INTEGRATION OF CLINICAL AND MICROARRAY DATA USING BAYESIAN NETWORKS

Olivier Gevaert [*] **Frank De Smet** [*,**]
Dirk Timmerman [***] **Yves Moreau** [*]
Bart De Moor [*]

[*] *Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium*
[**] *Medical Direction, National Alliance of Christian Mutualities, Haachtsesteenweg 579, 1031 Brussel, Belgium*
[***] *Department of Obstetrics and Gynecology, University Hospital Gasthuisberg, Katholieke Universiteit Leuven, Herestraat 49, 3000 Leuven, Belgium*

Abstract: Microarrays have revolutionized research in molecular biology especially in cancer research. They allow to measure the expression of thousands of genes and can be used to guide clinical management of cancer. However, mathematical models based on microarray data often ignore the available clinical data, instead of integrating clinical and microarray data. We present and evaluate three methods for integrating clinical and microarray data using Bayesian networks: full integration, partial integration and decision integration, and use them to predict prognosis in breast cancer. Partial integration performs best on the test set and is promising for other types of cancer and data.

Keywords: data models, medical applications, discrete systems, probabilistic models, biomedical systems

1. INTRODUCTION

In the past decade microarrays have changed research in molecular biology. A microarray is a collection of probes that represent a selection of genes on a solid surface. When RNA is extracted from a tumor sample for example and applied onto this surface, we can measure the expression of thousands of genes. Among other applications, microarrays can be used to guide the clinical management of cancer. Mathematical models built using microarray data can be used to model the phenotype of a tumour and, can predict the clinical behaviour. Because of the low signal-to-noise ratio of microarray data, integration of other sources of information in the clinical decision process is important. More reliable models can be built when multiple sources of information are combined. However, the current focus of attention is on microarray data. When available it is the only source of information that is modeled. The available clinical data is usually ignored although it contains useful and independent information. We propose methods that treat the clinical data on an equal footing with the microarray data. We also want to stress that this approach is rarely applied when studying microarray data.

Bayesian networks have been used to achieve the integration of clinical and microarray data. A Bayesian network (Pearl, 1988; Neapolitan, 2004) is a model situated in a probabilistic framework that can be used for any type of reasoning. The major difference between Bayesian networks and 'classical' system identification is that the model is non-dynamic but includes a causal interpretation. Furthermore this model is very flexible for integrating data sources. It is possible to combine data sources directly or by combining them at the decision level. Furthermore, due to the way Bayesian networks are learned from data, we will define a third method for integrating data sources. To the author's knowledge, the first two methods have not been previously applied in this context and the third method has not been previously defined. We will present these three methods for incorporating clinical and microarray data and we will evaluate them using Receiver Operator Characteristic curves (ROC). The best methods for integrating the clinical and the microarray data will be tested on an independent test set.

We will focus as an example on the prediction of the prognosis in lymph node negative breast cancer (without apparent tumor cells in local lymph nodes at diagnosis). We define the outcome as a variable that can have two values: poor prognosis or good prognosis. Poor prognosis corresponds to recurrence within 5 years after diagnosis and good prognosis corresponds to a disease free interval of at least 5 years. If we can distinguish between these two groups, patients could be treated accordingly thus eliminating over- or under-treatment.

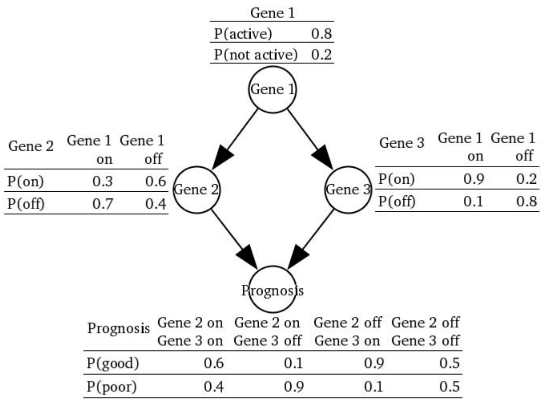
2. BAYESIAN NETWORKS

2.1 Model class

A Bayesian network is a probabilistic model (Pearl, 1988; Neapolitan, 2004) that consists of two parts: a dependency structure (also called a Directed Acyclic Graph) and local probability models. An example with four binary variables is shown in figure 1. The dependency structure specifies how the variables are related to each other by drawing directed edges between the variables without creating any directed cycles. The edges define the (in)dependency relations that exist between the variables. Usually each variable x_i only depends on a few other variables, called the parents:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i | Pa(x_i)) \qquad (1)$$

