Estimation of effect size posterior using model averaging over Bayesian network structures and parameters

Gabor Hullam
Budapest University of Technology and Economics, Hungary
gabor.hullam@mit.bme.hu

Peter Antal
Budapest University of Technology and Economics, Hungary
peter.antal@mit.bme.hu

Abstract

Bayesian networks provide multiple advantages in feature subset analysis, but the parametric properties of Bayesian networks are typically neglected in inductive approaches, as they are focused on existential inference of relevant variables. However, in many application fields, particularly in biomedicine, certain parametric properties, such as effect size measures, are essential and widely used in interpretation, because they provide a more detailed characterization of relevance. To cope with multiple hypothesis testing, the Bayesian statistical framework became popular, but the posterior distribution of these effect size measures derived by Bayesian model averaging over structures are analytically not tractable, thus their estimation requires Monte Carlo simulation methods. In this paper we compare the structural (existential) and parametric (quantitative) concepts of Bayesian relevance and overview Bayesian approaches to the estimation of effect size posteriors. Furthermore, we compare the high probability density regions of these posteriors with traditional frequentist confidence intervals. Methods are illustrated on an artificial data set simulating a genetic association study.

1 Introduction

The application of graphical models, especially Bayesian networks, became increasingly popular in the field of biomedicine and genomics, especially in feature subset analysis when the need for modeling potentially complex dependency structures between genomic, environmental, and clinical factors and disease state indicators emerged. Recently, it became apparent that the simple phenotype approach, which is applied in most genome-wide association studies, has considerable limits. Analyses based on standard statistical methods using the hypothesis testing framework were frequently unsuccessful. This phenomenon induced an extensive research for applicable methods, capable of handling the needs of genetic research.

Bayesian network based learning methods appeared in this field with the promise of revealing the possibly complex dependency relationships between the investigated factors. Several Bayesian network based methods focus on the learning of some structural properties or the whole structure from data. Even though learning structural properties provides invaluable information on the relationships between factors, and enables the selection of relevant factors with respect to a specified target, it neglects parametric properties such as effect size. One might argue, that one of the main benefits of Bayesian networks with conjugate priors in the context of structure learning is that the parametric level can be bypassed, i.e. analytically averaged out. However, in several domains, such as biomedicine, effect size as a quantitative descriptor plays a central role in the interpretation of the results from feature subset analysis.

Therefore, a Bayesian effect size measure applicable within the same framework as the structure learning methods could enhance the interpretation of such results. Earlier approaches include a Bayesian estimation of log odds ratios in presence of prior information (Demirhan and Hamurkaroglu, 2008), a Bayesian method for the non-informative prior case (Rahardja et al., 2010), and an inference of posteriors of mutual information using a Bayesian framework (Hutter and Zaffalon, 2005).

Although there are many possible measures for the relevance of a variable, three main approaches can be distinguished: the association based approach neglecting structural aspects, the causal approach assuming a fixed structure, and the existential approach averaging over structures. The widespread association based approach uses effect size measures that do not take structural, multivariate relationships into account. Odds ratio (OR) is the most widely used such measure, especially in case-control studies.

The causal approach measures the effect of a variable X on another Y given a structure describing causal relationships, i.e. the nature and the strength of a relationship between variables X and Y. Structural equation modeling and the average causal effect measure (Pearl, 2000) are two related methods providing a measure of effect size assuming a known causal structure. The obvious drawback of these methods is the lack of learning structures.

The structural (existential) uncertainty based approach uses Bayesian networks, which provide a graph based language for encoding relevance and representing dependency relationships. Assuming a Bayesian framework, a high posterior for strong relevance indicates a non-mediated relationship between X and Y, i.e. it indicates a close structural connection (Friedman and Koller, 2003).

Note that parametric relevance does not imply structural relevance (e.g. strong relevance), and vice versa a relatively high posterior of strong relevance does not imply relatively high parametric relevance, e.g. a quantitative measure of effect size e.g. odds ratio, denoted as $\mathrm{OR}(X, Y)$, may indicate the relevance of X with respect to a selected target Y just by being over a certain threshold e.g. $\mathrm{OR}(X, Y) \geq 2.5$ (the structural relation between X and Y is unknown and has no influence on this aspect of relevance). This shows that the structural strong relevance and parametric effect size aspects appear to be two separate dimensions of relevance. The main goal of this paper is to connect these two aspects of relevance by introducing a structure based Bayesian effect size measure $OR(X, Y|\theta, G)$ based on $p(\theta, G|D_N)$, where θ denotes the parametrization and G the structure of an underlying BN, and D_N denotes data (i.e. $OR(X, Y|\theta, G)$ is a random variable with distribution $p(\theta, G|D_N)$).

