Using Literature and Data to Learn Bayesian Networks as Clinical Models of Ovarian Tumors

Peter Antal [a] Geert Fannes [a] Dirk Timmerman [b] Yves Moreau [a]
Bart De Moor [a]

[a] *Department of Electrical Engineering, Katholieke Universiteit Leuven,*
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

[b] *Department of Obstetrics and Gynecology, University Hospitals Leuven,*
Herestraat 49, B-3000 Leuven, Belgium

Abstract

Thanks to its increasing availability, electronic literature has become a potential source of information for the development of complex Bayesian Networks (BN), when human expertise is missing or data is scarce or contains much noise. This opportunity raises the question of how to integrate information from free-text resources with statistical data in learning Bayesian networks. Firstly, we report on the collection of prior information resources in the ovarian cancer domain, which includes "kernel" annotations of the domain variables. We introduce methods based on the annotations and literature to derive informative pairwise dependency measures, which are derived from the statistical cooccurrence of the names of the variables, from the similarity of the "kernel" descriptions of the variables and from a combined method. We perform wide-scale evaluation of these text-based dependency scores against an expert reference and against data scores (the mutual information and a Bayesian score). Next, we transform the text-based dependency measures into informative text-based priors for Bayesian network structures. Finally, we report the benefit of such informative text-based priors on the performance of a Bayesian network for the classification of ovarian tumors from clinical data.

Key words: text mining, literature networks, Bayesian networks, prior incorporation, structure learning

1 Introduction

The increasing availability of electronic literature poses the question of how to use both domain knowledge and data efficiently in knowledge engineering, machine learning and decision support—theoretically as well as practically. This challenge is particularly acute in the complex and rapidly changing fields of medicine and genomics where much of the voluminous knowledge is only available as free text, scattered throughout the literature [22, 2]. The efficient extraction of domain knowledge from the literature and its incorporation into statistical models requires the extension of the representations and methods used for the integration of expert knowledge and data.

Bayesian networks (BN) are an attractive technique for tackling this task. Indeed, an important aspect of Bayesian networks is the possibility to incorporate many kinds of prior knowledge into learning—ranging from logical constraints on the model structure [43, 15, 33, 13] or qualitative monotonicity relations between the variables [47, 24] to prior distributions for network structures and parametrizations of local dependencies [10, 15, 25, 12]. For the incorporation of prior knowledge, we adopt the Bayesian framework, which allows the combination of probabilistic prior information with statistical data in a principled way.

The main research question in this paper is how to automatically construct an informative prior distribution over the space of Bayesian network *structures* from textual information. Previous approaches to define an informative prior distribution over Bayesian network structures can be grouped into *penalization* methods and *pairwise* methods. The *penalization* method defines the probability of a structure B_S based on the number of edges that are different from a prior network B_{Prior}, where a uniform probability is assumed for each missing or extra edge [25]. The *pairwise* methods define the probability of a structure B_S by combining the individual arc probabilities independently for each edge in B_S; they assume that being a parent of some node is independent from any other parental relation (as discussed in Section 7.2) and expect *a priori probabilities* for edges [10, 12]. In both approaches, the existence of informative edge priors is an important elementary component.

We follow here the pairwise approach to incorporate prior information into Bayesian network learning. The main goal is then to specify each such prior edge probability $P(X \rightarrow Y)$ using a text score $R_{\text{Text}}(X;Y)$ that expresses the relatedness between the two variables X and Y according to the domain literature. The rationale in the medical context is that (1) a significant body of medical research is devoted to the discovery of informational relationships between domain variables (mainly dependencies between variables are reported in the literature and independencies are largely ignored) and that (2) text-mining methods for relationship extraction have already proved useful in many domains as shown below. However, note that these earlier applications focused on providing results for the domain expert or data analyst, whereas our aim is to go one step further and use the results of these methods automatically in the statistical learning of *quantitative* models.

Beside the linguistic approaches, which ambitiously try to structurally analyze [36] and extract high level logical statements from free text [41, 16, 39], unstructured statistical approaches have similarly shown good performance in extraction of entity relationships. Two important types of unstructured methods are the methods based on *name cooccurrence* and the methods based on *kernel similarity.*

The methods based on *name cooccurrence* quantify the relatedness of two domain variables by the relative frequency of cooccurrence of their names (and possibly synonyms) in documents from a certain corpus. An early biomedical application of this approach was described by Swanson and Smalheiser [46]. A recent detailed evaluation of the cooccurrence of MeSH terms in MEDLINE abstracts against the manually curated UMLS Semantic Net similarly demonstrated the effectiveness of this approach [11]. In genomics, Stapley and Benoit [44] summarized the biological rationale for this simplified approach and performed a quantitative manual analysis for the model organism *Saccharomyces cerevisiae*, which indicated the usefulness of this approach for knowledge discovery in genomics. For human genes, Jensen et al. [28, 27] performed an extensive quantitative manual check of such pairwise scorings based on cooccurrence and concluded that the name cooccurrence in MEDLINE abstracts reflects biologically meaningful relationships with a practically acceptable reliability and that the main problem for the method in human genomics is synonym identification. In another experiment, the SUISEKI system [8] (which contains an advanced name detection subsystem) similarly achieved good performance using a frame-based approach to relationship extraction. For the use of occurrence and cooccurrence patterns to learn Bayesian network structures, de Campos et al. [17] used conditional mutual information scores defined on the occurrence patterns of words to learn a Bayesian network thesaurus from the literature.

The methods based on *kernel similarity* quantify the relatedness of two domain variables based on the vector representations of their textual descriptions (called kernels). Each component of this vector describes the weight of a certain word in the document and in the collection of documents (i.e., in the domain), which is derived from simple counting statistics or more elaborate weighting techniques. The relatedness of two variables can be based on either direct similarity or indirectly by the corelevance of their kernels to documents from a certain corpus (for the cognitive aspects of relevance, see [9]). Direct similarity means that domain variables (concepts) are related if their descriptions are similar while corelevance means that variables (concepts) are related if the same documents are similar to their descriptions. In a related approach, Reiner and Aszodi [40] defined the similarity measure of documents (annotations) by common concept presence. Shatkay et al. [42] proposed a probabilistic relevance measure between kernels and documents and they found good qualitative correspondence between the clustering of genes based on expression data and the clustering based on the simplified corelevance of the gene kernels.

We demonstrate the use of these text-based priors for Bayesian network learning on a medical domain related to clinical models of ovarian tumors and to the preoperative classification of ovarian tumors from patient history information, ultrasonography measurements, color Doppler measurements, and serum marker levels. This task serves as a well-understood evaluation domain for our methodology.

Figure 1 gives an overview of the different phases in using textual prior information to learn Bayesian networks. First, using the annotations of Bayesian network variables, we convert half a million MEDLINE abstracts to vectorial *literature data* ($\underline{\mathbf{P}}$), which indicates the presence (or relevance) of each variable in MEDLINE abstracts. Based on this *literature data*, we define cooccurrence and mutual information scores to quantify the textual dependencies between the variables. Additionally, we define another score based on the similarity of the kernel descriptions. Finally, we transform the automatically derived text scores into an informative prior for Bayesian network structures by normalizing and scaling it to satisfy certain higher-order statistical constraints and we apply this prior for the learning of Bayesian networks.

[Fig. 1 about here.]

To derive informative priors for Bayesian network structures from textual sources, we assume the existence of (1) annotations for the Bayesian network variables (which includes a textual name for the random variable, synonyms, a free text description (the kernel) and references to documents), (2) a collection of domain documents, and (3) domain vocabularies. Whereas the annotations of the Bayesian network variables could be seen only as an *ad hoc* solution to some engineering aspects of knowledge modeling, we exploit these text kernels to derive text-based relations $R_{\text{Text}}(X;Y)$ between the variables X and Y. Specifically, we introduce several pairwise dependency scores based on textual annotations, such as the pairwise cooccurrence of variable names in domain documents, the similarity of variable annotation, and the pairwise corelevance of variable descriptions to domain documents. These dependency scores are used in the derivation of formal informative priors for structure learning.

The enrichment of Bayesian networks with kernels for the variables (and implicitly with the text-based relations that can be derived from these kernels and the corpus) also fits in the trend of representing more and more contextual and qualitative information linked to the Bayesian network formalism. Leong [35] reported an extended network formalism enriching the Bayesian network with new types of relations between the variables related to qualitative Bayesian networks [50]. Koller and Pfeffer [29] and Laskey and Mahoney [34] suggested an object-oriented approach in which partial probabilistic knowledge fragments can be maintained in a modular way and can be reused for flexible Bayesian network construction within an evolving context. A further extension in this direction is the introduction of the probabilistic frame-based systems, which for example allows probabilistic relationships between the knowledge fragments [30].

In fact, annotations attached to a Bayesian network serve a broader purpose than the derivation of informative priors explored in this paper. Other applications include (1) linking data collection to data analysis using annotations from the data collection protocol in the Bayesian network [1], (2) supporting information retrieval by using the annotations of the random variables organized according to the Bayesian network structure in complex queries [6], or (3) supporting the knowledge engineering of complex Bayesian networks and explaining the Bayesian network structure and its inference [38, 32, 5]. Because of the multiple uses of annotations for Bayesian networks, we propose the term of *Annotated Bayesian Network* (ABN) to encompass the enhanced functionalities of textually enriched Bayesian networks. The ABN is defined as a *directed, acyclic graph whose nodes are annotated with local probabilistic models* (as in standard Bayesian networks) *and with textual attributes.* Furthermore, in the context of this paper where we derive informative priors, we informally require that the annotations and the document corpus be rich enough to give rise to meaningful scores $R_{\text{Text}}(X;Y)$ for X, Y pairs of random variables.

Admittedly, this working definition is simple and the growing complexity and amount of electronically available prior information will inevitably require a broader formalization to support the integrated application of (1) a priori textual information such as domain literature, domain vocabularies, taxonomies, ontologies, (2) a priori known qualitative domain relations and structural constraints for the Bayesian network and its parametric submodels, and even (3) textually annotated data sets. We believe that a full-scale principled use of such a wide scope of information resources can be achieved if these information resources are semantically organized around an "annotated Bayesian network" (in a broader sense than our current definition). Even though this formal, comprehensive, and semantic approach to information resources and ABNs remains a challenge, our work on deriving informative priors and other applications of annotations in Bayesian networks are systematic steps towards this goal. As we will demonstrate, even the definition of ABNs with text kernels already bridges the gap between the quantitative Bayesian networks and the textual resources.

