Learning Causal Bayesian networks from Literature Data

Péter Antal [1)] and András Millinghoffer [1)]

[1)] Budapest University of Technology and Economics,
Department of Measurement and Information Systems, Hungary

e-mail of contact author: milli@mit.bme.hu

Abstract: We propose two machine learning methods based on Bayesian networks to discover automatically real world causal relations from scientific publications. The first method assumes that the occurrence of causal mechanisms (and the corresponding entities) in the publications follows a transitive scheme, the second method assumes that the causal mechanisms occur independently. We perform an evaluation of these methods in the ovarian cancer domain, because of the availability of an expert causal model as gold-standard reference and various collections of scientific publications as source. The evaluation shows that the fully observable transitive model and the intransitive model with hidden variables perform comparable to the performance of a human expert and the second, computationally more complex method allowing hidden variables proved to be slightly better.

Keywords: Bayesian network, Literature data, Learning

Introduction

In biomedical sciences the emergence of the web-based collective electronic knowledge has posed new challenges for many disciplines such as for knowledge engineering to make available the voluminous, uncertain and frequently inconsistent knowledge fragments, for machine learning to cope with high-dimensional data and to incorporate a priori knowledge in various learning and discovery algorithms, for natural language processing to retrieve relevant raw information and to extract relevant implicit information [1], for philosophy of science to understand the nature of this new collective, distributed research, and for biomedical sciences themselves to investigate high dimensional and more complex hypotheses and on the other hand to integrate this frequently manually curated and constantly updated knowledge in medical practice also (e.g. in online decision support). Despite recent trends related to the concept of semantic web aiming to broaden the scope of formal knowledge bases in biomedical domains [2], even to advocate the use of a formal supplementary abstract for research articles, the free text electronic literature is still the central part of the web-based collective knowledge, and this central role probably remains in the near future, because of (1) the rapidly expanding frontiers, (2) the integration of separated levels such as of biochemical, biological, medical and clinical level, (3) the related refinement of current knowledge by specialization and conditionalization and finally (4) because of the spread of free access to full electronic papers (for the relation of classical free text and web-based knowledge, see [3, 4, 5, 6, 7, 8].

Fig. 1 shows our assumption that various measurement and experimental methods, statistical approaches and the subsequent analysis and publication methods behind publications can be modeled as a collection of binary Bayesian networks (see

Section “Bayesian belief networks”) reflecting different aspects (fragments) of domain causality with different noise and bias (for an overview of a related fragmentation by levels, see [9]). For a detailed derivation and interpretation see Section “Literature mining with Bayesian networks”, in short the interpretation of such a binary network basically stems from the generative view of causal relevance patterns in the domain literature, such as e.g. the report of a (tentatively) causally related set of entities or the report of a (tentatively) causally related chain of entities. Note that the transitivity of dependencies is satisfied in binary networks [10], fitting to an expectation about the transitivity of causal explanation. Another interesting property that such a binary network on one hand expresses decomposed uncertainty over domain mechanisms, and on the other hand expresses beliefs about the overall application of the method, including the design of the application (i.e. interventionist setups of variables) and even the habits of publishing results from the method. Such a binary network is a biased and noisy representation of the causal structure of the real domain. This later causal interpretation of the dependency model over the literature providing the generative (interventionist) interpretation is only an approximation, as there are certainly many factors outside the domain variables, e.g. describing general, domain independent constraints on publications.

Moreover, we assume that the generative mechanism of the discussion, citation or partial overview of the more authenticated domain knowledge similarly can be represented by a binary network representing the uncertainties over causal mechanisms in the domain, i.e. that this network can be similarly interpreted as a noisy and biased representation of the more or less explored causal mechanisms in the domain (shown in the lowest level in Fig. 1).