A further reason behind the structure based Bayesian effect size measure is that there is no closed form for the posterior of effect size measures, such as odds ratio. Thus, the posterior has to be estimated, e.g. by sampling the log-odds ratio of a Dirichlet distribution based on data. Another possible solution is to use the structural properties of BNs to guide the estimation process. Though instead of learning a whole Bayesian network from the data, the learning of relevant variables with respect to a target (i.e. the Markov blanket of the target) is sufficient.

In the following section we discuss the importance of structural properties of BNs and their relation to relevance and to effect size. Then we introduce the Bayesian odds ratio and describe various methods for its estimation. Finally, we demonstrate the properties of Bayesian odds ratio using an artificial data set.

2 Strong relevance and its relation to structural properties of BNs

Relevance is a central concept in all processes in which learning is involved. In case of learning structural properties of a Bayesian network from data, the identification of relevant elements is essential. In realistic cases, with at least a hundred variables, the learning of the whole network is practically untractable. A possible approach is to learn smaller substruc-

tures or structural properties that are statistically better confirmed.

Several methods were created for this purpose, starting with a Bayesian inference over structural properties of Bayesian networks (Buntine, 1991; Cooper and Herskovits, 1992). In (Madigan et al., 1996) a Markov Chain Monte Carlo (MCMC) scheme was proposed to approximate such Bayesian inference. Subsequently, Friedman et al. reported an MCMC scheme over the space of orderings (Friedman and Koller, 2003). Then in (Koivisto and Sood, 2004) a method to perform exact full Bayesian inference over modular features was introduced, and an ad hoc randomized approach was reported in (Pena et al., 2007). Recent applications in genetic association studies were discussed in (Jiang et al., 2010; Xing et al., 2011).

Despite the leading role of relevance, its relation to structural properties and to the concepts of association and effect size, seemed unclarified in some cases. Thus we focused our efforts towards investigating various structural properties of BNs related to relevance. We demonstrated the Bayesian application of Bayesian networks in relevance analysis (Antal et al., 2006), and proposed a Bayesian network based Bayesian multilevel analysis of relevance (Antal et al., 2008a). We carried out a comparative study of BN-BMLA against other methods in (Hullám et al., 2010), and applied the BN-BMLA method in a candidate gene association study of asthma (Ungvari et al., 2012).

Here we provide a brief summary of the most important aspects of relevance. First of all, the concept of relevance can be defined in multiple ways. On the one hand, it can be defined specific to the applied model class used as a predictor, the optimization algorithm, the data set, and the loss function (Kohavi and John, 1997).

Definition 1 (Strong and weak relevance)**.** A feature X_i is strongly relevant to Y, if there exists some $X_i = x_i, Y = y$ and $s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, s_i) > 0$ such that $p(y|x_i, s_i) \neq p(y|s_i)$. A feature X_i is weakly relevant, if it is not strongly relevant, and there exists a subset of features S_i' of S_i for which there exists some x_i, y and s_i' for which $p(x_i, s_i') > 0$ such that $p(y|x_i, s_i') \neq p(y|s_i')$. A feature is relevant, if it is either weakly or strongly relevant; otherwise it is irrelevant.

In contrast, association as a relevance type only means that there is a relationship between X_i and Y such that they are statistically dependent $p(y|x_i) \neq p(y)$, however association does not provide information to discriminate between strong and weak relevance. This is also indicated by the possibility that X_i is strongly relevant, yet there is no association found with Y. In other words, strong relevance entails strong criteria on the structure of an underlying BN, whereas association does not necessarily. Finally, another difference between strong relevance and association is that the strength of association can be characterized by several measures, such as odds ratio or mutual information, whereas there are no such quantitative descriptors for the strength of strong relevance. These differences also confirm the necessity of dual characterization of relevance. For this purpose we propose the structure based Bayesian odds ratio, which is a combination of effect size and structural information.

