JMLR: Workshop and Conference Proceedings 3: 249-256 NIPS 2009 workshop on causality

TIED: An Artificially Simulated Dataset with Multiple Markov Boundaries

Alexander Statnikov ALEXANDER.STATNIKOV@VANDERBILT.EDU
Discovery Systems Laboratory
Department of Biomedical Informatics
Vanderbilt University
Nashville, TN 37232, USA

Constantin F. Aliferis CONSTANTIN.ALIFERIS@NYUMC.ORG
Center for Health Informatics and Bioinformatics
New York University
New York, NY 10016, USA

Editors: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

We present an artificially simulated dataset (*TIED*) constructed so that there are many minimal sets of variables with maximal predictivity (i.e., Markov boundaries) and likewise many sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of the response variable. This dataset was used in the *Potluck Causality Challenge* to determine all statistically indistinguishable sets of direct causes and direct effects and all Markov boundaries of the response variable and also to predict the response variable in the independent test data. We also present baseline results of application of several algorithms to this dataset.

Keywords: local causal discovery, Markov boundary induction, variable selection, classification

1. Introduction

The problem of variable/feature selection is of fundamental importance in machine learning and applied statistics, especially when it comes to analysis, modeling, and discovery from high-dimensional data (Guyon and Elisseeff, 2003; Kohavi and John, 1997). In addition to the promise of cost-effectiveness, two major goals of variable selection are to improve the prediction performance of the predictors and to provide a better understanding of the data-generative process (Guyon and Elisseeff, 2003). An emerging class of algorithms proposes a principled solution to the variable selection problem by identification of a Markov blanket of the response variable of interest (Aliferis et al., 2009; Aliferis et al., 2003; Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003). A *Markov blanket* is a set of variables conditioned on which all the remaining variables excluding the response variable are statistically independent of the response variable. A related concept is a *Markov boundary* (or non-redundant Markov blanket) that is a Markov blanket such that no proper subset of it is a Markov blanket (Pearl, 1988). Under assumptions about the learner and loss function, a Markov boundary is the solution to the variable selection problem (Tsamardinos and Aliferis, 2003).

An important theoretical result states that if the distribution satisfies the *intersection property*[1], then it is guaranteed to have a unique Markov boundary of the response variable (Pearl, 1988). Furthermore, if the distribution satisfies common causal assumptions such as *faithfulness*, *Markov condition*, and *causal sufficiency*, then the Markov boundary is also unique and consists only of direct causes, direct effects, and direct causes of direct effects (also known as "spouses") of the response variable in the underlying causal graph (Tsamardinos and Aliferis, 2003). Even though there are several well-developed algorithms for learning a Markov boundary either in faithful distributions or in distributions where the intersection property holds (Aliferis et al., 2009; Peña et al., 2007; Aliferis et al., 2003; Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003), little research has been done in development of algorithms for learning *multiple* Markov boundaries from the same dataset when the above assumptions do not hold.

We present an artificially simulated dataset (*TIED*) that contains multiple Markov boundaries (and thus violates the intersection and faithfulness properties) and likewise many sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of the response variable. This dataset was used in the *Potluck Causality Challenge* to determine all statistically undistinguishable sets of direct causes and direct effects and all Markov boundaries of the response variable and also to predict the response variable in the independent test data. We also present baseline results of application of several algorithms to this dataset.

2. Dataset

Using the principles from (Lemeire, 2006), we constructed a discrete Bayesian network *TIED* with 1,000 variables (including a response variable T). Figure 1 shows a fragment of the network structure and specifies which variables contain the same information about T by the color of highlighting. The parameterization of the network fragment shown in Figure 1 is provided in Table 1. The network fragment contains a response variable T, all variables that participate in all Markov boundaries of the response variable T, and some other variables. The full network can be obtained by adding 10 children to each variable from the set $\{X_5, X_6, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{18}, X_{19}, X_{20}\}$ (a total of 110 variables) with conditional probability distribution defined in Table 2 and 860 variables that do not have a path to T in the network. If variables X and Y are shown with the same color in Figure 1, then (a) for every combination of values of X and T such that $P(T = t \mid X = x) = p$, there exists a value y of variable Y such that $P(T = t \mid Y = y) = p$, and (b) for every combination of values of Y and T such that $P(T = t \mid Y = y) = p$, there exists a value x of variable X such that $P(T = t \mid X = x) = p$. Such variables are interchangeable for prediction of T, and therefore if X belongs to a Markov boundary $\boldsymbol{M_1}$ of T, then $\boldsymbol{M_2} = (\boldsymbol{M_1} \setminus \{X\}) \cup \{Y\}$ is another Markov boundary of T. The work of (Lemeire, 2006) specifically describes why such variables violate the intersection property of the probability distribution. In summary, the network contains 72 Markov boundaries of T. Each of these Markov boundaries contains 5 variables: (i) X_9, (ii) X_4 or X_8, (iii) X_{11} or X_{12} or X_{13}, (iv) X_{18} or X_{19} or X_{20}, and (v) X_1 or X_2 or X_3 or X_{10}. Similarly, there are 72 sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of T. These sets of variables coincide with the Markov boundaries of T.