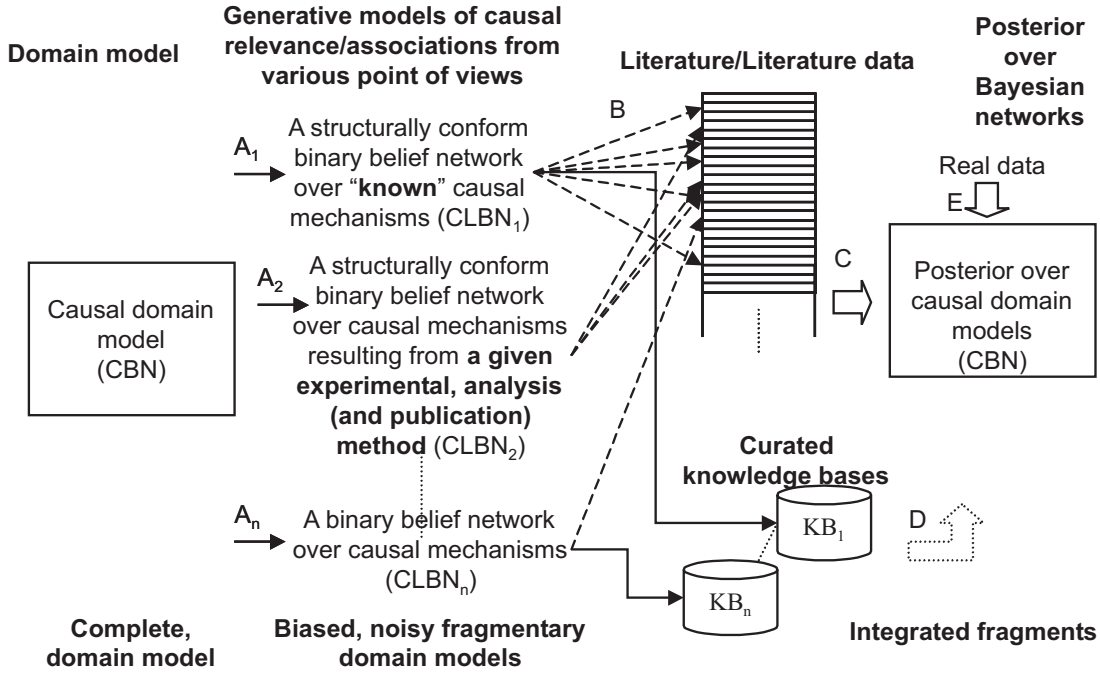

Figure 1: The overview of typical fragmentation of knowledge in biomedicine and a possible integration through literature data. Arrows $A_1, \ldots, A_n$ indicate generative causal models of report of causal relevancies from various point of views, such as different experimental setup, analysis method and publication style. These biased, noisy, fragmentary, consequently uncertain domain theories are represented by a special binary belief network, expressing beliefs over causal mechanisms from the

corresponding point of view, with respect to the current state-of-the-art of the domain theory. Arrow *B* shows the appearance of mechanisms in scientific publications. Arrow *C* indicates the usage of the overall publications to integrate the various fragments into a combined causal domain model through literature data. Arrow *D* indicates a not covered part here, when the accepted domain theories are represented in knowledge bases, which are later transformed into a priori distribution for the subsequent Bayesian learning. Arrow *E* shows the potential use of real data in the Bayesian learning.

Though to our knowledge this assumption has not been formalized earlier (for an overview of the biological rationale of a shallow statistical cooccurrence analysis see [11]), it was probably always tacitly assumed in the usage of the associative analysis of domain literature, such as in cooccurrence analysis or in clustering [11, 12, 13, 14, 15].

We propose two Bayesian network models to discover automatically real world dependency relations from scientific publications. The first method assumes that the reporting activity of causal mechanisms follows a transitive scheme, the second method assumes that the causal mechanisms in the domain are reported autonomously (i.e. more or less independently).

The application domain: ovarian cancer

We perform an evaluation of these methods in the ovarian cancer domain, in which various collections of scientific publications are available as source and an expert causal model as gold-standard reference. In the experiments we used a total of sixteen variables, which had been previously evaluated as highly relevant domain variables. Furthermore, a leading expert in the ultrasonography of ovarian tumors constructed a model with 'highly' and 'moderately relevant' relations, shown in Fig. 2 and provided a causal ordering of the variables used in this paper.

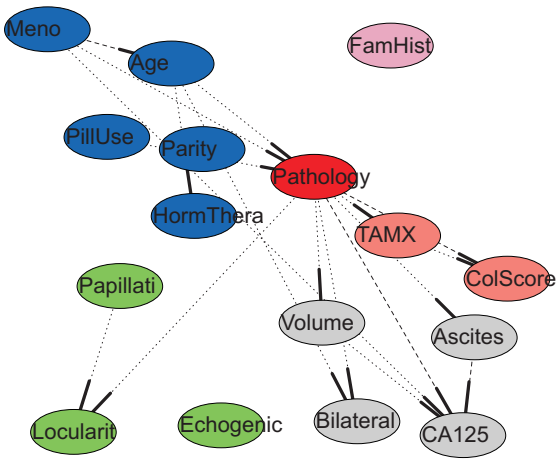

Figure 2: The expert model: edges occurring in the highly relevant model are indicated by dashed lines, edges in the moderately relevant model are indicated by dotted lines.