The relationship between structural properties of BNs and strong relevance is revealed by the following model-based, probabilistic definition of Markov blankets (Pearl, 1988) and by a subsequent theorem.

Definition 2 (Markov blanket)**.** A set of variables $\mathbf{X}' \subseteq \boldsymbol{V}$ is called a *Markov blanket* set of X_i w.r.t. the distribution $p(\boldsymbol{V})$, if $(X_i \perp\!\!\!\perp \boldsymbol{V} \setminus \mathbf{X}' | \mathbf{X}')_p$, where $\perp\!\!\!\perp$ denotes conditional independence.

The following theorem gives a sufficient condition for the unambiguous BN representation of the relevant structural properties (Tsamardinos and Aliferis, 2003).

Theorem 1. *For a distribution p defined by Bayesian network (G, θ) the variables* $\mathrm{bd}(Y, G)$ *form a Markov blanket of Y, where* $\mathrm{bd}(Y, G)$ *denotes the set of parents, children and the children's other parents for Y (Pearl, 1988). If the distribution p is stable with respect to the DAG G, then* $\mathrm{bd}(Y, G)$ *forms a unique and mini-*

mal Markov blanket of Y, denoted as $\mathrm{MBS}_p(Y)$. *Furthermore,* $X_i \in \mathrm{MBS}_p(Y)$ *iff* X_i *is strongly relevant.*

We also refer to $\mathrm{bd}(Y,G)$ as the Markov blanket set for Y in G using the notation $\mathrm{MBS}(Y,G)$ by the implicit assumption that p is Markov compatible with G. Based on $\mathrm{bd}(Y,G)$ a pairwise relation can be defined indicating whether X_j is a member of $\mathrm{bd}(Y,G)$. This relation is called Markov blanket membership and it is denoted as $\mathrm{MBM}(X_j,Y)$.

In other words, this means that an $\mathrm{MBS}(Y,G)$ contains all the strongly relevant variables X_j with respect to Y. Furthermore, estimation of effect size measure for the strongly relevant variables in $\mathrm{MBS}(Y,G)$ depends only from parameters in the Markov Blanket subGraph under multinomial sampling and global parameter independence, thus we can rely on this set of variables and parameters instead of taking the whole BN structure into consideration (see Eq. 4).

3 Bayesian effect size

In several fields, such as biomedicine and genetics, the most frequently used effect size measure is the odds ratio (Balding, 2006).

Let $X_1, X_2, ..., X_n$ denote discrete variables that encode single nucleotide polymorphism (SNP) states 0,1,2 that refer to common (wild) homozygote, heterozygote, rare (mutant) homozygote genotypes respectively. Then $X_i^{(s)}$ denotes SNP X_i in state s. In case of a disease indicator Y, i.e. $Y^{(0)}$: non-affected (control), $Y^{(1)}$: affected (case), an *odds* is defined as

$$\mathrm{Odds}(X_i^{(s)}) = \frac{p(Y^{(1)}|X_i^{(s)})}{p(Y^{(0)}|X_i^{(s)})} \qquad (1)$$

Consequently an *odds-ratio* e.g. heterozygous (1) versus common homozygous (0) is given as

$$\mathrm{OR}(X_i^{(1,0)}) = \frac{Odds_{X_i^{(1)}}}{Odds_{X_i^{(0)}}} \qquad (2)$$

Although odds ratio is a simple measure, the computation of significance of its value (and its confidence interval) is plagued by multiple testing. A possible approach is to apply univariate Bayesian methods using various priors as suggested by (Stephens and Balding, 2009). However, this was not extended to the multivariate case, i.e. all SNPs were treated as independent entities, thus prohibiting the analysis of joint effects.

A possible multivariate Bayesian approach is to utilize the underlying BN (G,θ) for odds ratio computation $\mathrm{OR}(\theta)$. However, assuming a Dirichlet prior $\theta \sim Dir(X_i, \alpha_i)$, this computation is analytically non-tractable. Therefore,we propose to sample the parameters in the 'relevant part' of the BN, instead of the whole structure. The structural property in which the relevant structural elements (with respect to a target) are encoded is called the Markov blanket graph (Acid et al., 2005; Antal et al., 2006).