where $Pa(x_i)$ stands for the parents of x_i; for example, the prognosis variable in figure 1 has two parents: gene 2 and gene 3. This means that the full joint distribution (the space of all possible patients) $p(x_1, ..., x_n)$ can be decomposed into independent factors. In this manner a Bayesian network is a sparse way of writing down the joint probability distribution instead of specifying the full joint distribution, which requires an intractable number of parameters. This is the most important idea behind Bayesian networks namely that they allow a dramatic decrease in the number of parameters that is needed to specify a probabilistic distribution over a number of variables. Otherwise for n binary variables $2^n - 1$ different probabilities would have to be specified; one probability for each instantiation of the variables. For example, the Bayesian network in figure 1 needs 9 parameters vs. 15 parameters that are needed for the full joint distribution. The second part of this model, the local probability models, specifies how the variables depend on their parents. We used discrete-valued Bayesian networks, which means that these local probability models can be represented with Conditional Probability Tables (CPTs). Such a table specifies the probability that a variable takes a certain value given the value of its parents. In figure 1 these tables are shown next to the nodes. The columns in each table represent the specific instantiation of the parents of a specific node. Gene 1 has no parents therefore this node's table specifies a priori probabilities.



Gene 1	
P(active)	0.8
P(not active)	0.2

Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Gene 3	Gene 1 on	Gene 1 off
P(on)	0.9	0.2
P(off)	0.1	0.8

Prognosis	Gene 2 on Gene 3 on	Gene 2 on Gene 3 off	Gene 2 off Gene 3 on	Gene 2 off Gene 3 off
P(good)	0.6	0.1	0.9	0.5
P(poor)	0.4	0.9	0.1	0.5

Fig. 1. A simple Bayesian network with 4 binary variables.

2.2 Model estimation

Learning discrete-valued Bayesian networks from data proceeds in two steps: structure learning and parameter learning.

2.2.1. Model structure selection First the dependency structure that best explains the data is constructed. This is done using a scoring metric combined with a search strategy. The scoring metric describes the probability of the structure

S given the data, D. When we have n variables $x_1, ..., x_i, ..., x_n$ with r_i the number of values of each variable and q_i the number of instantiations of the parents of each variable than the scoring metric is defined as (Cooper and Herskovits, 1992; Heckerman *et al.*, 1995):

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right], \quad (2)$$

with $p(S)$ the prior probability of the structure. We used an uninformative prior biased towards edges with the outcome variable for all developed models. Therefore edges with the outcome variable were more likely a priori than other edges. N_{ijk} are the number of cases in D having variable i in state k with the j-th instantiation of its parents in S. A superscript is added when necessary and refers to the data set the counts are taken from. Then $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$. N'_{ijk} are the prior counts and correspond with a prior for the parameters. When no knowledge is available they are estimated using $N'_{ijk} = N/(r_i q_i)$ (Heckerman *et al.*, 1995) with N the equivalent sample size. N corresponds to the importance of the prior counts.

Equation 2 allows to score structures and now we have to define a search strategy to find a good model. An exhaustive search is infeasible since the number of structures becomes intractably large when there are much variables. Therefore we used the greedy search algorithm K2 (Cooper and Herskovits, 1992). This algorithm uses a prior ordering of the variables to restrict the number of structures that can be built. This means that x_i can only become a parent of x_j if x_i precedes x_j in the ordering. Equation 2 also shows that the score decomposes into independent factors where each factor represents the addition to the score from each variable. Therefore K2 iteratively tries to find to best parents for each variable separately. This is done by starting with an empty set of parents for a certain variable and incrementally adding the parent that increases the score of the current variable the most, taking the ordering restriction into account. The algorithm stops when no more parents can be added that increase the score. Because the prior ordering of the variables is not known in advance we repeat the model building process for a set of random variable orderings. For each of these orderings, a structure is learned and the structure with the highest score is kept.