[1] We use notation $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ to denote that subset of variables $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$ in the underlying probability distribution. Let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{W}$ be any four disjoint subsets of variables. Then the probability distribution satisfies the intersection property if $\mathbf{X} \perp \mathbf{Y} \mid (\mathbf{Z} \cup \mathbf{W})$ and $\mathbf{X} \perp \mathbf{W} \mid (\mathbf{Z} \cup \mathbf{Y}) \Rightarrow \mathbf{X} \perp (\mathbf{Y} \cup \mathbf{W}) \mid \mathbf{Z}$.

The dataset *TIED* was obtained by sampling 3,750 instances from the above Bayesian network. 750 (20%) instances were used for discovery of multiple Markov boundaries (or sets of variables that are statistically indistinguishable from the set of direct causes and direct effects) of *T*, and the remaining 3,000 (80%) instances were used for validation of classification performance of *T*. We also computed the optimal Bayes classification performance of *T* which is 0.9663 weighted accuracy[2].

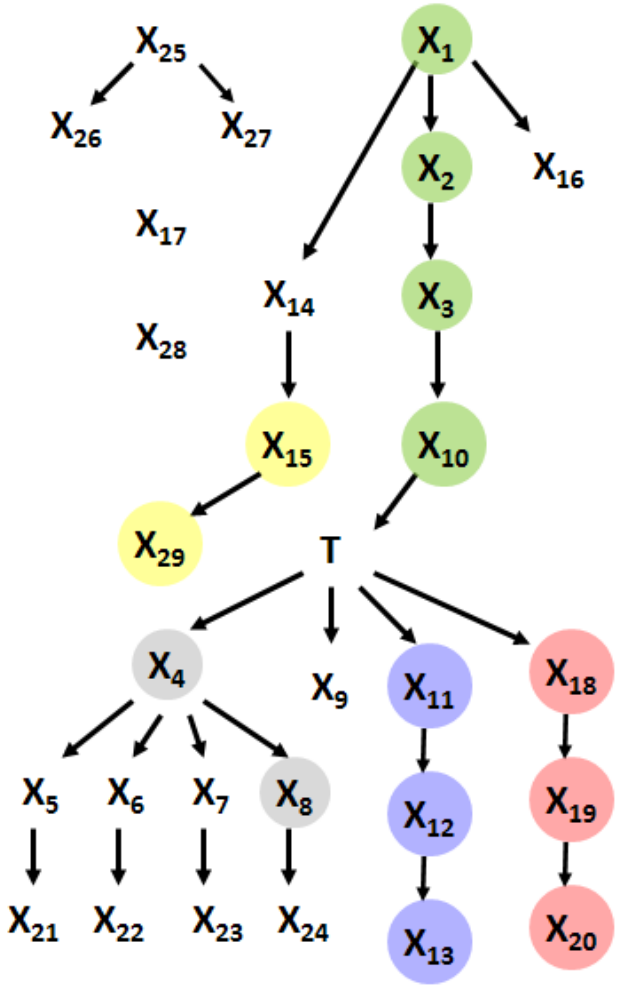

Figure 1: Graphical visualization of the fragment of a discrete Bayesian network *TIED*. Variables that contain exactly the same information about *T* are highlighted with the same color, e.g. variables X_{11}, X_{12}, and X_{13} provide exactly the same information about *T* and thus are interchangeable for prediction of *T*.

[2] Weighted accuracy is defined as the average proportion of correct classifications in each category/class of the response variable.