We asked medical experts to select the *most relevant* journals for the domain and performed the query 'ovarian cancer' in the PubMed database[1] between 1998 and 2002 which resulted 500 papers. These publications were converted to a vectorial representation providing the literature data used in the paper (for the description of the domain, model construction and conversion steps of literature, see [16]).

Bayesian belief networks

A belief network represents a joint probability distribution over a set of variables [10]. We assume that these are discrete variables. The model consists of a qualitative part (a directed graph) and quantitative parts (dependency models). The vertices V_i of the graph represent the random variables X_i and the edges define the direct dependencies (each variable is probabilistically independent of its non-descendants given its parents [10]). There is a probabilistic dependency model for each variable that describes its dependency on its parents.

Beside the parametric aspects of Bayesian network representation (i.e. providing an efficient representation of high dimensional joint distributions), it has further advantages with respect to the structure of the domain variables. It provides an efficient and graphical representation of the conditional independencies in the domain with standard probabilistic semantics and enables inferences on conditional independencies irrespectively of the underlying parametrization [10]. Furthermore, it provides a representation of causal domain models and enables causal inferences [17]. In the paper we follow a causal interpretation, and we use the single causal ordering over the variables to ensure the validity of this interpretation, because (1) the selected variables are high-level clinical variables, (2) only the sixteen most important variables are used out of the fifty-six from the study, (3) there are logical dependencies between certain values, e.g. exclusions (for a full fledged causal discovery and causal interpretation of Bayesian networks see also [18, 19]). By the careful selection of variables and definition of a causal order, we ensure the interpretation of the dependencies as autonomous, causal mechanisms in the domain.

In the Bayesian framework the uncertainty over the structure of the domain model is represented by a distribution over the space of directed graphs. Assuming structure independence [20, 21], the probability of a domain model can be decomposed into the product of probabilities of the dependencies in the domain, which fits in the causal interpretation of the structure. Additionally, it is frequently assumed that the belief in substructures (i.e. in parental sets) can be further decomposed into a product of probabilities corresponding to the belief that an individual parent is a member of the parental set (i.e. is a direct cause). Later, discussing the application of the noisy-OR canonical local dependency model in generative models of publications, we refer to this assumption as edge independence.

The existence of a closed-form formula for the structure reduces the structure learning to a discrete search. On the contrary, the learning from incomplete data is computationally more demanding as there is no closed formula for the score of a structure, consequently an embedded parameter optimization is necessary to determine a fittness score for each structure [22].

[1] http://www.ncbi.nlm.nih.gov/PubMed/

Literature mining with Bayesian networks

The first step in the investigation of possible Bayesian network models to analyze the literature is to consider the types of the variables and their values and interpretation. Our goal is to analyze the pattern of occurrences to discover latent causal domain models (c.f. information extraction, see [1]), so the literature is converted into a vectorial representation preserving only the occurrence of the variables in each scientific paper.

Adopting the centrality of causal understanding and explanation in scientific research [23, 24], we also assume respectively the centrality of causal explanations in scientific publications. We accept ‘causal relevance' as possible interpretation, more specifically the ‘explained' (explanandum) and ‘explanatory' (explanans), additionally, we allow the ‘described' status. This implicitly means that we assume that publications either contain descriptions of domain concepts without considering their relations or occurrences of entities participating in known or unknown (latent) causal relations (c.f. Causal Markov Condition [18, 17, 19]).

Now we consider the types of variables, local dependency models and structures to model the occurrence pattern of the accepted three roles of domain variables, as of causal relevance (explanatory and explained) and descriptional. Of course, this is a particularly ambitious attempt and serious simplifications have to be accepted, because a probabilistic or causal model over these roles of the domain variables means a generative model of scientific explanation in publications, with certain implications to scientific research itself (consider that research and publication can be modeled as governed by the discrepancy between the published and believed “truth"). Furthermore, beside the ‘description', we should model the transitive nature of causal explanation over mechanisms, e.g. that causal mechanisms with a common cause or with a common effects are surveyed in an article, or that chains of causal mechanisms are tracked to demonstrate a causal path. On the other hand, we have to model the lack of transitivity, i.e. the incompleteness of causal explanations, e.g. that certain variables are assumed as explanatory, others as potentially explained, except in survey articles that describe an overall domain model.

The simplest, atomistic approach is to assume complete independence of the report of the causal mechanisms and univariate descriptions. Indeed, this is the currently prevailing assumption, because all the information extraction methods that extract, analyze and provide result separately for the singular relations rely on this assumption. These methods also assume that the singular reports of the causal mechanisms and univariate descriptions can be sufficiently identified as shown in Fig. 3. Note that these methods are not intended to discover new latent dependencies or mechanisms that are conjectured and loosely articulated or indicated by only associative patterns.