2.2.2. Model parameter identification The second step of the model building process consists of estimating the parameters of the local probability models corresponding with the dependency structure. In section 2.1 we reported that we are using CPTs to model these local probability models. For each variable and instantiation of its parents there exists a CPT that consists of a set of parameters. Each set of parameters was given a uniform Dirichlet prior:

$$p(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1}, ..., N'_{ijr_i}) \quad (3)$$

with θ_{ij} a parameter set where i refers to the variable and j to the j-th instantiation of the parents in the current structure. θ_{ij} contains a probability for every value of the variable x_i given the current instantiation of the parents. Dir corresponds to the Dirichlet distribution with $(N'_{ij1}, ..., N'_{ijr_i})$ as parameters of this Dirichlet distribution. Parameter learning then consists of updating these Dirichlet priors with data. This is straightforward because the multinomial distribution that is used to model the data, and the Dirichlet distribution that models the prior, are conjugate distributions. This results in a Dirichlet posterior over the parameter set:

$$p(\theta_{ij}|D, S) = Dir(\theta_{ij}|N'_{ij1} + N_{ij1}, ..., N'_{ijr_i} + N_{ijr_i}) \quad (4)$$

with N_{ijk} defined as before. We summarized this posterior by taking the Maximum A Posteriori (MAP) parameterization of the Dirichlet distribution and used these values to fill in the corresponding CPTs for every variable.

2.3 Classification

After learning the model, we can use it for classification. This means that we can use a model to predict the outcome variable given the value of the other variables. In the context of Bayesian networks this is called inference. We used the Probability propagation in tree of cliques algorithm (PPTC) (Huang and Darwiche, 1996) to predict the probability of the outcome on the test set. The models were then evaluated using the Area Under the ROC curve (AUC) of the predictions for the outcome variable.

2.4 Model building

We evaluated the performance of the different methods (see section 3) for integrating both data sources using the training data. This was done by randomizing the training data 100 times for each method, in a stratified way, into a set of 70% of the patients used to build the model (model building data set) and a set of 30% to estimate the Area Under the ROC curve (AUC). This AUC is a measure for the independent data set performance of a model. Then these 100 AUCs were averaged and reported. In this manner we can evaluate the generalizing performance of a specific method and compare with other methods. Next, the method that performed best in the previous step was used to train 100 models using the complete training set with different initial orderings of the variables. The model with the highest AUC on the training data among these 100 models was chosen to predict the outcome on the test set. The AUC for this test set was calculated and represents the performance of this model on unseen data.

3. INTEGRATION OF DATA SOURCES

Bayesian networks allow to combine the clinical and microarray data in different ways. Apart from using them separately we will combine both data sources using three methods: full integration, decision integration and partial integration. D^c, D^m and D^{cm} refer to the clinical data, microarray data and combined clinical and microarray data respectively. Analogously for the references to the structures: S^c, S^m and S^{cm}.

3.1 Full integration

Full integration is equal to putting both data sources together and treating them as if it is one dataset, D^{cm}. This means that both the clinical variables (e.g. age, diameter, grade, etc.) and the microarrays variables (mRNA expressions for each gene) are offered as one data set to the Bayesian network learning algorithm. The structure is learned for the combined data set:

$$p(S^{cm}_{K2}|D^{cm}) \propto p(D^{cm}|S^{cm}_{K2})P(S^{cm}_{K2}) \tag{5}$$

using equation 2 to calculate the right hand side. Next, the parameters are learned by updating the Dirichlet priors using the data, D^{cm}:

$$p(\theta_{ij}|D^{cm}, S^{cm}_{K2}) = \\ Dir(\theta_{ij}|N'_{ij1} + N^{cm}_{ij1}, ..., N'_{ijr_i} + N^{cm}_{ijr_i}) \tag{6}$$

In this manner the developed model can contain any type of relationship between the clinical variables and the microarray variables.

3.2 Decision integration

The opposite is a weak integration of the two data sources and is called decision integration. This method starts with learning a Bayesian network structure for both data sources using K2 (S^c_{k2} and S^m_{k2}). Followed by updating the Dirichlet priors with the data (D^c and D^m):

$$p(S^c_{K2}|D^c) \propto p(D^c|S^c_{K2})P(S^c_{K2}) \tag{7a}$$

$$p(S^m_{K2}|D^m) \propto p(D^m|S^m_{K2})P(S^m_{K2}) \tag{7b}$$

$$p(\theta_{ij}|D^c, S^c_{K2}) \\ = Dir(\theta_{ij}|N'_{ij1} + N^c_{ij1}, ..., N'_{ijr_i} + N^c_{ijr_i}) \tag{7c}$$

$$p(\theta_{ij}|D^m, S^m_{K2}) = \\ Dir(\theta_{ij}|N'_{ij1} + N^m_{ij1}, ..., N'_{ijr_i} + N^m_{ijr_i}) \tag{7d}$$