T: P(T=0|X_{10}=0) = 1.0 P(T=0|X_{10}=1) = 1.0 P(T=0|X_{10}=2) = 1.0 P(T=1|X_{10}=3) = 0.3 P(T=2|X_{10}=3) = 0.3 P(T=3|X_{10}=3) = 0.4	X_5: P(X_5=0|X_4=0) = 0.6 P(X_5=1|X_4=0) = 0.2 P(X_5=2|X_4=0) = 0.2 P(X_5=0|X_4=1) = 0.5 P(X_5=1|X_4=1) = 0.25 P(X_5=2|X_4=1) = 0.25 P(X_5=0|X_4=2) = 0.8 P(X_5=1|X_4=2) = 0.1 P(X_5=2|X_4=2) = 0.1	X_{10}: P(X_{10}=0|X_3=0) = 1.0 P(X_{10}=0|X_3=1) = 1.0 P(X_{10}=1|X_3=2) = 0.3 P(X_{10}=2|X_3=2) = 0.7 P(X_{10}=3|X_3=3) = 1.0
X_1: P(X_1=0) = 0.25 P(X_1=1) = 0.25 P(X_1=2) = 0.25 P(X_1=3) = 0.25	X_6: P(X_6=1|X_4=0) = 0.5 P(X_6=2|X_4=0) = 0.5 P(X_6=0|X_4=1) = 0.8 P(X_6=1|X_4=1) = 0.2 P(X_6=0|X_4=2) = 0.2 P(X_6=1|X_4=2) = 0.3 P(X_6=2|X_4=2) = 0.5	X_{11}: P(X_{11}=0|T=0) = 1.0 P(X_{11}=0|T=1) = 1.0 P(X_{11}=0|T=2) = 1.0 P(X_{11}=1|T=3) = 0.5 P(X_{11}=2|T=3) = 0.5
X_2: P(X_2=0|X_1=0) = 0.8 P(X_2=1|X_1=0) = 0.2 P(X_2=0|X_1=1) = 0.1 P(X_2=1|X_1=1) = 0.9 P(X_2=2|X_1=2) = 1.0 P(X_2=3|X_1=3) = 1.0	X_7: P(X_7=0|X_4=0) = 0.9 P(X_7=1|X_4=0) = 0.1 P(X_7=0|X_4=1) = 0.7 P(X_7=1|X_4=1) = 0.2 P(X_7=2|X_4=1) = 0.1 P(X_7=0|X_4=2) = 0.6 P(X_7=1|X_4=2) = 0.3 P(X_7=2|X_4=2) = 0.1	X_{12}: P(X_{12}=0|X_{11}=0) = 1.0 P(X_{12}=1|X_{11}=1) = 0.5 P(X_{12}=2|X_{11}=1) = 0.5 P(X_{12}=1|X_{11}=2) = 0.5 P(X_{12}=2|X_{11}=2) = 0.5
X_3: P(X_3=0|X_2=0) = 0.3 P(X_3=1|X_2=0) = 0.7 P(X_3=0|X_2=1) = 0.8 P(X_3=1|X_2=1) = 0.2 P(X_3=2|X_2=2) = 1.0 P(X_3=3|X_2=3) = 1.0	X_8: P(X_8=1|X_4=0) = 1.0 P(X_8=2|X_4=1) = 1.0 P(X_8=0|X_4=2) = 1.0	X_{13}: P(X_{13}=0|X_{12}=0) = 1.0 P(X_{13}=1|X_{12}=1) = 0.5 P(X_{13}=2|X_{12}=1) = 0.5 P(X_{13}=1|X_{12}=2) = 0.5 P(X_{13}=2|X_{12}=2) = 0.5
X_4: P(X_4=1|T=0) = 0.9 P(X_4=2|T=0) = 0.1 P(X_4=0|T=1) = 0.8 P(X_4=1|T=1) = 0.1 P(X_4=2|T=1) = 0.1 P(X_4=0|T=2) = 0.1 P(X_4=1|T=2) = 0.8 P(X_4=2|T=2) = 0.1 P(X_4=0|T=3) = 0.1 P(X_4=1|T=3) = 0.1 P(X_4=2|T=3) = 0.8	X_9: P(X_9=0|T=0) = 0.1 P(X_9=1|T=0) = 0.8 P(X_9=2|T=0) = 0.1 P(X_9=1|T=1) = 0.1 P(X_9=2|T=1) = 0.9 P(X_9=0|T=2) = 0.1 P(X_9=1|T=2) = 0.8 P(X_9=2|T=2) = 0.1 P(X_9=0|T=3) = 0.2 P(X_9=1|T=3) = 0.7 P(X_9=2|T=3) = 0.1	X_{14}: P(X_{14}=0|X_1=0) = 0.8 P(X_{14}=1|X_1=0) = 0.1 P(X_{14}=2|X_1=0) = 0.1 P(X_{14}=0|X_1=1) = 0.1 P(X_{14}=1|X_1=1) = 0.8 P(X_{14}=2|X_1=1) = 0.1 P(X_{14}=0|X_1=2) = 0.8 P(X_{14}=1|X_1=2) = 0.1 P(X_{14}=2|X_1=2) = 0.1 P(X_{14}=0|X_1=3) = 0.1 P(X_{14}=1|X_1=3) = 0.1 P(X_{14}=2|X_1=3) = 0.8