The paper is organized as follows. Section 2 introduces the ovarian cancer domain, the variables and their relations, the textual resources, and the data set of clinical measurements. Section 3 introduces the text scores and Section 4 presents statistics of these text scores. Section 5 introduces the data scores. Section 6 reports on the correspondence between expert, text, and data scores. Section 7 presents methods for transforming the text scores into an a priori distribution over the Bayesian network structures. Finally, Section 8 reports the effect of these priors on Bayesian network performance for the classification of ovarian tumors. Sections 9 and 10 contain the discussion and conclusion.

2 Annotated Bayesian networks for ovarian cancer

We apply the techniques presented in this paper to clinical models of ovarian tumors incorporating patient history, ultrasonography measurements, color Doppler imaging measurements, and blood serum marker levels. This investigation takes place in the context of the International Ovarian Tumor Analysis Consortium[1] (IOTA), which is a multicenter study on ovarian tumors [49]. This study includes the multicenter collection of patient data and the corresponding data collection protocols.

2.1 Domain variables and edge priors

In the experiments we used a total of thirty-one variables (which had been previously evaluated as the most relevant domain variables), such as parity, drug treatment for infertility, use of oral contraceptives, family history of breast and ovarian cancer, age, bilaterality of the tumor, pain, descriptors of the morphology, echogenicity, and vascularization of the mass, or the level of CA125 tumor marker.

Furthermore, a leading expert in the ultrasonography of ovarian tumors selected the ‘most relevant’ relations between pairs of variables (57 relations), the ‘moderately relevant’ pairwise relations (56 relations) and the ‘weakly relevant’ pairwise relations (44 relations). S_3 denotes the set of ‘most relevant’ relations, S_2 denotes the set of both ‘most relevant’ and ‘moderately relevant’ relations, and S_1 denotes the set of all relations and $S^P_{1,2,3}$ denote the respective subsets of the relations corresponding to the central variable *Pathology*.

[1] http://www.iota-group.org

Furthermore, the expert provided rankings for the pairwise relations, which was manually transformed into a prior $R_{\mathrm{Expert}}(X;Y)$ for undirected edges as shown on Figure 2. We applied the scaling method described in Section 7.3 to the prior score to satisfy that the average pairwise direct relations per variable is 6, furthermore we set a lower limit ϵ to avoid the a priori exclusion of edges. For example, $R_{\mathrm{Expert}}(\mathrm{Pathology};Y)$ represents an assessment of the relevance of each domain variable Y with respect to the *Pathology* variable—that is, to discriminate between benign and malignant tumors.

[Fig. 2 about here.]

2.2 Text kernels

For the derivation of informative priors for Bayesian network structures from textual sources, we assume the existence of (1) annotations for the Bayesian network variables (which includes the name of the variable, synonyms, a free-text description (the kernel) and references to documents from the corpus). To ensure consistent, objective annotations and consequently the generality of our study and conclusions, we used the IOTA protocols without modification as primary sources. A corresponding Ph.D. thesis [48] provided an extension for the IOTA descriptions. Together, these compose the text kernels, on average a hundred-word description for each of the domain variables. Additionally, we let these kernels contain references to the Merck Manual[2], the On-line Medical Dictionary[3], the CancerNet Dictionary[4] and the MEDLINE collection of abstracts of the US National Library of Medicine[5].

2.3 Document collections

We asked medical experts to select the *most relevant* journals for the domain (2), the *highly relevant* (3), the *moderately relevant* (33) and the *relevant* journals (93). Based on these, we constructed four embedded collections of MEDLINE abstracts containing 5,367, 71,845, 231,582, and 378,082 abstracts denoted by C_3, C_2, C_1 and C_0 selected from the MEDLINE corpus dated between January 1982 and November 2000.

[2] http://www.merck.com/pubs/mmanual/

[3] http://www.graylab.ac.uk/omd/index.html

[4] http://thymoma.de/meddict.htm

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

2.4 Domain vocabularies

We constructed a domain vocabulary containing more than one million words ($D = 1,135,017$) by incorporating manually identified domain specific phrases and synonyms, statistically relevant words and manually curated general medical vocabularies, such as MeSH[6] . Furthermore, we constructed manually a smaller ($D = 700$), more specific vocabulary (which follows the IOTA terminology definitions and guidelines for controlled indices [49, 14]) for testing in this study (for further results, see [4]).

2.5 Data

In addition to the prior background information, data has been collected in the framework of the IOTA project [49], consisting of 68 parameters for 1,150 tumors at the moment of writing. In our experiments, we included the cases satisfying the IOTA protocol, excluded cases without measurement of the serum CA 125 level and use of oral contraceptives, which were not mandatory variables for the data collection but relevant to our goal. The data set contains the completely observed cases with respect to the selected variables (604 cases) denoted by $\underline{\mathbf{D}}$. Figure 3 shows the biplot of the data and the variables. The biplot plots variables and cases in the plane spanned by the first two principal components. In particular, a small angle between variables such as (Age, Meno, PostMenoY) points out that those variables are highly correlated. The observations of malignant tumors (indicated by $\diamondsuit$) tends to be correlated with high values for certain morphologic variables, such as Papillation or WallRegularity, but relatively low values for variables such as PillUse and Shadows.

[Fig. 3 about here.]

For the analysis we performed the following data transformation. Twenty of the variables are nominal or a nominal interpretation has been provided by the IOTA protocol. For the rest of the variables, a medical expert provided commonly used thresholds for their discretization.

3 Dependency scores based on annotations and domain literature

According to our assumption, a text kernel is available for each domain variable. The algebraic representation, called the *vector space model*, encodes a

[6] http://www.nlm.nih.gov/mesh/meshhome.html

document in a D-dimensional space where each component represents a corresponding word in the vocabulary described in Section 2.4. This approach thus neglects the grammatical structure of the text. We used the Porter stemmer to canonize the words [19], processed the essential domain specific phrases and synonyms appropriately and applied a standard stopword list to remove general words. The weights for the vector model were computed using the *term frequency–inverse document frequency* (tf-idf) term weighting schemel [7, 31]. The weighted frequency of term t_j in document d_i is

$$w_{ij}^{\text{tf-idf}} = f_{ij} \log(\frac{L}{n_i}) \tag{1}$$

where f_{ij} is the number of occurrences of t_j in d_i, L is the total number of documents and n_i is the number of documents containing term i in our largest MEDLINE corpus. We denote the presence of the name (and synonyms) of an ABN variable X_j in document d_i with a binary p_{ij}^N value. The vector $(p_{ij}^N)_{j=1,\ldots,N}$ is thus a binary vector of size N that describes document i by which of the domain variables (concepts) are present in this document. The vector $(p_{ij}^N)_{i=1,\ldots,L}$ is a binary vector of size L that describes the domain variable j by which of the documents from the corpus contain the name of this variable. $\underline{\mathbf{P}}^N$ denotes the complete matrix for a given document corpus. This matrix will be used in the name-cooccurrence methods. Note that this cooccurrence representation cannot handle repetition and proximity or separation into distinct paragraphs, sentence, and so on; but in our experiments this scheme gave satisfactory performance (for the comparison of such options, see [18]).

For the kernel methods, we need a measure expressing the similarity (relevance) of documents. The use of the vector representation of text was investigated intensively in the context of information retrieval. A principal goal in information retrieval is the definition of similarity metrics among the documents (or sets of documents), which express the semantic and information theoretic relation between the documents. A standard similarity metric for a pair of documents d_i, d_j is the cosine of the angle between their corresponding normalized tf-idf vector representation W_i, W_j:

$$\text{sim}(d_i, d_j) = \cos(W_i, W_j). \tag{2}$$

We define another binary representation of MEDLINE abstracts based on the kernel documents. It consists of binary variables defined as

$$p_{ij}^K = \begin{cases} 1 \text{ if } \tau < \operatorname{sim}(k_j, d_i) \\ 0 \ \text{ else} \end{cases}, \tag{3}$$

which expresses the relevance of kernel document k_j to document d_i. We will use an experimentally selected fixed value for τ (0.1) in this paper. $\underline{\mathbf{P}}^K$ denotes the complete matrix for a given corpus. This matrix will be used in the kernel-corelevance methods.

Next, we define two probabilistic models for the name cooccurrence and for the kernel corelevance for ABNs. Let $P(P_i^N = 1|\xi)$ represent the belief that the variable name X_i is reported in a random document from a given corpus (ξ describes the collection and other background conditions). Similarly, $P(P_i^N = 1|P_k^N = 1, \xi)$ represents the belief that the variable X_i is reported in a document given the presence of the name of X_k. Finally, $P(P_1^N, \dots, P_L^N|\xi)$ denotes the joint probability of presence of the names and synonyms of ABN random variables. For the kernel relevance, let $P(P_i^K = 1|\xi)$ represent the belief that a document from the corpus is relevant (in the sense defined in the previous paragraph) to the kernel document of variable X_i (ξ describes the threshold τ for relevance and other conditions). Similarly, $P(P_i^K = 1|P_k^K = 1, \xi)$ represents the belief that a document from a certain collection is relevant for the kernel document of variable X_i, if the kernel document of variable X_k is relevant. Finally, $P(P_1^K, \dots, P_L^K|\xi)$ denotes the joint probability of the relevance of the kernels of the random variables in the ABN for a certain document. Based on the previous definitions, we can define several text scores to quantify the dependency or correspondence of the pairs of random variables in the ABN. Let X and Y denote ABN variables. $R_{\text{COOC}}^{\text{AND},C_i}(X;Y)$ and $R_{\text{COREL}}^{\text{AND},C_i}(X;Y)$ denote a name-cooccurrence and a kernel corelevance score, $R_{\text{COOC}}^{\text{MI},C_i}(X;Y)$ and $R_{\text{COREL}}^{\text{MI},C_i}(X;Y)$ denote the mutual information scores based on name presence and kernel relevance over the collection C_i. Using the random variables P_i^N and P_i^K introduced above, the definitions are as follows (we denote the random variables for name presence and kernel relevance for the ABN variable X with P_X^N, P_X^K and their binary values with p_x):

$$R_{\text{COOC}}^{\text{AND}}(X;Y) \triangleq P(P_X^N = 1, P_Y^N = 1|(P_X^N = 1) \vee (P_Y^N = 1)) \tag{4}$$

$$R_{\text{COREL}}^{\text{AND}}(X;Y) \triangleq P(P_X^K = 1, P_Y^K = 1|(P_X^K = 1) \vee (P_Y^K = 1)) \tag{5}$$

$$R_{\text{COOC}}^{\text{MI}}(X;Y) \triangleq I(P_X^N; P_Y^N) = \sum_{p_x,p_y} P^N(p_x,p_y) \log(\frac{P^N(p_x,p_y)}{P^N(p_x)P^N(p_y)}) \tag{6}$$

$$R_{\text{COREL}}^{\text{MI}}(X;Y) \triangleq I(P_X^K; P_Y^K) = \sum_{p_x,p_y} P^K(p_x,p_y) \log(\frac{P^K(p_x,p_y)}{P^K(p_x)p^K(p_y)}) \tag{7}$$

These quantities are estimated from the frequencies, for example

$$R^{\text{AND}}_{\text{COOC}}(X_i;X_j) \approx \widehat{R}^{\text{AND}}_{\text{COOC}}(X_i;X_j) = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}, \tag{8}$$

where n_i and n_j are the number of documents in collection C_i containing the names of the ABN variables X_i and X_j, $n_{i,j}$ is the number of documents containing both of them. Additionally, we introduce a relevance scoring for X and Y inspired by information retrieval. We assume that the information need is defined by the kernel description of X (K_X) and the score expresses the relevance of the kernel descriptions of Y (K_Y). The definition is the following:

$$R_{\text{ASIM}}(X;Y) \triangleq \text{sim}(K_X, K_Y).$$

We refer to these text-based relevance scores in general with $R_{\text{Text}}(X;Y)$.