where equation 7b and equation 7b correspond with structure learning and are calculated using equation 4. Equation 7c and 7d correspond to parameter learning for both structures (S^c_{K2} and S^m_{K2}) separately. Then the probabilities predicted for the outcome variable by both models are combined using the weight parameter w:

$$p(Out|D^c, D^m) = \\ wp(Out|S^c_{K2}, \theta^c) + (1 - w)p(Out|S^m_{K2}, \theta^m) \tag{8}$$

where Out stands for the outcome variable and w is the weight parameter. θ^c and θ^m correspond to the complete set of parameters of the clinical model and microarray model respectively. The weight parameter is trained using only the model building data set (see section 2.4) of each randomization, in the context of decision integration called an outer randomization. This is done by performing again 100 inner randomizations of the model building data set within each outer randomization. For each inner randomization the weight is increased from 0.0 to 1.0 in steps of 0.1. Then the weight value with the highest average AUC over the 100 inner randomizations is chosen as weight for the outer randomization.

3.3 Partial integration

Bayesian networks also allow a third method, which we call partial integration. This is due to the fact that learning Bayesian networks is a two step process (see section 2.2). Therefore we can perform the first step, structure learning, separate for both data sources:

$$p(S^c_{K2}|D^c) \propto p(S^c_{K2})p(D|S^c_{K2}) \tag{9a}$$

$$p(S^m_{K2}|D^m) \propto p(S^m_{K2})p(D|S^m_{K2}) \tag{9b}$$

where equations 9b and 9b are again calculated according to equation 2. This results in a structure for the clinical data and a structure for the microarray data. These structures have only one variable in common: the outcome variable. Therefore we can join both structures using this variable. This combined structure will not contain an edge between a clinical variable on the one hand and a microarray variable on the other. Both structures are linked only through the outcome variable. Then the second step of learning

Bayesian networks (i.e. parameter learning) starts as if the structure was learned as a whole:

$$p(\theta_{ij}|D^m, S_{K2}^{c+m}) = Dir(\theta_{ij}|N'_{ij1} + N_{ij1}^{cm}, ..., N'_{ijr_i} + N_{ijr_i}^{cm}) \quad (10)$$

where S_{K2}^{c+m} is the combined structure. The parameter learning thus proceeds as normal because this step is independent of how the structure was built. Partial integration thus forbids links between both data sources. The developed model can now be used for classification.

4. DATA

4.1 Description

We used the data of (van 't Veer *et al.*, 2002) available at http://www.rii.com/publications/default.htm. This data set consists of two groups of patients. The first group of patients, which we call the training set, consists of 78 patients of which 34 patients belonged to the poor prognosis group and 44 to the good prognosis group. The second group of patients, the test set, consists of 19 patients of which 12 patients belonged to the poor prognosis group and 7 to the good prognosis group. DNA microarrays analysis was used to determine the mRNA expression levels of approximately 25000 genes for each patient. Every tumour sample was hybridized against a reference pool made by pooling equal amounts of RNA from each patient. The ratio of the sample and the reference was used as a measure for the expression of the genes and they constitute the microarray data set. Each patient also had the following clinical variables recorded: age, diameter, tumor grade, oestrogen and progesterone receptor status, the presence of angioinvasion and lymphocytic infiltration, which together form the clinical data.

4.2 Preprocessing

The microarray data consists of approximately 25000 expression values per patient, which was already background corrected, normalized and log-transformed. An initial selection was done (similar to (van 't Veer *et al.*, 2002)) by removing the genes that did not meet the following criteria using only the training data: at least a twofold increase or decrease and a P-value of less than 0.01 in more than 3 tumors. This resulted in a subset of approximately 5000 genes. Then we calculated the correlation between the expression values of these genes with the binary outcome and selected the genes with a correlation of ≥ 0.3 or ≤ -0.3. This resulted in 232 genes that where correlated with the outcome. Missing values were estimated using a 15-weighted nearest neighbours algorithm (Troyanskaya *et al.*, 2001). Then these genes were discretized into three categories: baseline, overexpression or under-expression according to two thresholds. These thresholds depended on the variance of the gene such that a gene with high variance receives a higher threshold than a gene with low variance. The data set that results from these steps was used as input for the Bayesian network software.

5. RESULTS

Model building was done as described in section 2.4 for the three integration methods (full, partial and decision integration) and for the clinical and microarray data separately for comparison. Table 1 shows the average AUCs for the developed models. Partial integration is significantly different from both data sources separately and full integration (P-value < 0.001, Wilcoxon rank sum test) and not significantly different from decision integration (P-value=0.0686, Wilcoxon rank sum test).