Table 1 (continued on the next page): Parameterization of the *TIED* network for variables shown in Figure 1 {T, X_1, X_2, X_3, X_4, …, X_{29}}. Only nonzero probabilities are shown in the table.

X_{15}: P(X_{15}=0|X_{14}=0) = 1.0 P(X_{15}=0|X_{14}=1) = 1.0 P(X_{15}=1|X_{14}=2) = 0.5 P(X_{15}=2|X_{14}=2) = 0.5	X_{20}: P(X_{20}=0|X_{19}=0) = 1.0 P(X_{20}=1|X_{19}=1) = 1.0 P(X_{20}=2|X_{19}=2) = 1.0	X_{25}: P(X_{25}=0) = 0.5 P(X_{25}=1) = 0.5
X_{16}: P(X_{16}=0|X_1=0) = 0.2 P(X_{16}=1|X_1=0) = 0.6 P(X_{16}=2|X_1=0) = 0.2 P(X_{16}=0|X_1=1) = 0.1 P(X_{16}=1|X_1=1) = 0.3 P(X_{16}=2|X_1=1) = 0.6 P(X_{16}=0|X_1=2) = 0.5 P(X_{16}=1|X_1=2) = 0.1 P(X_{16}=2|X_1=2) = 0.4 P(X_{16}=0|X_1=3) = 0.3 P(X_{16}=1|X_1=3) = 0.5 P(X_{16}=2|X_1=3) = 0.2	X_{21}: P(X_{21}=0|X_5=0) = 0.2 P(X_{21}=1|X_5=0) = 0.6 P(X_{21}=2|X_5=0) = 0.2 P(X_{21}=0|X_5=1) = 0.1 P(X_{21}=1|X_5=1) = 0.3 P(X_{21}=2|X_5=1) = 0.6 P(X_{21}=0|X_5=2) = 0.5 P(X_{21}=1|X_5=2) = 0.1 P(X_{21}=2|X_5=2) = 0.4	X_{26}: P(X_{26}=0|X_{25}=0) = 0.1 P(X_{26}=1|X_{25}=0) = 0.9 P(X_{26}=0|X_{25}=1) = 0.3 P(X_{26}=1|X_{25}=1) = 0.7
X_{17}: P(X_{17}=0) = 0.25 P(X_{17}=1) = 0.25 P(X_{17}=2) = 0.25 P(X_{17}=3) = 0.25	X_{22}: P(X_{22}=0|X_6=0) = 0.3 P(X_{22}=1|X_6=0) = 0.2 P(X_{22}=2|X_6=0) = 0.5 P(X_{22}=0|X_6=1) = 0.8 P(X_{22}=1|X_6=1) = 0.1 P(X_{22}=2|X_6=1) = 0.1 P(X_{22}=0|X_6=2) = 0.6 P(X_{22}=1|X_6=2) = 0.2 P(X_{22}=2|X_6=2) = 0.2	X_{27}: P(X_{27}=0|X_{25}=0) = 0.4 P(X_{27}=1|X_{25}=0) = 0.6 P(X_{27}=0|X_{25}=1) = 0.8 P(X_{27}=1|X_{25}=1) = 0.2
X_{18}: P(X_{18}=1|T=0) = 0.1 P(X_{18}=2|T=0) = 0.9 P(X_{18}=0|T=1) = 0.1 P(X_{18}=2|T=1) = 0.9 P(X_{18}=0|T=2) = 0.8 P(X_{18}=1|T=2) = 0.1 P(X_{18}=2|T=2) = 0.1 P(X_{18}=0|T=3) = 0.1 P(X_{18}=1|T=3) = 0.8 P(X_{18}=2|T=3) = 0.1	X_{23}: P(X_{23}=0|X_7=0) = 0.5 P(X_{23}=1|X_7=0) = 0.1 P(X_{23}=2|X_7=0) = 0.4 P(X_{23}=0|X_7=1) = 0.6 P(X_{23}=1|X_7=1) = 0.3 P(X_{23}=2|X_7=1) = 0.1 P(X_{23}=0|X_7=2) = 0.7 P(X_{23}=1|X_7=2) = 0.1 P(X_{23}=2|X_7=2) = 0.2	X_{28}: P(X_{28}=0) = 0.33 P(X_{28}=1) = 0.33 P(X_{28}=2) = 0.33
X_{19}: P(X_{19}=1|X_{18}=0) = 1.0 P(X_{19}=2|X_{18}=1) = 1.0 P(X_{19}=0|X_{18}=2) = 1.0	X_{24}: P(X_{24}=0|X_8=0) = 0.8 P(X_{24}=1|X_8=0) = 0.1 P(X_{24}=2|X_8=0) = 0.1 P(X_{24}=0|X_8=1) = 0.6 P(X_{24}=1|X_8=1) = 0.2 P(X_{24}=2|X_8=1) = 0.2 P(X_{24}=0|X_8=2) = 0.5 P(X_{24}=1|X_8=2) = 0.3 P(X_{24}=2|X_8=2) = 0.2	X_{29}: P(X_{29}=0|X_{15}=0) = 1.0 P(X_{29}=1|X_{15}=1) = 0.5 P(X_{29}=2|X_{15}=1) = 0.5 P(X_{29}=1|X_{15}=2) = 0.5 P(X_{29}=2|X_{15}=2) = 0.5

