Doctoral Thesis in the branch of study Mathematical Engineering in the meeting room No. ..... of the Faculty of Nuclear Sciences and Physical Engineering of the CTU in Prague.

Those interested may get acquainted with the doctoral thesis concerned at the Dean Office of the Faculty of Nuclear Sciences and Physical Engineering of the CTU in Prague, at the Department for Science and Research, Břehová 7, Prague 1, 115 19.

doc. Ing. Zuzana Masáková Ph.D.
Chairman of the Board for the Defence of
the Doctoral Thesis in the branch of study
Mathematical Engineering
Faculty of Nuclear Sciences and Physical Engineering
of the CTU in Prague

1 Introduction

The interest in computer-aided reasoning within computer science dates back to the very early days of *Artificial Inteligence*, when much work had been initiated for developing computer programs to solve problems that require a high degree of intelligence.

In 1956, at *Dartmouth Conference*, the term "Artificial Intelligence" was coined by *John McCarthy*, following by an influential proposal for building automated reasoning systems. This proposal calls for a system with two components: a *knowledge base*, which encodes what we know about the world, and a *reasoner* (inference engine), which acts on the knowledge base to answer queries of interest. While the knowledge base can be domain-specific, changing from one application to another, the reasoner is quite general and fixed. This aspect became the basis for a class of reasoning systems known as knowledge-based or model-based systems. We will also subscribe to this knowledge-based approach for reasoning, except that our knowledge bases will be multidimensional probability distributions (modeled by compositional models in our case) and our reasoning engine will be based on probability theory.

Size of joint probability distribution grows exponentially with the number of variables of interest. Many reasoning tasks arising in uncertainty reasoning in artificial intelligence can be considerably simplified if a suitable concept of relevance or irrelevance of symptoms or variables is taken into consideration. The conditional irrelevance in probabilistic reasoning is modeled by means of the concept of probabilistic *conditional independence* (CI) [6]. Since every CI-statement can be interpreted as a certain qualitative relationship among involved variables, the dimensionality of the problem can be reduced and a more effective way of storing the knowledge base may be found.

2 State of the Art

Compositional models are probabilistic models presenting an alternative to well-known Bayesian networks. But unlike the graphical models, the compositional models represent a purely algebraic approach based on directly assembling low-dimensional probability distributions with the aid of the *operator of composition*, without the necessity to employ graphs. Yet graphs (hypergraphs) can be used for the sake of visualization.

It can be shown that both approaches – Bayesian networks and compositional models – are equivalent in the sense that they can both represent the same class of probability distributions [2], but the compositional models appear to be less computationally demanding for the frequent task of computing marginal probability distributions [1]. Another advantage may be that, by redefining the operator of composition, three different frameworks for uncertainty description may be considered: probability and possibility theories, and Dempster-Shafer theory of belief functions. Special operators of composition are introduced

within all three frameworks in [4].

The basic properties of compositional models are described well in [3] and in the recent review paper [5]. But the approaches for understanding how CI is coded in a model structure are rather undeveloped. Similar to other probabilistic models, the structure of the model induces a system of CI relations. Note that the representation of CI is imperfect in case of compositional model structures. On the one hand, not all sets of CI relations that might be satisfied by a probability distribution can be represented by such a structure. On the other hand, two or more structures often represent the same CI relations. When model structures do represent the same independence relations, we say they are *independence equivalent* (sometimes it is called Markov equivalence.)

3 Goals of the Thesis

The so-called *equivalence problem* is how to recognize whether two given compositional model structures are independence equivalent. It is also of special importance to have a simple rule that allows us to recognize this equivalence, and an easy way to get from one structure to another in terms of some elementary operations on structures. Another very important aspect of the equivalence problem is the ability to generate all structures equivalent to a given one.

The thesis elaborates on two major topics in detail, namely, solution of the equivalence problem in theory of compositional models, and its usage in other areas like *conditioning* probability distributions represented by a compositional model.

The latter topic was motivated by an open problem of *model flexibility* posed by Radim Jiroušek in [3]. Note that in the case of conditioning of a probability distribution represented by a compositional model, the problem is converted to that of respective model flexibility.

4 Solution

Our program of investigation of equivalence problem in compositional models was directed in several successive steps. First, we have found various structural properties invariable in the class of equivalent structures. This means that they are necessary to guarantee the possible equivalence of given structures. The first two such properties were inspired by solution of the similar problem in the framework Bayesian networks and acyclic directed graphs. Unfortunately, these properties are not very suitable when one works with compositional models. That is why we derived *non-trivial sets* and related *weak* and *strong structure core*. We defined the *reduced structure* and and *formal ratio* of the structure. Thus we obtained a list of possible candidates for direct characterization of independence equivalence.