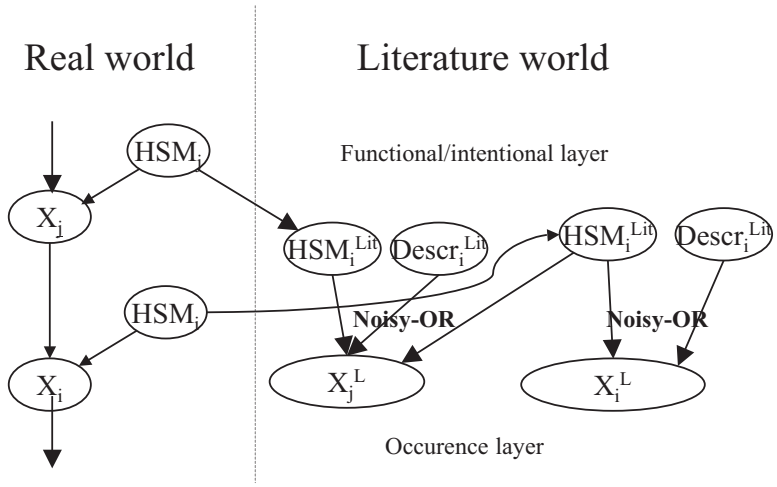

Figure 3: The separated extraction and analysis of the singular relations with the underlying assumption of complete independence of the report of the causal mechanisms and univariate descriptions. We assume that the belief in the (hidden sub-)mechanism (HSM) is an important factor influencing the publication, i.e. this factor establishes the link between a belief in real world mechanism and the frequency of occurrence in the literature world.

If the explanatory, explained and descriptive roles are not known and mainly unstructured causal relevance associations or tentative relations are reported, which cannot be identified sufficiently with linguistic methods, then the domain wide discovery methods can support the construction of consistent identification of relations from the simplified vectorial text representation. In the construction of a corresponding model, we maintain the assumption that the reporting of the causal mechanisms and univariate descriptions are independent, i.e. in the exploratory interpretation it means that we assume that a fragmentary domain theory corresponding to a given experimental, analytical and publication method results in such independent causal relevance associations.

We propose a two-layered Bayesian network structure. The upper layer contains variables corresponding to the possible causal roles of the entities, such as described, explained or explanatory (we treat explanatory as cause and explained as effect). In the explanatory interpretation these represent the authors' intentions, which are externalized possibly as occurrences of the entities in the publication. In the exploratory interpretation these represent the bias and incompleteness of a given experimental technique with respect to the causal relevance and causal roles.

The lower, external (textual) layer contains the observable occurrence of the entities. An external variable depends only on the variables denoting the causal roles related to the corresponding causal mechanism (i.e. it is independent of other external variables, such as the number of reported domain entities in the paper and it is independent of other non-external variables of neighboring causal mechanisms). The steps of derivation from the first atomistic model to this more entity oriented model are shown in Fig. 4. This model extends the individual mechanism oriented information extraction methods with supporting the domain wide, consistent interpretation of causal roles, but still cannot model dependencies (e.g. transitivity) between the reporting of the mechanisms (c.f. causation on the

experimental/intentional level).