Table 1 (continued from the previous page)

P(*Z* | *X*)	*X = 0*	*X = 1*	*X = 2*
Z = 0	0.3	0.4	0.3
Z = 1	0.3	0.3	0.4
Z = 2	0.4	0.3	0.3

Table 2: Conditional probability distribution of each of 110 variables (denoted by *Z*) mentioned in Section 2 that have a single parent from the set {X_5, X_6, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{18}, X_{19}, X_{20}} (denoted by *X*).

3. Experiments and Results

The experiments involved running several algorithms for discovery of multiple Markov boundaries:

- Four resampling-based techniques that apply a variable selection algorithm to bootstrap samples from the original dataset: The following variable selection methods were used: (i) SVM-based recursive feature elimination (SVM-RFE) (Guyon et al., 2002); (ii) SVM-RFE with additional application of McNemar's test (Everitt, 1977) to identify the most parsimonious variable set with classification performance statistically indistinguishable from the observed best one; (iii) backward wrapping with linear SVM classifier based on univariate ranking of variables by Kruskal-Wallis non-parametric ANOVA (Hollander and Wolfe, 1999); and (iv) backward wrapping with linear SVM classifier based on Kruskal-Wallis ANOVA with additional statistical comparison step, as in (ii). The above four methods are denoted as *Resampling-SVM-RFE1*, *Resampling-SVM-RFE2*, *Resampling-Univariate1*, *Resampling-Univariate2*, respectively. Since there is no natural termination criterion of these methods, they were run on 5,000 bootstrap samples from the original dataset.
- Three instantiations of KIAMB algorithm (Peña et al., 2007): KIAMB was applied with G^2 test, parameter $K = 0.8$, and three statistical thresholds $\alpha = 0.01$, $\alpha = 0.005$, and $\alpha = 0.001$ (denoted as *KIAMB1*, *KIAMB2*, *KIAMB3*, respectively). The first threshold was used by inventors of the method in the paper that introduced it (Peña et al., 2007). Since there is no natural termination criterion of these methods, they were run 5,000 times.
- *Iterative Removal* method (Natsoulis et al., 2005): This method works as follows: First, it extracts a Markov boundary from the original dataset and estimates its classification performance. Second, it removes all variables from the original dataset that were found to participate in the Markov boundaries, extracts a new tentative Markov boundary from the modified dataset, and estimates its classification performance. Finally third, if the classification performance of the tentative Markov boundary is statistically indistinguishable from the Markov boundary obtained in the first step, then this is also a true Markov boundary and the second and third steps of the algorithm are repeated. The implementation of this method used an algorithm HITON-PC (Aliferis et al., 2009; Aliferis et al., 2003) to learn a Markov boundary and McNemar's test to compare linear SVM classification performance of resulting variable sets (Everitt, 1977).