Then we started to look for operations that do not affect the induced independence

model. Knowledge of the weak core (and its properties and consequences) significantly narrow the space of possible operations. We have identified three elementary operations on structures. They allow us to convert any structure into another independence equivalent one and hence to generate the complete class of structures that are equivalent with the given one.

Finally, using a cycle of implications, we show that some of the invariant properties (or their combinations) are not only necessary, but also sufficient to guarantee the equivalence of considered structures. In other words, they are real direct characteristics of equivalence.

The second topis of the thesis had a more probabilistic nature. It is devoted to the study of applying an equivalence problem solution to other open problems. First we investigate the impact of generalized elementary operations, originally introduced only for structures, to the corresponding probability distributions from compositional model with the structure. We identified other necessary conditions guaranteeing that probability distributions represented by respective compositional models are identical. Finally, the thesis ends with a partial solution for determining *flexibility* and connected the *conditioning problem*, using a solution of the equivalence problem.

5 Thesis Contributions

Let us briefly summarize the main original results achieved in the thesis:

- introduction of three different direct characterizations of independence equivalence
- indirect characterization of independence equivalence consisting of three elementary operations on structures
- a unique representative of a class of independence equivalent structures – formal ratio
- partial solution of compositional model flexibility regarding a conditioning problem.

6 Conclusions

An integral part of the work with multidimensional probabilistic models, and in particular their learning, is the perfect knowledge of the conditional independence relations of the model - a list of conditional independence relations valid for the model and induced by its structure. Note that the representation of conditional independence relations by compositional model structure is imperfect - two or more structures may represent the same conditional independence relations - they are independence equivalent. How to recognize whether two structures represent the same set of independence relations, how to transform the one structure into another equivalent one in terms of some elementary operations, and

the ability to generate all structures equivalent with a given one - it's all part of the problem that is known as equivalence problem.

In this thesis we published the complete solution of the problem. Moreover, we illustrated its usage in partial solution of the conditioning problem. Above that, a unique representative of a class of equivalent structures was discovered.

References

[1] V. Bína, R. Jiroušek: *Marginalization in Multidimensional Compositional Models*, Kybernetika 42, (2006), pp. 405-422.

[2] R. Jiroušek: *What is the difference between Bayesian networks and compositional models?*, Proceedings of the 7th Czech–Japan Seminar on Data Analysis and Decision Making under Uncertainty, Osaka, Japan, (2004), pp. 191-196.

[3] R. Jiroušek: *Multidimensional Compositional Models*. Preprint DAR - ÚTIA 2006/4, ÚTIA AV ČR, Prague, (2006).

[4] R. Jiroušek: *Conditional independence and factorization of multidimensional models* Fuzzy Systems, IEEE World Congress on Computational Intelligence, Hong Kong, (2008), pp. 2359-2366.

[5] R. Jiroušek: *Foundations of compositional model theory*, International Journal of General Systems - Volume 40, Issue 6, (2011), pp. 623-678.

[6] J. Pearl: *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference*, Margan Kaufmann, San Mateo, CA, (1988).

List of Author's Publications

Journal articles

[1] V. Kratochvíl: *Characteristic Properties of Equivalent Structures in Compositional Models*, International Journal of Approximate Reasoning vol. 52,5 (2011), pp. 599-612.

Other publications

[2] V. Kratochvíl: *Conditioning and Flexibility in Compositional Models*, Proceedings of 14th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty CJS 2011 , Eds: Barták Roman, Hejnice (2011)

[3] V. Kratochvíl: *Relationship between properties characterizing independence equivalence in Bayesian networks and compositional models*, Proceedings of the 13th Czech-Japan Seminar on Data Analysis and Decision Making in Service Science, Eds: Itoh Takeshi , Suzuki Kenichi, Otaru - Japan (2010)

[4] V. Kratochvíl: *Equivalence Problem in Compositional Models*, WUPES'09, Eds: Kroupa T., Vejnarová J., WUPES'09, Liblice (2009).

[5] V. Kratochvíl: *Motivation for different characterization of Equivalent Persegrams*, Proceedings of the 12th Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty, Eds: Novák V., Pavliska V., Štěpnička M., Litomyšl (2009).