4 Descriptive statistics of literature scores

The usefulness of the R_{COOC} name-cooccurrence scores is essentially determined by the quality of the phrases and synonyms related to the names of the ABN variables [28]. Similarly, the kernel methods R_{COREL} and R_{ASIM} depend on the quality of the vector representation of the kernels. To check the quality of the vector representation of the ABN kernels defined by Equation 1, we verified the statistics of the (tf-idf) weights of the words corresponding to the smaller controlled vocabulary and verified the $R_{\text{ASIM}}(X;Y)$ relation using a hierarchical clustering with Ward linkage to create a clustering tree, as shown on the left of Figure 4. As another quality check, we used the same vector representation in an evaluation of an information retrieval language based on ABNs and the quantitative evaluation has indicated good performance [6].

[Fig. 4 about here.]

Table 1 shows the percentages of abstracts where a certain variable name occurs and the percentages of abstracts that are closer to the kernel of domain variables than the specified threshold in the *relevant* (largest) and the *most relevant* (smallest) MEDLINE corpora. Furthermore, Figure 5 shows the percentages of abstracts with $0, 1, 2, \ldots$ name occurrences and the percentages of abstracts that are close to $0, 1, 2, \ldots$ kernels of domain variables with respect to the specified threshold for the *relevant* (largest) MEDLINE corpus.

[Table 1 about here.]

[Fig. 5 about here.]

5 Dependency scores based on data

Based on the data, we can introduce similar data scores $R_{\text{Data}}(X;Y)$ to quantify the informational relevance of X, Y. Under the assumptions that the stochastic variables are discrete and the cases in the data set are complete, a natural choice is to use the conditional mutual information estimated from the data $\underline{\mathbf{D}}$, similarly to what we defined for the literature data $\underline{\mathbf{P}}^{\mathbf{N}}$ and $\underline{\mathbf{P}}^{\mathbf{K}}$:

$$R^{\text{MI}}_{\text{Data}}(X;Y) \triangleq I(X;Y) = \sum_{x_i,y_j} p(x_i,y_j)\log(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}) \tag{9}$$

For an information theoretic approach to learning Bayesian network structures using conditional mutual information, see [13] and for learning Bayesian network thesaurus from textual data, see [17].

A Bayesian approach to score the parental relations $\pi_i \to X_i$ in a Bayesian network was proposed by [15] (π_i denotes the parental set for X_i). Under the assumptions that (1) the stochastic variables are discrete and (2) the cases in the data are complete and exchangeable and (3) the prior over the parametrization is Dirichlet, a closed-form solution was derived for the probability of parental relations $P(\pi_i \to X_i|\underline{D},\xi)$. Assuming the BDeu parameter prior, which is a uniform prior resulting likelihood equivalent Bayesian metrics [10, 25], a symmetric, pairwise data score $R^{\text{BD}}_{\text{Data}}(X;Y)$ can be defined, which expresses the probabilities of the individual pairwise structures $P(Y \to X|\underline{D},\xi)$:

$$R^{\text{BD}}_{\text{Data}}(X;Y) \propto \prod_{j=1}^{r_Y}\prod_{k=1}^{r_X}\Gamma(N^{YX}_{jk} + \frac{1}{r_X + r_Y}). \tag{10}$$

Here r_X, r_Y denotes the number of discrete values of variables X and Y and N^{YX}_{jk} denotes the number of times we observe value j for variable Y and value k for variable X in the data $\underline{\mathbf{D}}$.

6 Correspondence of expert priors, text scores and data scores

To investigate the use of integrated text and data scores for learning Bayesian networks, we compared the text scores introduced in Section 3 against (1) the expert score $R_{\text{Expert}}(X;Y)$ and (2) the data scores. The correspondence is illustrated on the right of Figure 4, the domain variables are positioned on the coordinates $(R_{\mathrm{Expert}}(\mathrm{Pathology}; X_i), R^{\mathrm{MI},C_3}_{\mathrm{COREL}}(\mathrm{Pathology}, X_i))$ (marked by 'o') and $(R_{\mathrm{Expert}}(\mathrm{Pathology}; X_i), R^{\mathrm{MI}}_{\mathrm{Data}}(\mathrm{Pathology}, X_i))$ (marked by 'x') to illustrate their relation.

To quantitatively evaluate the different text scores and understand their relations, we computed (1) the Area Under the ROC curve (AUC) [23] to detect the relevant relations identified by the expert and (2) the Spearman rank correlation coefficient R_S with the expert score and with the data scores. The first column of Table 2 shows the AUC values for detecting the S_3, S_2, and S_1 relations. The second column of Table 2 shows the specificity values for detecting these sets corresponding to 50% sensitivity. The third column of Table 2 shows the sensitivity values for detecting these sets corresponding to 50% specificity. The upper triangle of Table 3 presents the Spearman rank correlation coefficients for all pairs of the expert score, text scores and data scores as introduced in Sections 2.1, 3, and 5. Beside the Spearman rank correlation coefficients for all the relations, the lower triangle of Table 3 shows the Spearman rank correlation coefficients for the relations of the variable *Pathology*. Bold indicates significant monotonic relationship between the ranks with $p < 0.05$, underscore indicates it with $p < 0.001$, and bold underscore with $p \ll 0.001$.

[Table 2 about here.]

[Table 3 about here.]

7 Transforming text scores to priors for BN structures

The evaluation reported in Sections 4 and 6 indicates the potential for the integration of text scores in a Bayesian statistical learning of Bayesian network structures (another study on the correspondence of a wider range of text scores can be found in [4]). We investigate the incorporation of the text scores $R_{\mathrm{Text}}(X;Y)$ in the Bayesian framework (i.e., to combine the text scores with $R^{\mathrm{BD}}_{Data}(X;Y)$) because it provides a more flexible and still principled foundation than the information theoretic approach and its corresponding score $R^{\mathrm{MI}}_{\mathrm{Data}}(X;Y)$.

7.1 Learning of BN structures using prior information and data

As we already mentioned in Section 5, under reasonable assumptions, a closed-form Bayesian formula was derived for the probability of a Bayesian network structure B_S given a data set $\underline{D}$ (for details, see [15]):

$$P(B_S|\underline{D},\xi) \propto P(B_S|\xi) \prod_{i=1}^{n} R_{\text{Data}}^{\text{BD}}(X_i, \pi_i). \tag{11}$$

Assuming the independence of beliefs for substructures $(\pi_i \rightarrow\ X_i)$ and $(\pi_j \rightarrow\ X_j)$ for all $i \neq j$, the prior $P(B_S)$ can be decomposed [10, 15] as

$$P(B_S|\underline{D},\xi) \propto \prod_{i=1}^{n} P(\pi_i^S \rightarrow\ X_i|\xi) R_{\text{Data}}^{\text{BD}}(X_i, \pi_i). \tag{12}$$

Note that with this assumption on the prior $P(B_S|\xi)$, the probability of a Bayesian network structure given a complete data set decomposes into a product of independent parts $P(\pi_i^S \rightarrow\ X_i|\xi)R_{\text{Data}}^{\text{BD}}(X_i, \pi_i)$, each expressing the probability of the local dependency model of variable X_i with parents π_i conditioned on the data. Despite the decomposition of the learning to the selection of appropriate parental sets, the amount of data needed for statistically significant identification of networks is still frequently insufficient. One potential solution is to define an informative a priori distribution $P(B_S|\xi)$ or $P(\pi_i^S \rightarrow\ X_i|\xi)$ for each variable. However, even the later task can be difficult for human experts or even for automatic methods, so the *pairwise* methods [10, 12] and the *penalty* methods were suggested [25].

7.2 BN structure priors from edge probabilities

The *pairwise* methods define the probability of a structure B_S by combining the individual arc probabilities independently. As we assume the independence of $(\pi_i \rightarrow\ X_i)$ and $(\pi_j \rightarrow\ X_j)$, our goal is to derive an estimate for $P(\pi_i \rightarrow\ X_i|\xi)$ based on the introduced text scores $R_{\text{Text}}(X_;\pi_i)$. Furthermore, we assume the independence of $X_k \in \pi_i$ and $X_j \in \pi_i$ (i.e., the independence of presence of the edges in the graph at a vertex), which allows the decomposition of $P(\pi_i \rightarrow\ X_i|\xi)$, where $\pi_i = \{\pi_{i,1}, \ldots, \pi_{i,L}\}$:

$$P(\pi_i \rightarrow\ X_i|\xi) = \prod_{k=1}^{L_i} P(\pi_{ik} \rightarrow\ X_i|\xi) \prod_{Y \notin \pi_i} (1 - P(Y \rightarrow\ X_i|\xi)).$$

Previous methods expected a priori probabilities for edges corresponding to a fixed ordering of the variables [10] or for directed edges [12]. However, we use the symmetric text scores for the derivation of the prior, which means that an edge is scored independently of its direction (for a possible approach to determine directionality in such case, see [17]). Therefore $P(\pi_{ik} \rightarrow\ X_i|\xi)$

(using p_{ij} as a shorthand notation assuming $\pi_{ik} = X_j$) can be defined by the pairwise text scores:

$$p_{ij} \triangleq P(\pi_{ik} \rightarrow \ X_i|\xi) \sim R_{\text{Text}}(X_i, \pi_{ik}). \tag{13}$$

Note that for all text scores $0 \leq R_{\text{Text}}(X_i, \pi_{ik}) \leq 1$ and that we guarantee a lower limit ϵ and an upper limit $1 - \epsilon$ for all p_{ij} to avoid the a priori exclusion or inclusion of edges and consequently structures. This relative definition of edge probabilities are refined subsequently to satisfy prior knowledge on higher-order statistics. Note also that because the text-based priors are symmetric, the *prior equivalence* holds (i.e., the structures encoding the same informational relevance model have the same prior) [25].