All methods were applied to the 750-instance training dataset to indentify Markov boundaries of the response variable *T*. Once the Markov boundaries were indentified, a linear SVM classifier was trained with these variable sets in the training dataset and it was applied to the 3,000-instance validation dataset. The classification performance was measured by the weighted accuracy metric

(Guyon et al., 2006). In independent tests (not shown here) the choice of a linear SVM versus non-linear one was validated as not compromising classification performance.

The results of experiments are presented in Table 3. The following are observed: (i) *Iterative Removal* identifies only one Markov boundary because all other Markov boundaries have a common variable (X_9) and thus cannot be detected by this method. This is s structural deficiency of that method. (ii) *KIAMB* fails to identify any true Markov boundaries due to its sample inefficiency (its sample requirements are of exponential order to the number of variables in the Markov boundary), and because of the same reason its output Markov boundaries have poor predictivity; (iii) Resampling-based methods either miss many true Markov boundaries and/or output many false positive variables in the identified Markov boundaries.

Method	Total number of output Markov boundaries	Number of variables in an average output Markov boundary	Number of true Markov boundaries		Average number of false positive variables in identified true Markov boundaries	Average classification performance in validation data	CPU time in minutes
			identified exactly	identified with false positive variables			
Iterative Removal	3	5.67	0	1	2.00	0.959	0.04
KIAMB1	5000	2.82	0	0	-	0.798	285.42
KIAMB2	5000	2.81	0	0	-	0.796	285.45
KIAMB3	5000	2.80	0	0	-	0.796	285.48
Resampling + Univariate1	5000	11.10	0	72	12.29	0.942	5999.64
Resampling + Univariate2	5000	5.58	0	0	-	0.934	6000.41
Resampling + RFE1	5000	8.70	0	72	6.38	0.952	6235.28
Resampling + RFE2	5000	4.24	0	29	5.76	0.947	6235.93

Table 3: Results of experiments with artificial dataset *TIED*. All experiments were executed on a cluster with Intel 2.4 GHz Xeon CPU's.

4. Conclusion

This report introduced an artificially simulated dataset (*TIED*) with multiple Markov boundaries and multiple sets of variables that are statistically indistinguishable from the set of direct causes and direct effects of the response variable. We also presented baseline results of several algorithms in this dataset. The results demonstrate that *TIED* is a challenging problem and many methods fail to discover multiple Markov boundaries from this dataset. Therefore, there is a need to create new algorithms to indentify multiple Markov boundaries.

References

Aliferis,C.F. et al. (2009) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification. Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*. (to appear)

Aliferis,C.F., Tsamardinos,I. and Statnikov,A. (2003) HITON: a novel Markov blanket algorithm for optimal variable selection. *AMIA 2003 Annual Symposium Proceedings*, 21-25.

Everitt,B. (1977) *The analysis of contingency tables*. Chapman and Hall, London.

Guyon,I. et al. (2006) *Feature extraction: foundations and applications*. Springer-Verlag, Berlin.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.

Guyon,I. et al. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, 46, 389-422.

Hollander,M. and Wolfe,D. (1999) *Nonparametric statistical methods*. Wiley, New York, NY, USA.

Kohavi,R. and John,G.H. (1997) Wrappers for feature subset selection. *Artificial Intelligence*, 97, 273-324.

Lemeire,J. (2006) The representation and learning of equivalent information in causal models. *Technical Report IRIS-TR-0099*.

Natsoulis,G. et al. (2005) Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. *Genome Res.*, 15, 724-736.

Pearl,J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, California.

Peña,J. et al. (2007) Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45, 211-232.

Tsamardinos,I. and Aliferis,C.F. (2003) Towards principled feature selection: relevancy, filters and wrappers. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AI & Stats)*.

Tsamardinos,I., Aliferis,C.F. and Statnikov,A. (2003) Time and sample efficient discovery of Markov blankets and direct causal relations. *Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining (KDD)*, 673-678.