Definition 3 (Markov blanket graph)**.** A Markov blanket graph $MBG(Y,G)$ of variable Y is a subgraph of a Bayesian network structure G, which contains the nodes of the Markov blanket set of Y, that is $MBS(Y,G)$ and the incoming edges into Y and its children. Given a target node, which corresponds to the target variable Y, $MBG(Y,G)$ as a (sub)graph structure consists of nodes that are (1) parents of Y, (2) children of Y or (3) "other parents" of the children of Y.

Note, that from all the possible edges between these nodes, $MBG(Y,G)$ only contains those that end in Y or one of its children. Mechanism boundary graph is another term for Markov blanket graph due to its significant role in describing causal relationships (i.e. mechanisms). As we discussed in section 1, the causal or mechanism based approach is one of the main approaches to effect size computation. According to which, given a causal relationship $A \rightarrow B$ the effect size measure defines the amount of effect a value of variable A has on specific values of B. Since the Bayesian network representation can be interpreted in a causal context under the Causal Markov Condition (for details see (Pearl, 2000)) then an edge between vertices X_i and X_j can denote a causal relationship $X_i \rightarrow X_j$.

¿From this perspective $MBG(Y, G)$ describes the direct causal relationships of Y (i.e. edges from nodes of type (1) and edges to nodes of type (2)), and also some of the indirect relationships of Y that form a special dependency pattern, called a *v-structure* (Pearl, 2000).

Note that $\mathrm{MBG}(Y, G)$ includes all the mechanisms, in which Y is involved. This property makes the Markov blanket graph an ideal candidate to serve as a base for measuring effect size both from the causal aspect and the structural relevance aspect. The latter is due to the fact that the nodes of $\mathrm{MBG}(Y, G)$ represent the variables of $\mathrm{MBS}(Y, G)$, that is the strongly relevant variables with respect to Y.

Since the goal is to measure the effect of a factor X_i (e.g. SNPs) on a target variable Y (e.g. disease susceptibility) the fact that X_i is a member of $\mathrm{MBG}(Y, G)$ is a relevant information. If X_i is a member, then the probability of a certain value of the target Y (e.g. in case of Y as a disease status the values are: "case" and "control") can be estimated based on the $\mathrm{MBG}(Y, G)$ and a specific instantiation of X_i. This in turn allows the computation of a Markov blanket graph based odds ratio

$$\mathrm{OR}(X_i^{(q,r)} | X_i \in \mathrm{MBG}(Y, G)) = \frac{p(Y^{(1)}|mbg, X_i^{(q)})}{p(Y^{(0)}|mbg, X_i^{(q)})} \cdot \frac{p(Y^{(0)}|mbg, X_i^{(r)})}{p(Y^{(1)}|mbg, X_i^{(r)})} \quad (3)$$

where mbg refers to $X_{j \neq i} \in \mathrm{MBG}(Y, G)$. In the opposite case, when X_i is not a member of $\mathrm{MBG}(Y, G)$ then its effect on Y is negligible, because the mechanisms in $\mathrm{MBG}(Y, G)$ shield Y from the effect of X_i.

4 The estimation of Bayesian odds ratio

In terms of BN learning, the first step of effect size estimation is the learning of structures, that is identifying the direct and indirect relationships between entities of a domain, i.e. between the variables of a data set. In order to characterize the relationships, a second step, the learning of conditional distribution parameters is needed. As we discussed in section 2, instead of learning the whole BN, the learning of structural properties is a viable choice. For this purpose we used BN-BMLA (Antal et al., 2008b) which relies on a directed acyclic graph based Markov Chain Monte Carlo method (Madigan et al., 1996) to estimate posteriors of structural properties, such as Markov blanket graphs.

A simple approximate solution to estimate the distribution of the log odds ratio is to rely on the maximum a posteriori $\mathrm{MBG}(Y, G)$ with a mean parametrization learned from data and a Gaussian approximation for its variance. A more sophisticated approach takes the parametric uncertainty into account by sampling the conditional distribution parameters. Based on this sampling of parameterizations a distribution of Bayesian odds ratios arises. This allows the estimation of the credible interval of Bayesian odds ratios, which is the Bayesian analogue of the confidence interval used in the frequentist approach. The drawback of this method is that it assumes that $\mathrm{MBG}(Y, G)$ was identified previously or that the maximum a posteriori Markov blanket graph has such a high posterior that all others can be neglected.