7.3 Scaling the edge probabilities

The a priori distribution on the edges defined by Equation 13 is relative, so an appropriate scaling can be achieved by noting that the expectation of the number of edges L is given by $\sum_{0<i<j<n} p_{ij}$. Assuming that there is an a priori estimate for the number of edges in the overall model or connected to a single variable, the p_{ij} can be scaled by an exponent ν to approximate this edge density in the prior Bayesian network. By denoting the value that scales the expectation of the number of parental edges to L_0 with $\nu(L_0)$ we define the following scaling (it is always possible because of the lower limit $\epsilon < p_{ij}$):

$$q_{ij} \triangleq p_{ij}^{\nu(L_0)} \text{ with } \nu(L_0) \text{ so that } \sum_{0<i<j<n} q_{ij} = L_0. \tag{14}$$

For a given ordering of variables, these rescaled edge probabilities define the following posterior probability for a structure B_S:

$$P(B_S|\underline{D}, \xi) \propto \prod_{i=1}^{n} \prod_{j=1}^{i-1} q_{ij}^{I_{\{j \in \pi_i\}}} (1 - q_{ij})^{I_{\{j \notin \pi_i\}}} R_{\text{Data}}^{\text{BD}}(X_i, \pi_i). \tag{15}$$

7.4 Combination of text-based edge probabilities with a prior structure

If a prior network structure is available, there is an easy method to combine it with the introduced informative edge probabilities. The penalization method [25] derives the prior from an a priori network structure B_0 by modeling each missing or extra edge e_{ij} independently with a uniform probability κ:

$$P(B_s|\xi) \propto \kappa^{\delta}, \text{where } \delta = \sum_{1\leq i<j\leq n} I_{\{(e_{ij}\in B_S)\wedge(e_{ij}\notin B_0)\vee(e_{ij}\notin B_S)\wedge(e_{ij}\in B_0)\}}.$$

This formula can be developed further into an informative penalization formula by replacing the uniform κ with an edge-specific pairwise prior based on the text scores $R_{\text{Text}}(X;Y)$:

$$P(B_S|B_0, p_{ij}, \nu, \xi) \propto \prod_{1\leq i<j\leq n} q_{ij}^{I_{\{(e_{ij}\in B_S)\wedge(e_{ij}\notin B_0))\}}}(1-q_{ij})^{I_{\{(e_{ij}\notin B_S)\wedge(e_{ij}\in B_0)\}}}$$

Note that the scaling of p_{ij} provides an option to control the penalization (i.e., to express the prior beliefs in the prior structure and in the literature).

8 Learning and classification using text-based priors

To evaluate the value of the text-based prior distributions for Bayesian network structures, we report results for the classification of ovarian tumors, as described in Section 2. (For previous results about the application of Bayesian belief networks and multilayer perceptrons to classify ovarian tumors, see [3].) We report the classification performance of a Bayesian belief network using the Area Under the Receiver Operating Curve (AUC). Since we work in the Bayesian framework, we have a posterior distribution $P(B_S|\underline{D})$ over the network structures and a conditional posterior $P(B_P|B_S, \underline{D})$ over its parameters, resulting in a posterior distribution of the AUC. We report the mean of this AUC using either an informative text-based prior, the expert prior, or a noninformative uniform prior over the structure space:

$$E[\text{AUC}_{B_S,B_P}(\underline{D}_{\text{te}})|\underline{D}_{\text{tr}}] = \sum_{B_S} P(b_S|\underline{D}_{tr}) \int_{B_P} \text{AUC}_{b_S,b_P}(\underline{D}_{\text{te}}) dP(b_P|b_S, \underline{D}_{\text{tr}}).$$

where $\underline{D}_{tr}$ and $\underline{D}_{te}$ denote the training and test data. Because we want to focus on the usage of textual prior knowledge for learning belief network *structures*, we used always the noninformative Bayesian Dirichlet prior BD_{eu} for the parameters [25]:

We approximate the summation over the network structures with a Monte-Carlo approximation using 200 networks with a high posterior probability. We evaluate 200 randomly drawn orderings for the variables. Using a set of complete and discrete samples, we learn a Bayesian network structure by maximizing the R^{BD}_{Data} score for each variable for the given ordering [15]. For each fixed ordering, the parents are selected using an exhaustive search up to three parents. If this exhaustive search finds three parents, the greedy (not exhaustive) K2 algorithm continues the search [15]. The probabilities for the belief network substructures are computed using both the training data and the edge probabilities according to Equation 15. The edge probabilities derived from the text were scaled with a ν value that results in prior networks with 3 parents for each node on average.

For these learned structures, the parameters are set to the maximum a posteriori value using the noninformative BD$_{eu}$ prior for the parameters and the training set $\underline{D}_{\mathrm{tr}}$. Predictions for the test samples are generated using the probability propagation in tree of cliques (PPTC) algorithm and these predictions are used to compute the AUC value on the test set $\underline{D}_{\mathrm{te}}$. The AUC values reported in Figure 6 are the averages over 300 cross-validation sessions with random training–test partitioning of the data set $\underline{D}$ into $(\underline{D}_{\mathrm{tr}}, \underline{D}_{\mathrm{te}})$. The x axis indicates the number of samples in the training set, ranging up to 150 samples (out of a total of 604), the y axis contains the AUC averages for that specific training–test proportion.

The upper part of Figure 6 reports the learning curves for the cooccurrence- and corelevance-based text priors ($R^{\mathrm{AND},C_0}_{\mathrm{COOC}}$ and $R^{\mathrm{AND},C_3}_{\mathrm{COREL}}$), together with the kernel similarity prior R_{ASIM}, the expert prior R_{Expert}, and no prior, all scaled by $\nu(3)$. The bottom part shows the effect of scaling the best performing prior based on the kernel similarity score R_{ASIM} by $\nu(0.1)$, $\nu(0.5)$, $\nu(1)$, $\nu(2)$, and $\nu(3)$. The noninformative case is again reported for comparison.

[Fig. 6 about here.]

9 Discussion

The main goal of our analysis is to understand the characteristics and usability of the text scores in learning Bayesian networks. First we compare the constructed text scores against the expert score R_{Expert} and the data scores $R^{\mathrm{BD}}_{\mathrm{Data}}$ and $R^{\mathrm{MI}}_{\mathrm{Data}}$. In the comparison, we performed a manual analysis as illustrated on Figure 4 and applied two quantitative evaluation methods: the efficiency of detecting the pairwise relations from the expert and the Spearman rank correlation. Next we evaluated the effect of the text-based prior for learning Bayesian networks using the AUC in classifying ovarian tumors.

In general, we can characterize the R_{Expert} as an expert reference, the annotation similarity R_{ASIM} as a kind of textual expression of expert belief, the cooccurrence relation R_{COOC} as an unbiased literature relation, the corelevance relation R_{COREL} as a mixture of expert belief and literature, and finally the data scores R_{Data} as objective references.

The relation detection means that we try to find back a set of important relations (specified by the medical expert) using a score for these pairwise relations and some threshold. (We use the sets S_1 and S_3 as defined in Section 2.1, S_2 is omitted for simplicity, S_1^P and S_3^P are the respective subsets containing only the relations corresponding to the variable *Pathology*.)

First, using the AUC values and sensitivity-specificity values from Table 2, we examine which of the text prior or the data can select better the relations from the expert. We expect the domain to be known enough, thus we expect the highest correspondence between the prior and the data scores (we expect the text scores to be less accurate due to the noise and bias). Surprisingly, the text scores performed better than expected; for example the $R_{\mathrm{COREL}}^{MI,C3}$ achieved an AUC value of 82.01 and $R_{\mathrm{Data}}^{\mathrm{MI}}$ achieves AUC=85.95 for selecting the S_1 relations. Although the data scores are slightly better, the differences are not statistically significant. The opposite behavior of S_1^P is investigated below. Another unexpected result is that $R_{\mathrm{COREL}}^{MI,C3}$ outperforms the R_{ASIM} relation (AUC=65.83), although the corelevance methods is a mixture of experts belief and literature, while the annotation similarity is closer to the expert belief.

Second, we examine the effect of increasing the size of the document collection from C_3 to C_0, which basically means a broader scope with less domain specificity and thus a higher noise level. As Table 2 shows, the (name) cooccurrence-based scores perform better on a larger collection—that is, they gain more from the larger number of publications than they lose from the fact that the documents are less domain specific (e.g., AUC=61.61 for C_0 versus AUC=64.95 for C_3 of the $R_{\mathrm{COOC}}^{\mathrm{MI}}$ for the set S_1). This is probably caused by the scarcity of names (i.e., the lack of a nomenclature), which can be seen from Table 1 that presents the name occurrence patterns for various collections, for full abstracts and for titles only (for example, non-established names such as 'Papillation flow' are very rare). In reverse, the corelevance methods perform better on the smaller, more specific collection C_3 (e.g., AUC=75.17 for $R_{\mathrm{COREL}}^{\mathrm{MI},C_0}$ versus AUC=82.01 for $R_{\mathrm{COREL}}^{\mathrm{MI},C_3}$ for the S_1 set). It means that the vector representation and the applied relevance measure cannot cope with the broader scope of the corpus, while still much better than the simpler cooccurrence methods.

Third, we examine the effect of detecting the 'most relevant' relations S_3 and all the relevant relations S_1. As we expected, the 'most relevant' relations are more easy to identify for all the text scores and data scores in the case of S_1 versus S_3. It also holds for the data scores, which means that the on average the expert score is in close correspondence with what the data says.

Interestingly, this trend is mixed in the case of the relations including variable *Pathology* (S_1^P versus S_3^P), in which the data scores are less effective to select the most relevant variables than a broader scope of related variables. A preliminary evaluation has shown that the expert ranking of certain factors as 'most relevant' and 'moderately relevant' is responsible for this, for example the top-rated papillation related variables were rated lower by the data (see Figure 4). Furthermore, the cooccurrence scores R_{COOC}, which can be seen as objective literature scores beside the objective data scores, are similarly less effective to select the most relevant variables than to select the broadest scope of variables. Note that this is not the case for the annotation-based score R_{ASIM}, which reflects the expert's textual ranking. However, in a detailed analysis of the ranking of the expert, data and literature, the limitations of the pairwise approach should be taken into consideration also, because the variables are strongly dependent that made difficult for the expert to select pairwise relations.