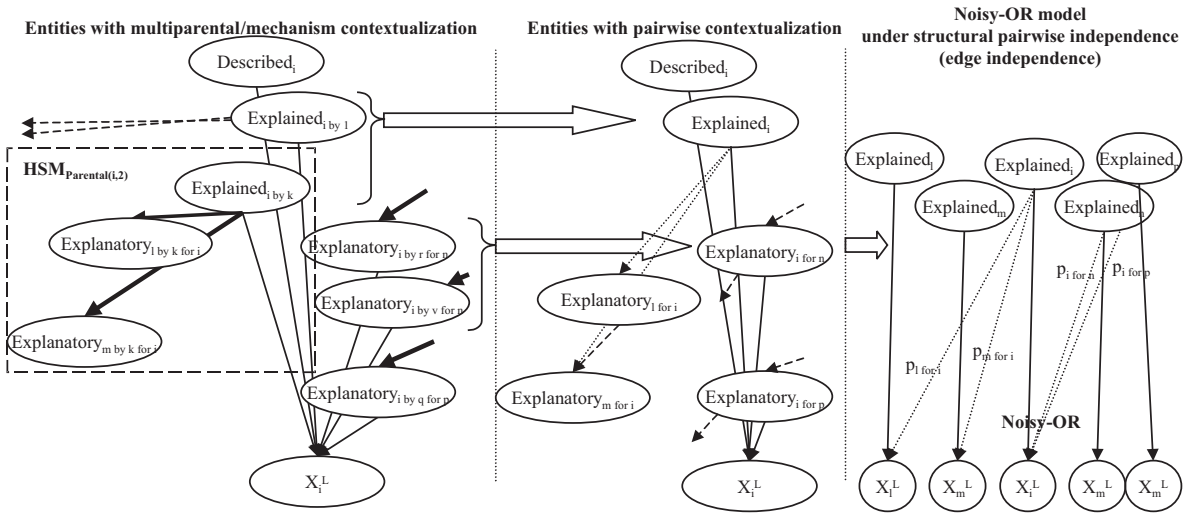

Figure 4: The steps of derivation of the intransitive model with noisy-OR local dependencies from the first atomistic model.

A further assumption, mainly motivated by the explanatory interpretation, is that the parental sets are composed of independent factors, in other words that the belief in a mechanism is the product of the individual beliefs in causes. Consequently we use noisy-OR canonic distributions for the children in the lower layer and interpret the occurrence of a variable in a paper as described or explanatory or explained. In a noisy-OR local dependency [17]), the edges can be labeled with an inhibitor parameter, inhibiting the OR function (i.e. the probability of an implicative edge). We set this parameter to zero for the 'explained to occurrence' edges, i.e. we assume that if the intended function is explained, then the variable is mentioned.

To devise a model more advanced with respect to the explanatory and exploratory interpretation, we relax the assumption of independence between the variables in the upper layer representing causal roles, but maintain that an external variable depends only from the variables in the upper layer that participate within the same causal mechanism. First we consider if the reporting of causal mechanisms is dependent in a causally transitive way, i.e. if we allow dependencies between the explained and the explanatory roles of the variables. In the explanatory interpretation this means, i.e. if a variable is explained, then it influences its explanatory role for other variables. In the exploratory interpretation, if a variable arises as an effected variable, then it influences its arise as cause. If this transitivity dependency (explained to explanatory) is uniform in each pairwise context, i.e. the explanatory role is not pairwise context dependent, then a single explanatory variable can represent this role (earlier we merged the multiparental contextualization, now the pairwise contextualization for the explanatory variables). Full transitivity means that this is an equivalence. In the explanatory interpretation this means, if a variable is explained, then it can be explanatory for any other variable. In the exploratory interpretation, that variables arise in general as causally relevant, both as an effective and causative variable. In full transitive case the variables representing various causal roles such as the status of being explained and being explanatory for another variable can be merged into one variable. Furthermore we assume full transparency, i.e. the full observability of

causal relevance. Fig. 5 shows these steps.