A third possible approach to effect size estimation takes the structural uncertainty of $\mathrm{MBG}(Y, G)$ into account. This is necessary because the cardinality of Markov blanket graphs is even greater than that of Markov blanket sets, which is super exponential in the number of variables. This means that the sufficient sample size for the estimation of posteriors of Markov blanket graphs is also higher. Typically, in a practical case there are thousands of Markov blanket graphs with relatively low posteriors. Therefore, selecting the $\mathrm{MBG}_k(Y, G)$ with the highest posterior is not the best option. A more viable solution is to apply model averaging on Markov blanket graphs. In case of the estimation of the Bayesian odds ratio of X_i with respect to Y based on a specific $\mathrm{MBG}_k(Y, G)$, the measure $OR(X_i^{(q,r)} | X_i \in \mathrm{MBG}_k(Y, G))$ has to be weighted according to the posterior probability of the specific Markov blanket graph $p(\mathrm{MBG}_k(Y, G))$. Finally, the averaging of these

measures over possible Markov blanket graphs leads to the Bayesian MBG-based odds ratio (BOR) as

$$BOR(X_i^{(q,r)}, Y) = \sum_{k=1}^{m} OR(X_i^{(q,r)} | Mbg_k, \theta^{mbg_k}) \cdot p(Mbg_k) \cdot 1(X_i \in \mathrm{MBG}_k(Y, G)), \quad (4)$$

where Mbg_k denotes $\mathrm{MBG}_k(Y, G)$, m is the number of MBGs with a posterior $p(Mbg_k) > 0$, and $1()$ is the indicator function.

Even though BOR implements a fully Bayesian approach, its application is limited due to its computationally intensive nature. It can be a reasonable compromise to estimate BOR based on a selection of the M number of $\mathrm{MBG}_k(Y, G)$ with the highest posteriors.

5 Results

We investigated the properties of Bayesian odds ratio on an artificial data set (5000 samples, 115 variables) simulating a candidate gene association study. This asthma related data set contained 11 relevant variables with high posteriors of strong relevance forming the reference Markov blanket graph seen on figure 1.

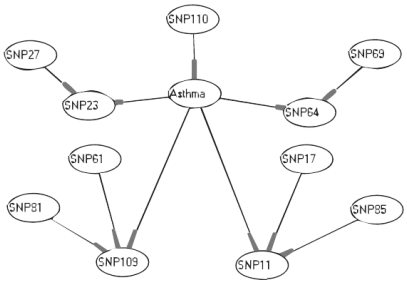

Figure 1: Reference Markov blanket graph containing all the strongly relevant variables of the data set.

The distribution of odds ratios is shown in figure 2. As depicted by figure 3 not all relevant variables had high odds ratios, in fact the majority of these relevant variables had odds ratios close to 1. A high posterior for strong relevance accompanied by an irrelevant odds ratio can indicate that the variable is in an interaction with other variables. This is in accordance with original data set, which contained several interacting elements.

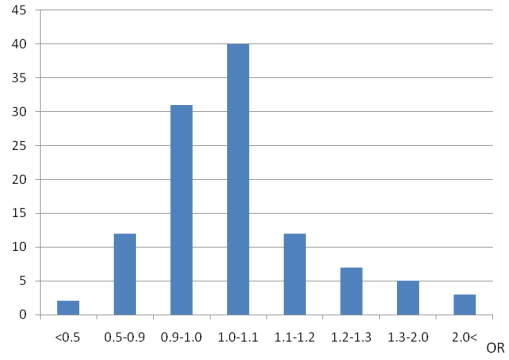

Figure 2: Distribution of odds ratios in the artificial data set.

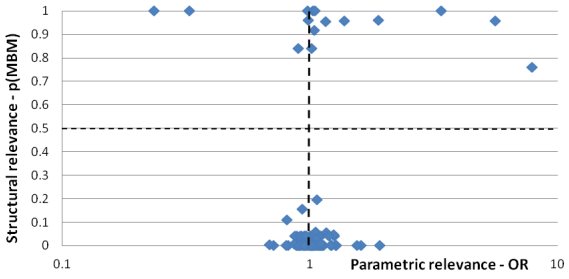

Figure 3: Structural and parametric relevance of variables of the data set. Posteriors of strong relevance are shown on the vertical, odds ratios are shown on the horizontal axis.