Finally, we examined the effect of using the mutual information (MI) and the cooccurrence (AND) formulas. Because the name cooccurrence method in our domain is prone to generating extreme relations (i.e., with uncommon variable names that never occur), the corelevance method is more appropriate for this investigation, but as Table 2 illustrates we could not find a significant difference or qualitative difference along this dimension.

The other quantitative method for the comparison of the scores is the comparison of the correspondence of their ranking by the Spearman ranking coefficient R_S (note that the scaling of the scores defined in Section 7.3 is monotonic, so does not influence ranking). Table 3 presents all the cross comparisons both for all of the relations and for only the *Pathology* relations (the AND options are not shown for simplicity, because they are not different from the MI case). We investigated the following hypotheses (by indicating the strength of a rank correlation in increasing order with $\sim$, $\approx$, $\simeq$, and $\cong$):

(1) *Quality of expert prior.* The expert score R_{Expert} strongly rank correlates with the data scores $R_{\mathrm{Data}}^{\mathrm{BD}}$ and $R_{\mathrm{Data}}^{\mathrm{MI}}$:

$$R_{\mathrm{Data}}^{\mathrm{BD}} \cong R_{\mathrm{Data}}^{\mathrm{MI}} \text{ and } R_{\mathrm{Data}} \simeq R_{\mathrm{Expert}}. \tag{16}$$

(2) *Quality of text-based priors.* The text scores R_{Text} strongly rank correlate with the expert score R_{Expert} and somewhat weakly with the data scores R_{Data}:

$$R_{\mathrm{Text}} \simeq R_{\mathrm{Expert}} \text{ and } R_{\mathrm{Text}} \approx R_{\mathrm{Data}}. \tag{17}$$

(3) *Subjectivity of text scores.* The annotation-based score R_{ASIM} is the most subjective (i.e., closest to the expert prior R_{Expert}) and R_{COOC} is the most

objective:

$$R_{\mathrm{ASIM}} \simeq R_{\mathrm{Expert}},\ R_{\mathrm{COREL}} \approx R_{\mathrm{Expert}}\ ,\text{and } R_{\mathrm{COOC}} \sim R_{\mathrm{Expert}}. \qquad (18)$$

In other words, the hybrid corelevance method R_{COREL} is between the expert (subjective) R_{ASIM} and the literature (objective) R_{COOC}:

$$R_{\mathrm{COOC}} \sim R_{\mathrm{ASIM}},\ R_{\mathrm{COOC}} \approx R_{\mathrm{COREL}} \text{ and } R_{\mathrm{ASIM}} \approx R_{\mathrm{COREL}}. \qquad (19)$$

From Table 3, we can conclude that the expert score R_{Expert} is significantly, strongly rank correlated with the data scores, so its reference status is corroborated (see Equation 16). Similarly, the text scores, more specifically the corelevance R_{COREL} and the annotations similarity R_{ASIM} are significantly, strongly rank correlated with the prior and somewhat weakly with the data scores (see Equation 17). Furthermore, the corelevance R_{COREL} and the annotation similarity R_{ASIM} are really better rank correlated with the expert score than with the ‘objective’ literature-based cooccurrence score (see Equation 18). However, contrary to our expectations, the corelevance relation R_{COREL} outperforms the annotation similarity R_{ASIM} (see Equation 18), which indicates that the annotations does not reflect completely the expert prior and can be refined in this respect using the literature by the corelevance method. Finally, the R_{COREL} score is strongly rank correlated with R_{ASIM} but not with R_{COOC}, and similarly R_{ASIM} is not rank correlated with R_{COOC} (see Equation 19).

In the classification task, the automatically constructed text-based prior for Bayesian network structures is beneficial in the small sample range, while it is not restrictive and vanishes in the middle and large sample range, that is it provides advantages comparable to those of a manually constructed expert prior. Note that the advantage of the prior over the structures can be fully exploited and they are better comparable if a prior is available for the parametrization, which is transformed appropriately [25].

10 Conclusion

As more and more domain knowledge becomes available as electronic literature, knowledge engineering and machine learning increasingly need representations and methods that integrate textual domain knowledge directly into knowledge modeling to assist the elicitation, learning, application, and maintenance of complex models. The *Annotated Belief Network*, which links textual information to the statistical model, is a response to this challenge. The ABN naturally supports the definition of various text scores to quantify the dependencies of the variables. On the one hand it provides an efficient overview of the textual information in the ABN and in the domain itself; on the other hand it can automate the knowledge acquisition for machine learning.

We presented a new application of methods for the extraction of relationship among domain variables from text to support Bayesian network structure learning, beside performing an analysis of the extracted relationships. From a medical point of view, the text representation and the text scores performed well in the manual qualitative analysis as illustrated in Figure 4 and in the quantitative evaluation when the text scores are used to detect pairwise relations rated as relevant by an expert (Table 2). Also the text scores proved to be significantly rank-correlated with the expert ratings and with the data scores (Table 3). These results indicate that the performance of the name-cooccurrence methods can be further improved by using kernel descriptions, which can be essential in domains without established nomenclature.

Overall, the positive results in the ovarian cancer domain, which we used as a relatively simple and well-known test domain, demonstrate that the automatically derived text scores can support the learning of informational relevance models, such as the structure of the Bayesian network. Indeed, the automatically derived text-based priors for the network structures improve the classification performance of the learned Bayesian networks comparably to a manually constructed expert prior.

However, our approach still has limitations. The first is the use of unstructured text representations, whereas the annotations are often already structured into various fields. A more refined linguistic analysis similarly could improve the text scores and the text priors also. Another limitation of the presented approach is its pairwise nature, so we are investigating the derivation of multiparental text scores and probabilities $P(\pi_i \rightarrow X_i|\xi)$ as well as the extraction of general irrelevance statements ($I\langle X|Z|Y\rangle$, see [37, 21, 45]) as a richer prior for the learning algorithms of Bayesian network structures. We are also investigating prior derivation methods based on the combination of the expert score and the text scores, as introduced in Section 7.4. Another research direction we are pursuing is the detailed medical evaluation of the data scores against the expert prior and the literature scores.

Despite these simplifications and constraints, the work presented here contributes to the recent efforts to better integrate data, electronic domain literature, and human expertise in medicine and genomics.

Acknowledgements

Bart De Moor is a full professor at the K.U. Leuven, Belgium. Dirk Timmerman is an associate professor at the University Hospitals K.U. Leuven. Yves Moreau is an assistant professor at the K.U. Leuven and a postdoctoral researcher with the Belgian Fund for Scientific Research (F.W.O. - Vlaanderen). Peter Antal is a Research Assistant with the K.U.Leuven. Geert Fannes is a Research Assistant with the F.W.O. - Vlaanderen. Our research is supported by grants from: Research Council KUL: Concerted Research Action GOA-Mefisto 666, IDO (IOTA Oncology, Genetic networks); Flemish Government: Fund for

Scientific Research Flanders (projects G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0407.02, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s , STWW-Genprom, GBOU-McKnow , Eureka-Impact, Eureka-FLiTE); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP-TR-18)

References

[1] S. Aerts, P. Antal, B. De Moor, and Y. Moreau. Web-based data collection for ovarian cancer: a case study. In *Proc. of the 15th IEEE Symp. on Computer-Based Medical Systems (CBMS02)*, pages 282–287, 2002. Maribor, Slovenia.

[2] R. B. Altman and T. E. Klein. Challenges for biomedical informatics and pharmacogenics. *Annual Review, Pharmacological Toxicology*, 42:113–133, 2002.

[3] P. Antal, G. Fannes, Y. Moreau, B. De Moor, J Vandewalle, and D. Timmerman. Extended Bayesian regression models: a symbiotic application of belief networks and multilayer perceptrons for the classification of ovarian tumors. In *Lecture Notes in Artificial Intelligence (AIME 2001)*, pages 177–187, 2001. Cascais, Portugal.

[4] 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-KDD02*, pages 405–414, 2002. Edmonton, Canada.

[5] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Annotated Bayesian networks: a tool to integrate textual and probabilistic medical knowledge. In *Proc. of the 13th IEEE Symp. on Comp.-Based Med.Sys. (CBMS01)*, pages 177–182, 2001. Bethesda, Maryland.

[6] P. Antal, T. Meszaros, B. De Moor, and T. Dobrowiecki. Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain. In *Proc. of the 15th IEEE Symp. on Computer-Based Medical Systems (CBMS02)*, pages 213–218, 2002. Maribor, Slovenia.

[7] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.

[8] C. Blaschke and A. Valencia. The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems*, 17(2):14–20, 2002.

[9] T. A. Brooks. The semantic distance model of relevance assessment. In *Proc. of the 61st Annual Meeting of ASIS, Pittsburgh, PA, USA*, pages 33–44, 1998.

[10] W. L. Buntine. Theory refinement of Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence (UAI-*

91), pages 52–60, 1991.

[11] A. Burgun and O. Bodenreider. Methods for exploring the semantics of the relationships between co-occurring umls concepts. In *Proc of the Tenth World Congress on Health and Medical Informatics (MedInfo 2001)*, pages 171–175, 2001.

[12] R. Castelo and A. Siebes. Priors on network structures. biasing the search for Bayesian networks. *International Journal of Approximate Reasoning*, 24(1):39–57, 2000.

[13] J. Cheng, D. A. Bell, and W. Liu. Learning belief networks from data: an information theory based approach. In *Proc. of the 6th ACM International Conference on Information and Knowledge Management, CIKM'97*, pages 325–331, 1997.

[14] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5):394–403, 1998.

[15] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[16] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118(1-2):69–113, 2000.

[17] L. M. de Campos, J. M. Fernández, and J. F. Huete. Query expansion in information retrieval systems using a Bayesian network-based thesaurus. In *Proc. of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98), Wisconsin, USA*, pages 53–60, 1998.

[18] J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. Mining medline: Abstracts, sentences, or phrases? In *Proc. of Pacific Symposium on Biocomputing (PSB 2002)*, pages 326–337, 2002.

[19] W. B. Frakes. *Information retrieval: Data Structures and Algorithms*, chapter Stemming Algorithms. Prentice Hall, 1992.

[20] N. Friedman and Z. Yakhini. On the sample complexity of learning Bayesian networks. In *Proc. of the 12th Uncertainty in Artificial Intelligence Conference (UAI96), Portland, Oregon, USA*, pages 274–282, 1996.

[21] David Galles and Judea Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

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

[23] D. J. Hand. *Construction and Assessment of Classification Rules.* Wiley & Sons, 1997.

[24] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, 17(2):37–43, 2002.

[25] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[26] 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*, volume 58(2-3), pages 384–395, 2001.