[6] V. Kratochvíl: *An Effective Algorithm to Search Reductions in Compositional Models*, Proceedings of Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty /10./ , Eds: Kroupa T., Vejnarová J., Liblice (2007)

[7] R. Jiroušek, V. Kratochvíl: *Marginalization algorithm for compositional models*, Information Processing and Management of Uncertainty in Knowledge-Based Systems, p. 2300-2307 , Eds: Bouchon-Meunier B., Yager R. R., Paris (2006)

[8] V. Kratochvíl: *Different Approaches of Study Direct Equivalence Characterization*, Doktorandské dny 2009, sborník workshopu doktorandů FJFI oboru Matematické inženýrství, Eds: Ambrož P., Masáková Z., ČVUT Praha (2009), pp. 101-110. ISBN 978-80-01-04436-0

7 Summary

In the case when one tries to represent a knowledge in a form of probability distribution, one hit the problem that the size of probability distribution grows exponentially with the number of variables of interest. This problem is usually bypassed by involving one of the probabilistic models. They represent the multidimensional probability distribution in the form of differently interrelated collection of low-dimensional probability distributions using the notion of conditional independence.

Compositional model theory (originally developed by Radim Jiroušek) represents an alternative approach to probabilistic models, mainly graphical ones. A compositional model may be defined as a multidimensional distribution assembled from a sequence of low-dimensional unconditional distributions (the so-called generating sequence), with help from the operator of composition. Fragmenting the multidimensional distribution into a generating sequence brings forth several complications. While a model is put together, a system of (un)conditional independencies is simultaneously introduced by the structure of the generating sequence. This system of independencies - the so-called induced independence model – is valid for any compositional model defined by a generating sequence with this structure.

This thesis familiarizes the reader with new results in this theory, namely with the complete solution of the equivalence problem. The equivalence problem is the problem of recognizing whether two given structures over the same set of variables induce the same independence model. In this case, we present three different simple rules to recognize that two structures are equivalent. We also present three elementary operations on structures such that we can easily convert one structure into an equivalent one in terms of these operations. Moreover, we show that one can generate all structures equivalent to a given one using these operations, and we identified a unique representative of such a class of equivalent structures.

Using above mentioned elementary operations and our knowledge of equivalence problem solution, we were able to examine the problem of conditioning a probability distribution represented by a compositional model, as well as a connected problem of generating sequence flexibility, from a different perspective. A partial solution of this problem is published.

8 Resumé

Aby mohla být teorie pravděpodobnosti použita při řešení praktických problémů, je nutno vypořádat se s problémem, který je některými autory nazýván jako prokletí multidimensionality. Při řešení praktických úloh je totiž potřeba pracovat alespoň se stovkami znaků (proměnných) a použijeme-li pravděpodobnostní distribuci jako prostředek k uchování znalostí, narazíme na problém, že velikost distribuce roste exponenciálně s počtem proměnných. Tento problém lze úspěšně obejít použitím některého z pravděpodobnostních modelů jako jsou například Bayesovské sítě.

Takovým pravděpodobnostním modelem jsou i kompozicionální modely, jenž představují alternativu zejména ke grafickým modelům. Kompozicionální model lze definovat jako mnohodimenzionální pravděpodobnostní distribuci sestavenou z posloupnosti málo-dimenzionálních distribucí (tzv. generující posloupnosti) pomocí operátoru kompozice. Rozložení mnohodimenzonální distribuce do takovéto posloupnosti sebou přináší několik komplikací. Postupným skládáním modelu je do reprezentované distribuce zaváděn systém podmíněných nezávislostí, indukovaný právě strukturou generující posloupnosti. Tento systém nezávislostí - tzv. indukovaný nezávislostní model - platí pro libovolný model definovaný generující posloupností s touto strukturou.

Tato disertační práce přináší nové výsledky v teorii kompozicionálních modelů, zejména však úplné řešení problému ekvivalence. Problémem ekvivalence rozumíme problém jak rozpoznat zda dvě dané struktury nad stejnou množinou proměnných indukují stejný nezávislostní model. Publikované řešení obsahuje jednak tři různá pravidla k určení případné ekvivalence daných struktur, jednak sadu elementárních operací nad strukturou pomocí kterých lze jednoduše konvertovat strukturu na jinou, libovolnou - s ní ekvivalentní. Dále byl zaveden unikátní reprezentant takové třídy ekvivalentních struktur.

Praktické použití řešení problému ekvivalence je ilustrováno na problému podmiňování pravděpodobnostní distribuce reprezentované kompozicionálním modelem. Díky znalosti elementárních operací zmíněných výše, a jejich zobecněním na generující posloupnosti, jsme byli schopni problém podmiňování částečně vyřešit.