We computed odds ratios and corresponding confidence intervals using 300, 500, 1000 and 5000 samples. Confidence intervals were corrected for multiple hypothesis testing using the number of a priori known, strongly relevant variables (11, see figure 1). Bayesian odds ratios and related credible intervals were estimated using Markov blanket graphs learned by the BN-BMLA method. Table 1 compares the properties of two selected variables (S11 and S23) which are both strongly relevant and have high odds ratios. As figure 4 shows, in case of 5000 samples the length of Bayesian credible intervals is smaller and are contained within the cor-

rected confidence intervals in these examples.

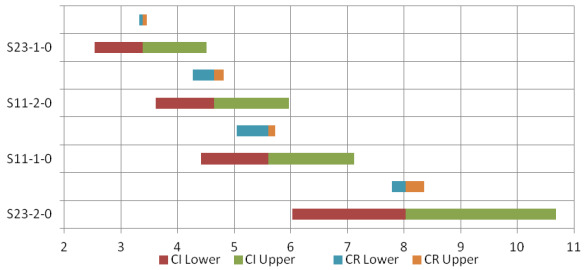

Figure 4: Comparison of confidence intervals (CI) and credible intervals (CR) of selected variables based on the data set of 5000 samples. S011-1-0 and S011-2-0 denote odds ratios of variable S011, 1 vs.0 and 2 vs.0 respectively.

With decreasing sample size the length of confidence intervals increases, whereas the length of Bayesian credible intervals remains relatively constant (see figure 5).

Table 1: Comparison of confidence intervals (CI) and Bayesian credible intervals (CR) in case of a data set of 5000 and 1000 samples. Suffixes L and U denote the Lower and the Upper half of the interval respectively.

S-5000	OR	CI-L	CI-U	CR-L	CR-U
S11 1 vs 0	5.60	4.41	7.12	5.05	5.73
S11 2 vs 0	4.65	3.62	5.97	4.27	4.82
S23 1 vs 0	3.38	2.54	4.51	3.32	3.46
S23 2 vs 0	8.02	6.03	10.68	7.78	8.36
S-1000	OR	CI-L	CI-U	CR-L	CR-U
S11 1 vs 0	6.71	3.79	11.87	4.42	5.08
S11 2 vs 0	5.39	2.96	9.82	3.65	4.28
S23 1 vs 0	3.23	1.76	5.91	2.93	3.49
S23 2 vs 0	6.87	3.77	12.53	6.02	7.56

However, Bayesian odds ratio has its own limitation in the small sample size domain. This is due to the insufficiency of the data for learning Markov blanket graphs. This results in remarkably different Markov blankets even among the most relevant ones. Figure 6 illustrates the case when the Bayesian odds ratio is computed separately for the 10 most relevant Markov blanket graphs for 300 samples. As it can be seen, a joint (i.e. based on all MBGs) estimation of a Bayesian credible interval is rather problematic in this case, although a possible solution is to cover the whole concerned region.

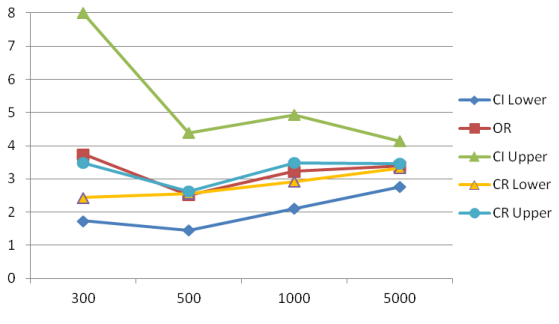

Figure 5: The effect of sample size on the confidence interval (CI) and the credible interval (CR) of a selected variable.

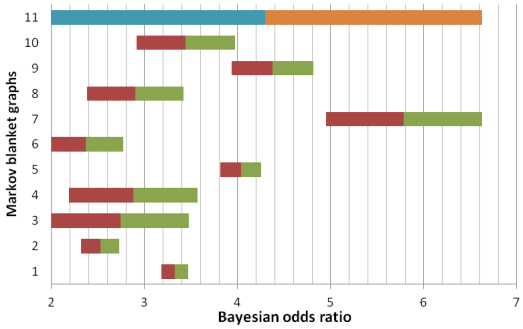

Figure 6: Bayesian Markov blanket based odds ratios of a selected variable in case of a data set of 300 samples. The vertical axis denotes the 10 highest ranking Markov blankets.