[27] T. Jenssen, L. M. J. Oberg, M. L. Anderson, and J. Komorowski. Methods for large-scale mining of networks of human genes. In *Proc. of 1st SIAM International Conference on Data Mining*, 2001.

[28] 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*, 28:21–28, may 2001.

[29] D. Koller and A. Pfeffer. Object-oriented Bayesian networks. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97), Providence, Rhode Island, USA*, pages 302–313, 1997.

[30] D. Koller and A. Pfeffer. Probabilistic frame-based systems. In *Proc. of the 15th National Conference on Artificial Intelligence (AAAI), Madison, Wisconsin*, pages 580–587, 1998.

[31] R. Korfhage. *Information Storage and Retrieval.* Wiley, 1997.

[32] C. Lacave and F. J. Diez. A review of explanation methods for Bayesian networks. Technical Report Tech. Report IA-2000-01, Dept. Int. Art. UNED, Madrid, 2000.

[33] W. Lam and F. Bacchus. Using causal information and local measures to learn Bayesian networks. In *Proc. of the 9th Conference on Uncertainty in Artificial Intelligence (UAI-93)*, pages 243–250, 1993.

[34] K. Laskey and S. Mahoney. Network fragments: Representing knowledge for constructing probabilistic models. In *Proc. of the 13th Conference on Uncertainty in Artificial Intelligence (UAI-97), Providence, Rhode Island, USA*, pages 334–341, 1997.

[35] T. Y. Leong. Representing context-sensitive knowledge in a network formalism: A preliminary report. In *Proc. of the 8th Conference on Uncertainty in Artificial Intelligence (UAI-92), Stanford, California, USA*, pages 166–173, 1992.

[36] J. C. Park. Using combinatory categorial grammar to extract biomedical information. *IEEE Intelligent Systems*, 16(6):62–67, 2001.

[37] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

[38] M. Pradhan, G. Provan, B. Middleton, and M. Henrion. Knowledge engineering for large belief networks. In *Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence*, pages 484–490, 1994.

[39] D. Proux, F. Rechenmann, and L. Julliard. A pragmatic information extraction strategy for gathering data on genetic interactions. In *Proc. of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'2000), LaJolla, California*, pages 279–285, 2000.

[40] A. Renner and A. Aszodi. High-throughput functional annotation of novel gene products using document clustering. In *Proc. of Pacific Symposium on Biocomputing (PSB00)*, volume 5, pages 54–65, 2000.

[41] T. C. Rindflesch, L. Tanabe, and J. N. Weinstein. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proc. of Pacific Symposium on Biocomputing (PSB00)*, volume 5, pages 514–525, 2000.

[42] H. Shatkay, S. Edwards, and M. Boguski. Information retrieval meets gene analysis. *IEEE Intelligent Systems*, 17(2):45–53, 2002.

[43] S. Srinivas, S. Russell, and A. Agogino. Automated construction of sparse Bayesian networks for unstructured probabilistic models and domain information. In *Proc. of the 5th Conference on Uncertainty in Artificial Intelligence (UAI-1990)*, pages 295–308. North-Holland, 1990.

[44] B. Stapley and G. Benoit. Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline asbtracts. In *Proc. of Pacific Symposium on Biocomputing (PSB00)*, volume 5, pages 529–540, 2000.

[45] M. Studeny. Semigraphoids and structures of probabilistic conditional independence. *Annals of Mathematics and Artificial Intelligence*, 21(1):71–98, 1997.

[46] D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183–203, 1997.

[47] A. Tanay and R. Shamir. Computational expansion of genetic networks. *Proc. of International Conference on Intelligent Systems for Molecular Biology (ISMB'01)*, 17(Suppl. 1):270–278, 2001.

[48] D. Timmerman. *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology*. Ph.D. dissertation, Leuven University Press, D/1997/1869/70, 1997.

[49] D. Timmerman, L. Valentin, T. H. Bourne, W. P. Collins, H. Verrelst, and I. Vergote. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. *Ultrasound Obstetrics Gynecology*, 16(5):500–505, Oct 2000.

[50] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44:257–303, 1990.

List of Figures

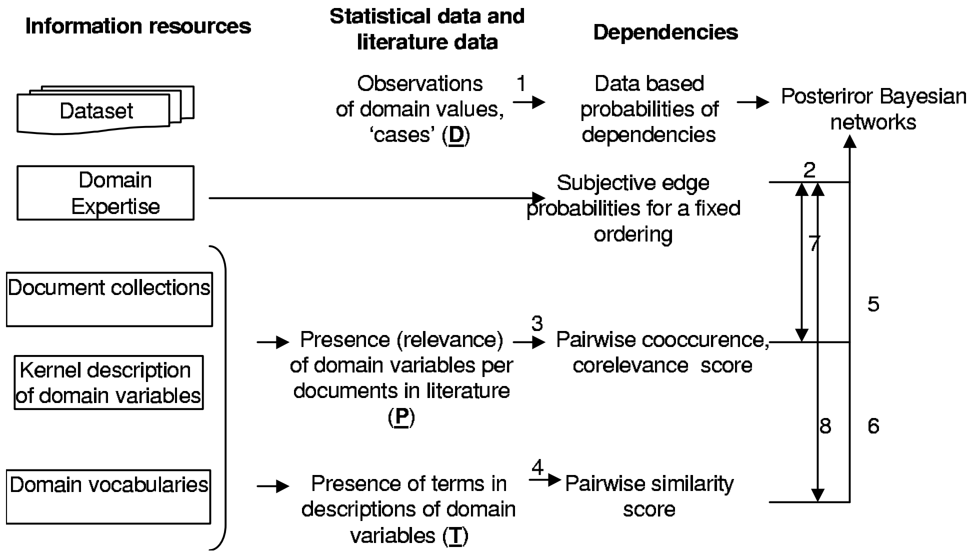
Information resources
Statistical data and literature data
Dependencies
Dataset
Observations of domain values, 'cases' (D)
1
Data based probabilities of dependencies
Posteriror Bayesian networks
Domain Expertise
Subjective edge probabilities for a fixed ordering
2
7
5
Document collections
Kernel description of domain variables
Presence (relevance) of domain variables per documents in literature (P)
3
Pairwise cooccurence, corelevance score
8
6
Domain vocabularies
Presence of terms in descriptions of domain variables (T)
4
Pairwise similarity score

Fig. 1. A framework for the combination of data, expert knowledge, and literature information. Arrow 1 shows the *tabula rasa* method for learning BNs when only data is available [20], Arrow 2 indicates methods for incorporating expert prior knowledge on structures and parametrizations [25], Arrow 3 indicates methods for quantifying the relations of domain variables based on their cooccurrence in documents [44, 28], and Arrow 4 denotes methods for quantifying the relations of domain variables based on their kernel descriptions [26, 42, 4]. We compare these text-based dependency scores against a prior from a medical expert (as indicated by Arrows 7 and 8) and investigate methods for transforming the text scores into an a priori distribution over Bayesian network structures (Arrows 5 and 6).

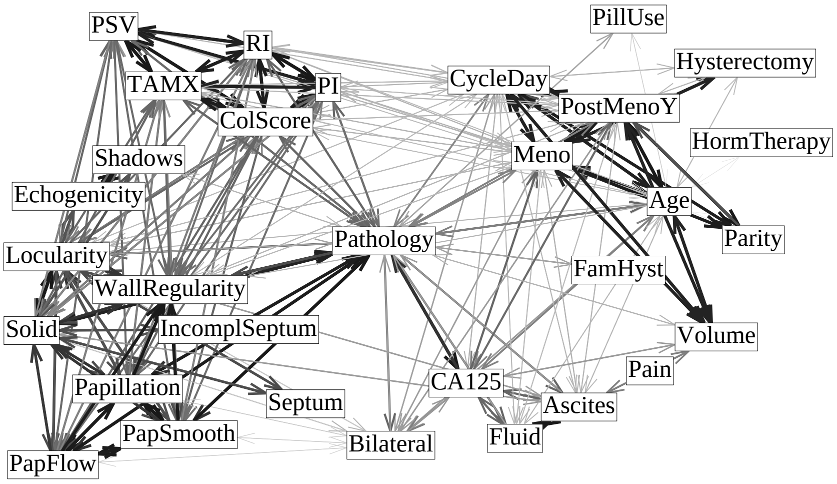

Fig. 2. The variables used in this paper and the relevant relations between them. The 'most relevant' pairwise relations, the 'moderately relevant' relations and the 'weakly relevant' relations selected by a medical expert are represented by edges with decreasing thickness.

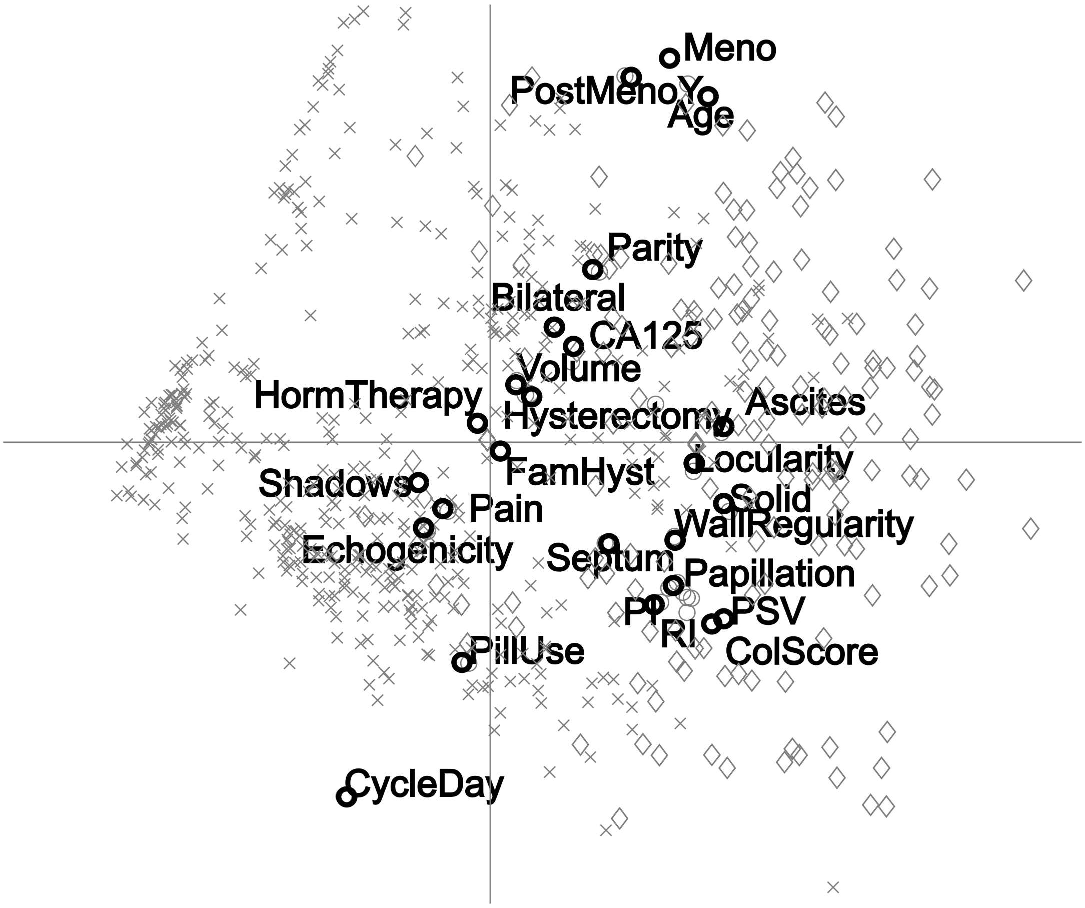

Fig. 3. The biplot of the domain variables and 604 cases used in this paper (not all of the thirty-one variables are shown). The variables are denoted by '**o**', the malignant cases by '◇' and the benign cases by 'x'.