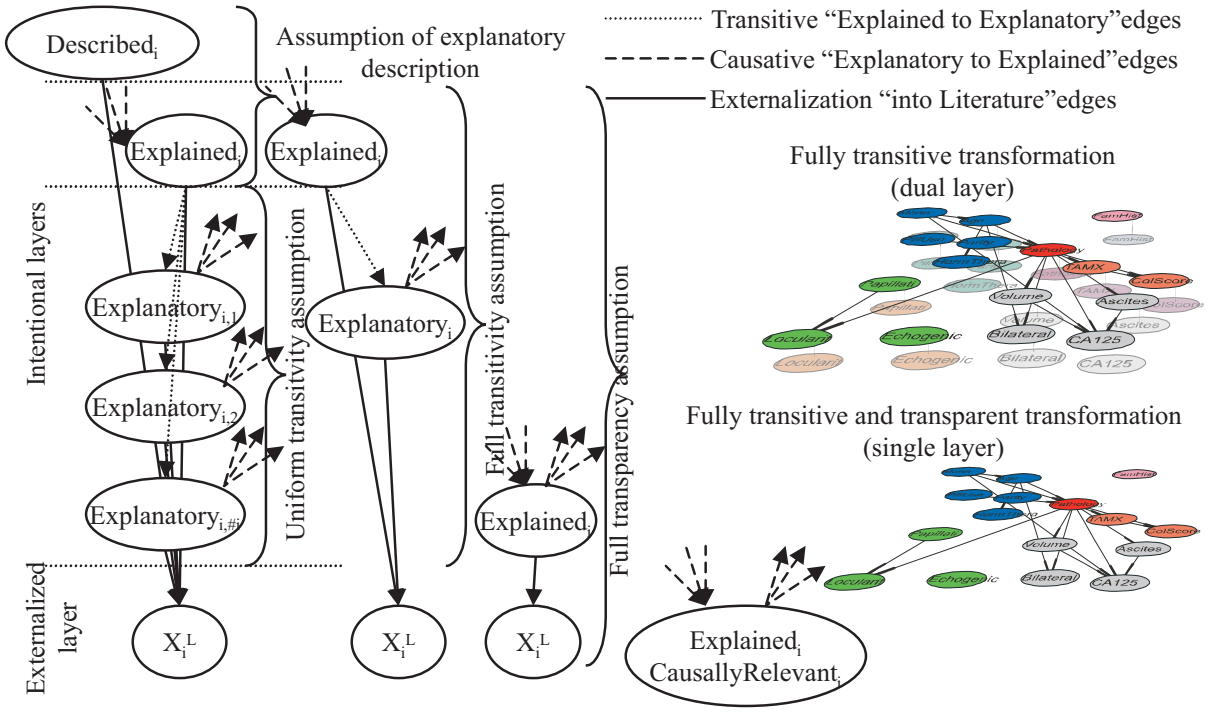

Figure 5: The steps of derivation of the transitive model from the first atomistic model.

A consequence of the full transparency is e.g., that under this interpretation the lack of occurrence of an entity in a paper means causal irrelevance and not a neutral omission, in other words there are no missing values. With full transitivity assumption this would also imply that we model only full survey papers, but the general, unconstrained multinomial dependency model used in the transitive Bayesian network provides enough freedom to avoid this as discussed below. A possible semantics of the parameters of the binary, transitive Bayesian network $P(X_i|Parents(X_i))$ can be derived from our causal stance that the presence of an entity X_i is influenced only by the presence of its potential explanatory entities, i.e. its parents. Consequently, $P(X_i=1|Parents(X_i)=\underline{x_i})$ can be interpreted as the belief that the parental variables that are present can explain the entity X_i as causes. A more strict interpretation requires necessity beside sufficiency, in which $P(X_i=1|Parents(X_i)=\underline{x_i})$ denotes a belief that the parental variables that are present are the sufficient and necessary causes. The multinomial model allows that at each node there are entity specific constants combined into the parameters that are not dependent on other variables, permitting the deviation from this interpretation and modelling (1) the description of the entities, (2) the initation of the transitive scheme of the causal explanation (the assumption of causally not explained entities) and (3) the reverse effect of not continuing the transitive scheme. The detailed discussion of this model is outside the scope of this paper, so we conclude here that this model allows partial explanations also. Note that a "backward" model using an effect-to-cause orientation is similarly an interesting model of publications (c.f. means-ends analysis), in which the noisy-OR dependency model can be also used as in the intransitive model.

Independently of the selected model, the result of learning of Bayesian networks from literature data can be manifold, e.g. an a posteriori distribution over the structures or substsructures (‘features', see [25]) or a maximum a posteriori network structure and the corresponding parameters. Note that in the later parametric case, because of the special structural interpretation of the binary network (i.e. ‘causal relevance') the parameters and the standard parametric inference in such a network can be interpreted structurally and can be converted into an a priori distribution for a subsequent learning. Another approach is to use the a posteriori distribution over the structures of the binary Bayesian literature networks as an a priori distribution over the structures of the real Bayesian networks with possibly multivalued or continuous variables.

Results

The structure learning of the transitive model is achieved by an exhaustive evaluation of parental sets using BD_{eu} score [26] up to maximum three parents, which was a technical choice to be compatible with the learning of the intransitive model with hidden variables. The final network is shown in Fig. 6.

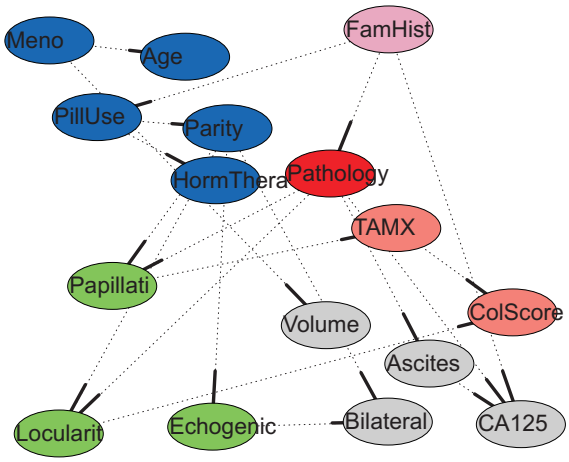

Figure 6: The transitive Bayesian network model with multinomial conditional tables.