6 Conclusion

Bayesian networks and the Bayesian statistical framework provide multiple advantages in feature subset analysis, particularly in the exploration of interactions and in the characterization of relevance types (Ungvari et al., 2012). Beside the multivariate aspect, Bayesian networks in the Bayesian statistical framework can also be used to explore the quantitative aspect of relevance by estimating the distribution of effect size parameters. In the paper we showed that the posterior probability of strong relevance and the distribution of effect size conditioned on the strong relevance of variables provide a rich characterization of relevance. Due to the lack of analytic methods, Monte Carlo methods and approximations are

necessary. These results indicate that multiple aspects of relevance can be characterized coherently in a full Bayesian approach performing model averaging over model structures. Currently we are exploring intermediate levels between the structural (existential) and parametric (quantitative) levels, such as the monotonicity of the relation of the variables.

Acknowledgments

This work was supported by the National Innovation Office NKTH TECH 08-A1/2-2008-0120, by the R3-COP project (Robust&Safe Mobile Co-op Syst) of ARTEMIS JU and the National Dev. Agency, and by the grant TAMOP-4.2.2.B-10/1–2010-0009.

References

S. Acid, L. M. de Campos, and J. G. Castellano. 2005. Learning bayesian network classifiers: searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235.

P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. 2006. Learning complex bayesian network features for classification. In *Proc. of third European Workshop on Prob. Graph. Models*, p. 9–16.

P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus. 2008a. A bayesian multilevel analysis of feature relevance. In *Workshop on New challenges for feature selection FSDM08*,p.63–74.

P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus. 2008b. A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. *JMLR Proceeding*, 4:74–89.

D. J. Balding. 2006. A tutorial on statistical methods for population association studies. *Nature*, 7:781–91.

W. L. Buntine. 1991. Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in AI*, pages 52–60. Morgan Kaufmann.

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

H. Demirhan and C. Hamurkaroglu. 2008. Bayesian estimation of log odds ratios from rxc and 2x2xk contingency tables. *Statistica Neerlandica*, 62(4):405–424.

N. Friedman and D. Koller. 2003. Being Bayesian about network structure. *Mach Learn*, 50:95–125.

G. Hullám, P. Antal, Cs. Szalai, and A. Falus. 2010. Evaluation of a bayesian model-based approach in ga studies. *JMLR W&C Proc.*, 8:30–43.

M. Hutter and M. Zaffalon. 2005. Distribution of mutual information from complete and incomplete data. *Comp Stat& Data An.*, 48(3):633–657.

X. Jiang, M. M. Barmada, and S. Visweswaran. 2010. Identifying genetic interaction in genome-wide data using bayesian networks. *Genetic Epidemiology*, 34:575–581.

R. Kohavi and G. H. John. 1997. Wrappers for feature subset selection. *AI*, 97:273–324.

M. Koivisto and K. Sood. 2004. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research*, 5:549–573.

D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. 1996. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25:2493–2520.

J. Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA.

J. Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

J.M. Pena, R. Nilsson, J. Bjrkegren, and J. Tegnér. 2007. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45:211–232.

D. Rahardja, Y. D. Zhao, and H. H. Zhang. 2010. Bayesian inference of odds ratios in misclassified binary data with a validation substudy. *Commun. Stat.-Simulation&Computation*, 39(10):1845–1854.

M. Stephens and D.J. Balding 2009. Bayesian statistical methods for genetic association studies. *Nature Review Genetics*, 10(10):681–690.

I. Tsamardinos and C. Aliferis. 2003. Towards principled feature selection: Relevancy,filters, and wrappers. In *Proc. of the AI&Stat.*, p. 334–342.

I. Ungvari, G. Hullam, P. Antal, and C. Szalai et al. 2012. Evaluation of a partial genome screening of two asthma susceptibility regions using bayesian network based bayesian multilevel analysis of relevance. *PLOS ONE*, 7:1–14.

H. Xing, P. D. McDonagh, J. Bienkowska, and J. Carulli et al. 2011. Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLoS Comp. Biol.*, 7(3):1001105.