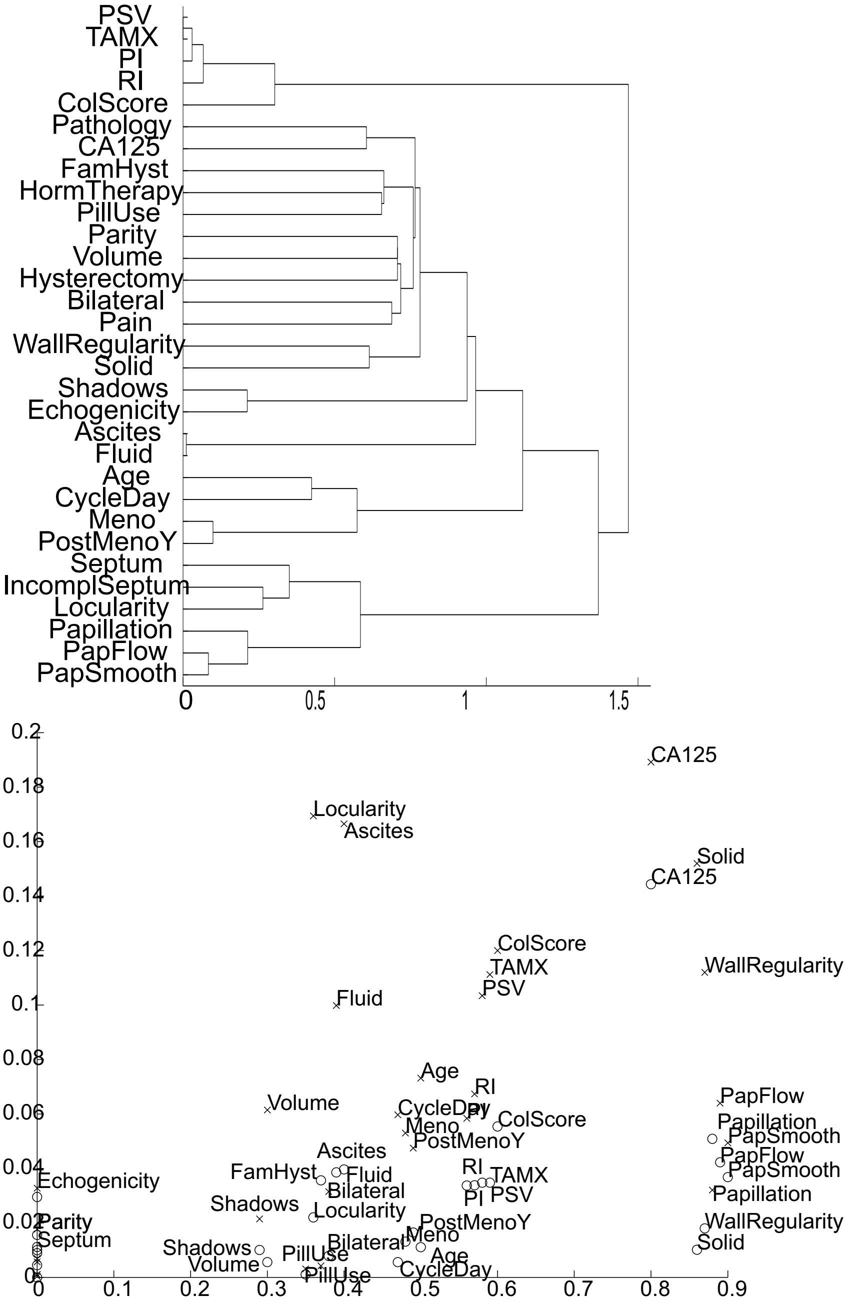

Fig. 4. Top: The hierarchical cluster tree of the relation score based on the cosine of the *tf-idf* vector representation of the kernel documents $R^{\mathrm{MI},C_3}_{\mathrm{COREL}}$. Bottom: The scatterplot of the *Pathology* relations for the domains variables X_i: the x axis holds the prior expert score $R_{\mathrm{Expert}}(\mathrm{Pathology}, X_i)$, while the y axis contains both the text score R_{ASIM} ('o') and the data score $R^{\mathrm{BD}}_{\mathrm{Data}}$ ('x').

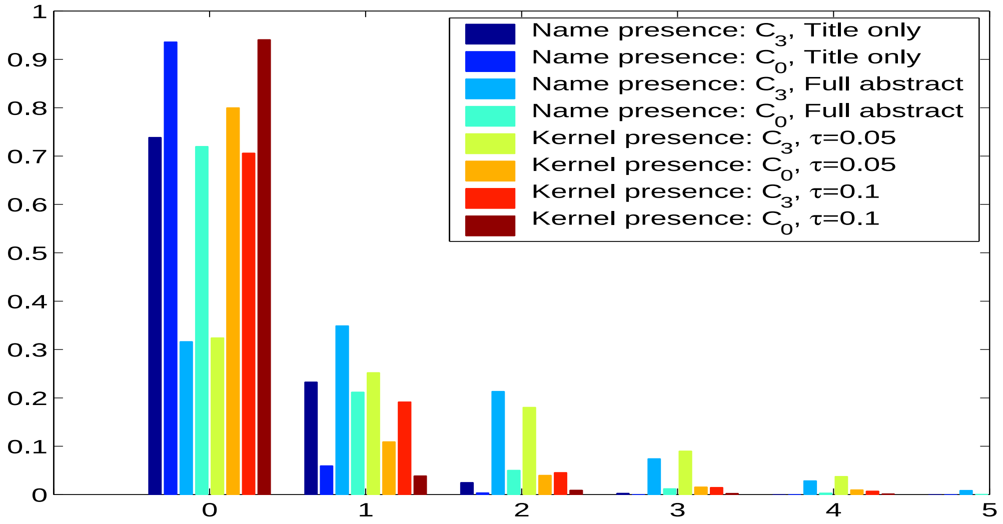

Fig. 5. For the four left bars of every group of eight bars: The percentage of abstracts (y axis) where 0, 1, 2, 3, 4, or 5 (indicated by the x axis) different variable names are present, only in the title or in the whole abstract using the small C_3 and large C_0 document sets. For the four right bars of every group of eight bars: the percentages of abstracts that are closer to 0, 1, 2, 3, 4, or 5 kernels of different variables than a threshold τ= 0.05 and τ = 0.1 as defined in Equation 3.

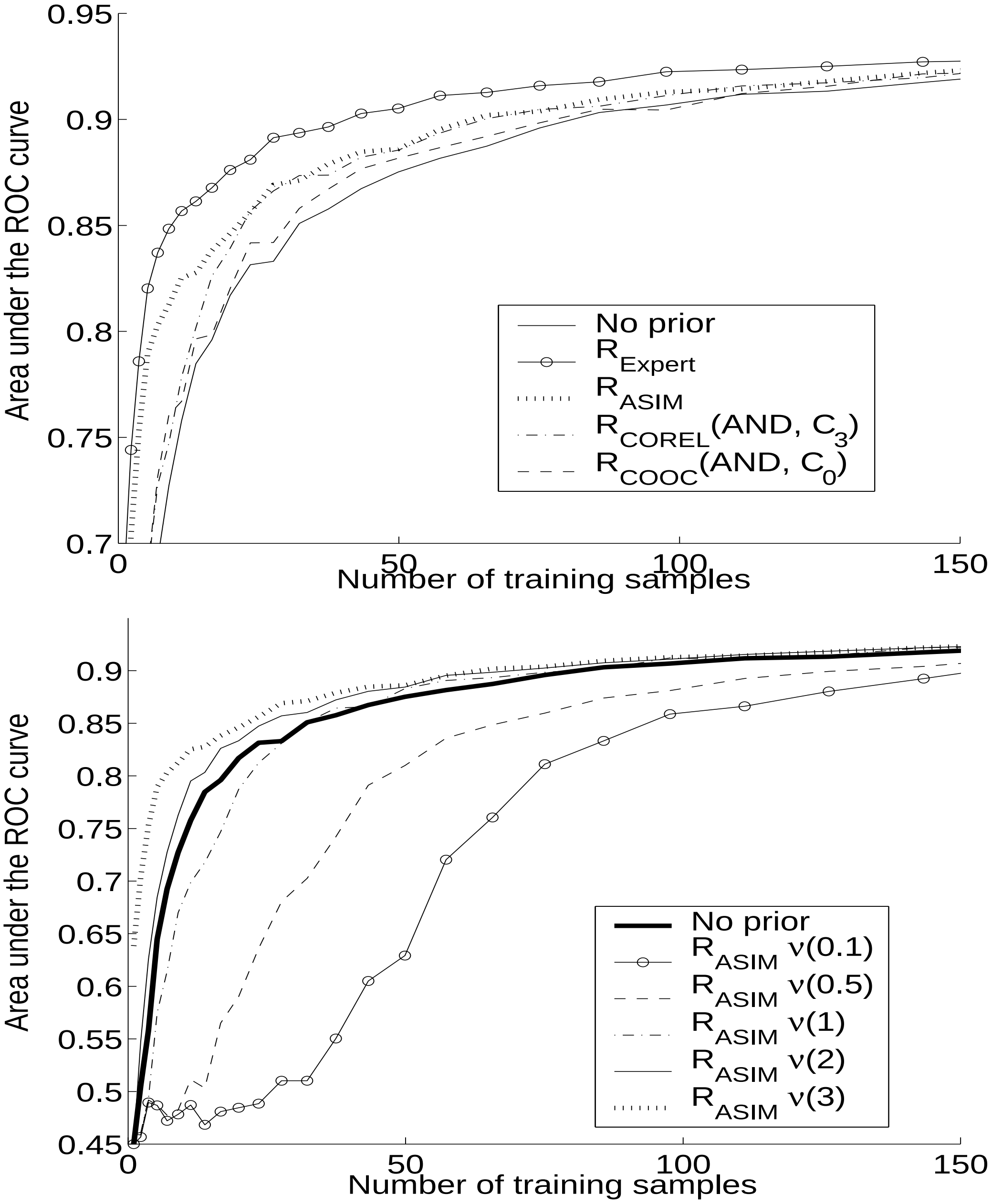

Fig. 6. Top: The AUC performance for BNs using different text-based priors ($R_{\text{COOC}}^{\text{AND},C_0}$, $R_{\text{COREL}}^{\text{AND},C_3}$), and R_{ASIM}), the expert prior R_{Expert}, and no prior. The priors are scaled to an average of 3 parents per variable. Bottom: The AUC performance for BN using the R_{ASIM} prior for different ν scalings (average number of parents is scaled to 0.1, 0.5, 1, 2, and 3) together with the performance without any prior.