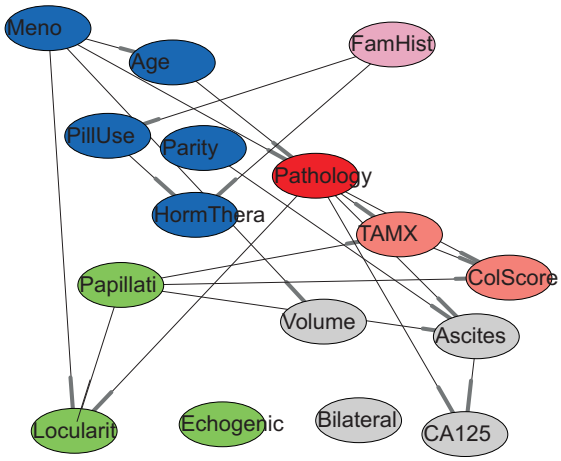

Figure 7: The intransitive Bayesian network with noisy-OR local conditional dependency models (Note that this is the conversion of the two-layered Bayesian network with hidden variables).

The structure learning of the two-layered model is computationally more complex, because the evaluation of a structure requires the optimization of parameters, which can be performed e.g. by a gradient-descent algorithm. The possible (examined) structures have to meet that (1) variables have less than a fixed number of parents, limited to four in this experiment, because of the computational complexity (2) only those variables in the upper layer can be the parents of an external variable that precede it in the causal order. Note that beside the optional three parental edges for the the external variables, we always force a deterministic edge from the corresponding non-external variable. During the parameter learning of a fixed network structure the non-zero inhibitory parameters of the lower layer variables are

adjusted according to a gradient descent method to maximize the likelihood of the data (see [22]). After the best structure is found, it has to be converted into the ordinary real world model by merging the corresponding pairs of nodes in lower and upper layer. The final network is shown in Fig. 7.

We compared the trained models to the expert model using a quantitative score i.e. based on the comparison of the types of the pairwise relations in the models. Exploiting the causal interpretation of the structure we use the following types of pairwise relations:

- *Causal path (P):* There is a directed path from one of the nodes to the other.
- *Causal edge (E):* There is an edge between the nodes.
- *(Pure) Confounded (Conf):* The two nodes have a common ancestor. The relation is said to be pure, if there is no edge or path between the nodes.
- *Independent (I):* None of the previous (i.e. there is no causal connection between the nodes).

The difference of two model structures can be represented in a matrix containing the number of relations with a fixed type in the expert model and in the trained model (the type of the relation in the expert model is the row index and the type in the trained model is the column index). E.g. the element *(I, Conf)* shows the number of those pairs, which are independent in the reference model and are confounded in the examined. These matrices (i.e. the comparison of the transitive and the intransitive models to the expert's) are shown in Tables 1 and 2 respectively.

Table 1: Causal comparison of expert (row) and transitive (column) domain models

	I	Conf	P	E
I	14	14	12	12
Conf	6	14	0	2
P	44	48	24	14
E	14	6	4	12

Table 2: Causal comparison of expert (row) and intransitive (column) domain models

	I	Conf	P	E
I	44	0	0	8
Conf	14	8	0	0
P	82	18	20	10
E	8	4	2	22

Scalar scores to evaluate the goodness of the trained model can be derived from this matrix, e.g. a standard choice is to sum the elements with different weights. One possibility e.g. if we take the sum of the diagonal elements as a measurement of similarity. By this comparison, the intransitive model achieves 94 points, while the transitive only 64, so the intransitive preserves more faithfully the pairwise relations. Particularly important is the *(E,E)* element according to which 22 of the 36 edges of the expert model remains in the two-layered model, on the contrary the transitive model preserves only 12 edges.

Another penalizing score, which penalizes only the incorrect identification of independence (i.e. those and only those weights have a value of *1* which belong to the elements *(I,.)* or *(.,I)*, the others are zero), gives a score *102* and *112* for the transitive model and the intransitive respectively, suggesting that the intransitive model is too conservative and results overly sparse models.

Conclusion

We investigated the applicability of Bayesian network learning methods to discover a causal domain model. We proposed two machine learning methods based on Bayesian networks, the first method assumes that the reporting activity of causal mechanisms follows a transitive scheme, the second method assumes that the causal mechanisms in the domain are reported autonomously (i.e. more or less independently). We performed an evaluation of these methods in the ovarian cancer domain. The evaluation shows that the fully observable transitive model and the intransitive model with hidden variables performs comparable to the performance of a human expert and the second, computationally more complex method proved to be slightly better than the first one. In future, we plan to test more complex transitive models and extend these methods to incorporate more information extracted by linguistic techniques.