Table 1. Average AUC performance and standard deviation of the three methods for integrating clinical and microarray data and each data source separately with 100 randomizations on the training data.

Method	AUC	Std
Clinical data	0.751	0.086
Microarray data	0.750	0.073
Decision integration	0.773	0.071
Partial integration	0.793	0.068
Full integration	0.747	0.099

Next, both decision integration and partial integration were chosen and 100 models were built using the complete training data. Then the best performing model for each method was used to predict the outcome on the test data. In case of decision integration, the weight parameter was trained on the training data, similar to the inner randomizations as described in section 3.2. This resulted in a weight of 0.6 for the probability of the outcome predicted by the clinical model and thus a weight of 0.4 for the probability of the outcome predicted by the microarray model, slightly favouring the clinical model. Table 2 shows the AUC of these two models on the test set and the number of patients assigned a poor prognosis in: the test set, the set of true poor prognosis patients and the set of true good prognosis patients.

6. CONCLUSION

We have developed Bayesian networks to integrate clinical and microarray data. As an example we

Table 2. The AUC and the number of patients assigned a poor prognosis for the complete test set and for the true poor and good prognosis patients using the test set.

	AUC (std)	Total test set n=19	Relapse n=12	Disease free n=7
Partial integration†	0.845 (0.132)	13/19	11/12	2/7
Decision integration†	0.810 (0.118)	11/19	9/12	2/7

† The operating point is determined by maximizing the sum of the sensitivity and specificity on the training set.

used the data of (van 't Veer *et al.*, 2002) and investigated if an improvement was made for the prediction of recurrence in breast cancer. We investigated three methods for integrating the clinical and microarray data with Bayesian networks: full integration, partial integration and decision integration. Table 1 showed that partial integration and decision integration perform significantly better than full integration and each data source separately. We believe that this is due to the different nature of the data sources. Clinical data has a low noise level, in most cases there are fewer variables than observations and there are both discrete and continuous-valued variables. Microarray data on the other hand has a much higher noise level. There are much more variables than observations and all the variables are continuous. Therefore, it could be better to treat them separately in some way when the amount of data is limited. Partial integration uses separate structure learning while decision integration builds separate models but fuses the outcome probabilities. Full integration does not make a distinction between these two heterogeneous data sources. Both data sources are combined into one data set and used for Bayesian network learning. Because there are much more microarray variables than clinical variables (232 vs. 8), the chance that a clinical variable is added as a parent is small. The clinical variables are submerged by the microarray variables and mostly have few connections. Therefore full integration behaves similar to using only the microarray data (see table 1).
Table 2 showed that partial integration performed better than decision integration on the test set. We believe this is due to the fact that partial integration uses combined parameter learning. This integrates the clinical and microarray variables at the parameter level instead of at the decision level.

We have shown that both data sources are complementary and that an integrated approach can improve the prediction of the prognosis of breast cancer. Therefore this approach is promising for the use of Bayesian networks to integrate data sources for other types of cancer and data. When more data become available, the developed models can be prospectively validated. Finally, moving towards integrating several independently gathered data sources is necessary to increase the reliability of models based on microarray data.

7. ACKNOWLEDGMENTS

This research is supported by: the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen), research Council KUL: GOA AMBioRICS, IDO (Genetic networks), several PhD/postdoc, fellow grants Flemish Government: - FWO: PhD/postdoc grants, projects G.0115.01 (microarrays/oncology), G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.02 41.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), research communities (ICCoS, ANMMM, MLDM); - IWT: PhD Grants, GBOU-McKnow (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope, Silicos; Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling, 2002-2006) EU-RTD: FP5-CAGE (Compendium of Arabidopsis Gene Expression); ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain

REFERENCES

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

Heckerman, D., D. Geiger and D.M. Chickering (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20**, 197–243.

Huang, C. and A. Darwiche (1996). Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning* **15**(3), 225–263.

Neapolitan, R.E. (2004). *Learning Bayesian networks*. Prentice Hall. Upper Saddle River, NJ.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers. San Matteo, California.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman (2001). Missing value estimation methods for dna microarrays. *Bioinformatics* **17**, 520–525.

van 't Veer, L., H. Dai, M.J. van de Vijver, U.D. He, A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards and S.H. Friend (2002). Gene expression profiling predicts clinical outcome in breast cancer. *Nature* **415**, 530–536.