List of Tables

Table 1
The percentage of abstracts where a certain variable name occurs and the percentages of abstracts that are closer to the kernel of domain variables than a threshold τ in the 'relevant' (largest) C_0 and the 'most relevant' (smallest) C_3 MEDLINE corpora. Name presence in collection C_i are denoted by $NP^T_{C_i}$ and $NP^A_{C_i}$ depending that only the title or the full abstract is used. The kernel presence in collection C_i are denoted by $KP^{\tau}_{C_i}$, in which τ denotes the required similarity of the kernel and the abstract, as defined in Equation 3.

Variable	$NP^T_{C_3}$	$NP^T_{C_0}$	$NP^A_{C_3}$	$NP^A_{C_0}$	$KP^{0.05}_{C_3}$	$KP^{0.1}_{C_3}$	$KP^{0.05}_{C_0}$	$KP^{0.1}_{C_0}$
Age	1.25	1.01	40.57	16.73	4.46	0.57	1.41	0.16
Ascites	0.55	0.10	2.66	0.37	2.64	1.22	0.60	0.18
Bilateral	0.48	0.13	5.41	0.84	2.27	0.28	0.72	0.09
CA125	2.29	0.15	4.61	0.27	19.87	3.84	2.16	0.32
ColScore	2.18	0.95	6.40	2.89	6.85	3.49	1.69	0.61
CycleDay	0.02	0.03	0.37	0.38	2.14	0.33	1.22	0.12
Echogenicity	0.18	0.07	0.98	0.99	1.94	0.50	0.62	0.10
FamHyst	0.07	0.01	0.26	0.11	9.94	2.58	1.49	0.32
Fluid	0.63	0.67	2.40	2.31	3.10	1.20	0.63	0.17
HormTherapy	1.05	0.29	2.14	0.64	8.45	2.44	3.80	1.12
Hysterectomy	2.12	0.28	11.55	1.03	9.11	3.12	1.53	0.50
IncomplSeptum	0.00	0.00	0.02	0.00	1.57	0.63	0.60	0.26
Locularity	1.40	0.64	3.56	1.29	3.60	0.90	0.84	0.30
Meno	0.13	0.21	1.92	0.99	2.97	0.85	1.23	0.30
Pain	0.17	0.38	1.92	1.39	2.53	0.44	1.48	0.58
PapFlow	0.00	0.00	0.00	0.00	5.09	1.09	1.06	0.27
Papillation	1.20	0.09	2.68	0.30	5.70	1.33	1.20	0.33
PapSmooth	0.00	0.00	0.00	0.00	4.83	0.96	1.01	0.26
Parity	0.09	0.02	1.40	0.77	1.03	0.28	1.25	0.28
Pathology	13.41	0.70	20.55	1.19	51.22	16.20	7.83	1.34
PI	0.11	0.03	1.75	0.52	4.65	2.45	1.28	0.48
PillUse	0.11	0.28	0.68	0.50	1.00	0.24	1.04	0.39
PostMenoY	1.20	0.34	3.36	0.83	3.71	0.87	1.42	0.27
PSV	0.00	0.00	0.50	0.06	4.67	2.47	1.29	0.48
RI	0.04	0.02	1.13	0.24	4.63	2.42	1.28	0.48
Septum	0.02	0.02	0.26	0.16	2.80	0.52	0.72	0.21
Shadows	0.00	0.01	0.06	0.08	2.08	0.44	0.65	0.11
Solid	0.07	0.14	1.42	0.81	2.60	0.31	0.79	0.18
TAMX	0.00	0.00	0.28	0.01	4.67	2.47	1.29	0.48
Volume	0.46	0.22	2.86	1.43	1.07	0.37	0.41	0.08
WallRegularity	0.00	0.00	0.00	0.00	2.60	0.48	0.91	0.17

Table 2
The AUC values for detecting important expert relations using the different text scores and the data scores (S_3 contains only the most important relations identified by the expert, S_0 contains a broader range of relevant relations as described in Section 2.1, S_1^P and S_3^P are their respective restrictions to the pairwise relations involving the variable *Pathology*). The specificity column presents the specificity values corresponding to the 50% sensitivity (i.e., it shows the percentage of *not* relevant relations that are correctly classified as *not* relevant when we demand that 50% of the relevant relations are correctly detected). The sensitivity column presents the sensitivity values corresponding to the 50% specificity (i.e., it shows the percentage of relevant relations that are correctly detected when we allow only 50% of the *not* relevant relations to be incorrectly classified as relevant). In each column, the three best values are indicated with bold.

	Area under the ROC curve (%)				Specificity (%)		Sensitivity (%)	
Settings	S_1	S_3	S_1^P	S_3^P	S_1	S_1^P	S_1	S_1^P
$R_{\text{COREL}}^{\text{MI},C_3}$	**82.01**	**93.24**	**78.26**	**95.83**	**90.43**	71.43	**90.74**	**82.61**
$R_{\text{COREL}}^{\text{MI},C_0}$	75.17	88.79	68.32	**91.67**	80.53	71.43	86.42	73.91
$R_{\text{COREL}}^{\text{AND},C_3}$	**82.10**	92.68	**78.26**	79.17	89.44	**100.00**	**90.74**	**82.61**
$R_{\text{COREL}}^{\text{AND},C_0}$	75.71	89.86	67.70	**90.97**	81.85	71.43	83.95	**82.61**
$R_{\text{COOC}}^{\text{MI},C_3}$	61.61	66.95	54.04	37.50	73.27	71.43	66.05	52.17
$R_{\text{COOC}}^{\text{MI},C_0}$	64.95	72.05	65.84	42.36	81.52	85.71	72.84	73.91
$R_{\text{COOC}}^{\text{AND},C_3}$	67.36	68.70	60.87	39.58	84.49	71.43	76.54	65.22
$R_{\text{COOC}}^{\text{AND},C_0}$	64.58	72.15	63.35	42.36	73.60	85.71	69.75	69.57
R_{ASIM}	65.83	88.48	75.78	88.89	80.20	**100.00**	67.28	69.57
$R_{\text{Data}}^{\text{BD}}$	75.99	**95.64**	**91.30**	75.69	**94.39**	**100.00**	77.16	**91.30**
$R_{\text{Data}}^{\text{MI}}$	**85.95**	**97.53**	**93.17**	72.92	**94.72**	**100.00**	**93.21**	**91.30**

Table 3
The Spearman rank correlation coefficients for the cross comparison of the expert score, the text scores, and the data scores (because of symmetry, the upper triangle presents the coefficients for comparing all the relations and the lower triangle presents the coefficients for comparing the relations related to the variable *Pathology*). The level of significance is indicated by underscore ($p< 0.05$), bold ($p< 0.001$), and bold underscore ($p\ll 0.001$).

settings	$R^{\mathrm{MI},C_3}_{\mathrm{COREL}}$	$R^{\mathrm{MI},C_0}_{\mathrm{COREL}}$	$R^{\mathrm{MI},C_3}_{\mathrm{COOC}}$	$R^{\mathrm{MI},C_0}_{\mathrm{COOC}}$	R_{ASIM}	$R^{\mathrm{BD}}_{\mathrm{Data}}$	$R^{\mathrm{MI}}_{\mathrm{Data}}$	R_{Expert}
$R^{\mathrm{MI},C_3}_{\mathrm{COREL}}$		$\underline{\mathbf{0.726}}$	$\underline{0.101}$	**0.111**	$\underline{\mathbf{0.508}}$	$\underline{\mathbf{0.385}}$	$\underline{\mathbf{0.408}}$	$\underline{\mathbf{0.507}}$
$R^{\mathrm{MI},C_0}_{\mathrm{COREL}}$	$\underline{\mathbf{0.787}}$		0.028	$\underline{0.081}$	$\underline{\mathbf{0.555}}$	$\underline{\mathbf{0.363}}$	$\underline{\mathbf{0.346}}$	$\underline{\mathbf{0.413}}$
$R^{\mathrm{MI},C_3}_{\mathrm{COOC}}$	-0.042	-0.117		$\underline{\mathbf{0.766}}$	-0.022	$\underline{\mathbf{0.139}}$	$\underline{\mathbf{0.193}}$	$\underline{\mathbf{0.175}}$
$R^{\mathrm{MI},C_0}_{\mathrm{COOC}}$	0.021	0.003	**0.684**		0.035	$\underline{\mathbf{0.179}}$	$\underline{\mathbf{0.268}}$	$\underline{\mathbf{0.237}}$
R_{ASIM}	**0.672**	**0.677**	-0.109	-0.006		$\underline{\mathbf{0.427}}$	$\underline{\mathbf{0.271}}$	$\underline{\mathbf{0.297}}$
$R^{\mathrm{BD}}_{\mathrm{Data}}$	$\underline{0.572}$	$\underline{0.473}$	0.010	0.160	$\underline{0.541}$		$\underline{\mathbf{0.629}}$	$\underline{\mathbf{0.471}}$
$R^{\mathrm{MI}}_{\mathrm{Data}}$	$\underline{0.513}$	$\underline{0.439}$	0.037	0.223	$\underline{0.534}$	$\underline{\mathbf{0.968}}$		$\underline{\mathbf{0.546}}$
R_{Expert}	**0.627**	$\underline{0.527}$	-0.119	0.009	$\underline{0.537}$	**0.640**	**0.650**	