References

[1] L. Hirschman, J. C. Park, J. Tsujii, L. Wong and C. H. Wu: “Accomplishments and challenges in literature data mining for biology”, Bioinformatics, vol 18, pp 1553–1561, 2002.

[2] A. D. Baxevanis: “The molecular biology database collection: 2002 update”, Nucleic Acid Research, vol 30(1), pp1-12, 2002.

[3] T. Berners-Lee and J. Hendler: “Publishing on the semantic web”, Nature, vol 410, pp 1023-1024, 2001.

[4] M. Gerstein and J. Junker: “Blurring the boundaries between scientific 'papers' and biological databases”, Nature (web debate, on-line 7 May 2001), 2001.

[5] N. Shadbolt: “What does the science in e-science”, IEEE Intelligent Systems, vol 17(May/June), pp 2-3, 2002.

[6] H. Pearson: “The future of the electronic scientific literature”, Nature, vol 413 pp 1-3, 2001.

[7] R. J. Roberts et al.: “Building a 'genbank' of the published literature”, Science, vol 291, pp 2318-2319, 2001.

[8] Vanessa Speding: “Xml to take science by storm”, Scientific Computing World, Supplement (Autumn), pp 15-18, 2001.

[9] M. Vidal: “A biological atlas of functional maps”, Cell, vol 104 pp 333-339, 2002.

[10] J. Pearl: “Probabilistic Reasoning in Intelligent Systems”, Morgan Kaufmann, San Francisco, CA., 1988.

[11] B. Stapley and G. Benoit: “Biobibliometrics: Information retrieval and visualization from cooccurrences of gene names in medline abstracts”, In Proc. of Pacific Symposium on Biocomputing (PSB00), vol 5, pp 529-540, 2000.

[12] I. Iliopoulos, A. J. Enright, and C. A. Ouzounis: “Textquest: document clustering of MEDLINE abstracts for concept discovery in molecular biology”, In Proc. of Pacific Symposium on Biocomputing (PSB01), Hawaii, vol 58(2-3), pp 384-395, 2001.

[13] T. K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig: “A literature network of human genes for high-throughput analysis of gene expression”, Nature Genetics, vol 28, pp 21-28, 2001.

[14] P. Antal, P. Glenisson, G. Fannes, J. Mathijs, Y. Moreau, and B. De Moor: “On the potential of domain literature for clustering and Bayesian network learning”, In Proc. of the 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (ACM-KDD-2002), pp 405-414, 2002.

[15] Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade: “Association of genes to genetically inherited diseases using data mining”, Nature, vol 31 pp 316-319, 2002.

[16] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor: “Using literature and data to learn Bayesian networks as clinical models of ovarian tumors”, Artificial Intelligence in Medicine, 30. Special issue on Bayesian Models in Medicine, 2004.

[17] J. Pearl: “Causality: Models, Reasoning, and Inference”, Cambridge University Press, 2000.

[18] P. Spirtes, C. Glymour, and R. Scheines: “Causation, Prediction, and Search”, MIT Press, 2001.

[19] C. Glymour and G. F. Cooper: “Computation, Causation, and Discovery”, AAAI Press, 1999.

[20] W. L. Buntine: “Theory refinement of Bayesian networks”, In Bruce D'Ambrosio and Philippe Smets, editors, Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991), pp 52-60, Morgan Kaufmann, 1991.

[21] G. F. Cooper and E. Herskovits: “A Bayesian method for the induction of probabilistic networks from data”, Machine Learning, vol 9, pp 309-347, 1992.

[22] Stuart J. Russell, John Binder, Daphne Koller, and Keiji Kanazawa: “Local learning in probabilistic networks with hidden variables”, In IJCAI, pp 1146-1152, 1995.

[23] P. Thagard: “Explaining disease: Correlations, causations and mechanisms”, Minds and Machines, vol 8, pp 61-78, 1998.

[24] J. Woodward: “Scientific explanation”, In E. N. Zalta, editor, The Stanford Encyclopedia of Philosophy, 2003.

[25] N. Friedman and D. Koller: “Being Bayesian about network structure”, Journal of Machine Learning Research, vol 2, pp 1-30, 2002.

[26] D. Heckerman, D. Geiger, and D. Chickering: “Learning Bayesian networks: The combination of knowledge and statistical data”, Machine Learning, vol 20, pp 197-243, 1995.