Optimizing variable order, selection and cost using a genetic algorithm for modelling ovarian masses with Bayesian networks

Olivier Gevaert, Dirk Timmerman and Bart De Moor

Abstract—**The pre-operative assessment of adnexal ovarian masses is still a major challenge for the gynecologist. This assessment is critical because the rupture of a stage I ovarian cancer during operation is life threatening and should be done non-invasively since approximately 75% of patients present with benign disease. The IOTA project aims to address this challenge by collecting data on 68 variables in a standardized manner and using methods from statistics and artificial intelligence to build models. An important problem however is to take into account the cost of variables that are used in a model while keeping the predictive performance intact. The cost is defined here to reflect a combination of the subjectivity, the financial cost and the time cost of each variable. In this contribution, we propose to tackle the problem of finding a model containing a set of low cost variables with high predictive performance using a combination of Bayesian networks and genetic algorithm optimization. The resulting model was used to predict the malignancy of ovarian masses which illustrated that our method effectively reduces the total cost of all variables in the model while keeping the drop in performance small. The model with and without cost optimization had an Area Under the ROC curve (AUC) on the test set of 0.958 and 0.966 respectively. We conclude that we have developed a method which concurrently optimizes predictive performance and cost of the selected variables.**

Index Terms—**Bayesian networks, genetic algorithm, ovarian cancer.**

I. Introduction

THE pre-operative assessment of adnexal ovarian masses is still a major challenge for the gynecologist. This assessment is critical because the rupture of a stage I ovarian cancer during operation is life threatening and should be done non-invasively since approximately 75% of patients present with benign disease [1], [2]. In 1998, the International Ovarium Tumor Analysis (IOTA) project was initiated for the pre-operative characterization of ovarian tumors based on artificial intelligence methods to provide solutions for this problem. The IOTA project started with the collection of a data set containing in total 68 variables related to adnexal ovarian masses in a standardized way from nine centers in five countries. The first phase of the IOTA data collection finished in 2002 and resulted in a data set consisting of 1066 patients. A number of models based on this data set have been proposed [1], [3]. An important challenge that remains unanswered is how to address variables that have a high cost associated with them. The cost of a variable is here defined to reflect the subjectivity, financial or time cost associated with it. For example, a well known biomarker for ovarian cancer is CA125 which has a financial cost and a large time cost because this test takes a few days to process and the patient has to return for consultation. Models based on this biomarker thus have an increased cost. Age, on the other hand is a variable which has no time cost, no financial cost and is relatively objective and thus has a low cost. It remains an open question however if models can be built that disregard high cost variables automatically without seriously harming predictive performance [4]. This problem can have tremendous consequences in three areas. First, much money can be saved by eliminating costly tests which would have significant consequences for health care budgets. Secondly, the patient would be aided much faster decreasing any emotional or physical stress. Finally, the model would be more objective and thus potentially generalizes easier to new populations.

The use of mathematical modeling for the prediction of diagnosis or prognosis of complex diseases has a history of more than three decades [5], [6] and has gained widespread use [1], [7], [8]. When the number of variables and the number of data points increase, it becomes hard to interpret the data manually. Mathematical models can aid clinicians to interpret the data by learning from the data and building models to predict the clinical outcome of future patients. A few examples of often used mathematical or statistical methods are, among others, logistic regression, support vector machines or Bayesian networks [1], [7], [8], [9]. These and other methods provide an unbiased way of analyzing clinical data and allow modeling many variables. Additionally, some of these methods (e.g. support vector machines (SVMs) or Bayesian networks), allow to model complex (i.e. non-linear) relationships in the data.

In this contribution, we propose to tackle the problem of reducing the total cost of variables participating in a mathematical model with high predictive performance using a combination of Bayesian networks and genetic algorithm optimization. Based on previous work in our group [9] a Bayesian network is used to model the malignancy of adnexal masses based on clinical data from the IOTA project. Next, to reduce the cost of the variables in the model we use a powerful optimization approach which is inspired by nature. A Genetic Algorithm (GA) mimics natural evolution by representing the model in genetic form and defining appropriate cross-over

O. Gevaert and B. De Moor are with the Department of Electrical Engineering (ESAT-SCD-Sista), K.U.Leuven, Leuven, Belgium.
E-mail: see http://www.esat.kuleuven.be/sista/ D. Timmerman is with the Department of Obstetrics and Gynecology, University Hospitals, K.U.Leuven, Leuven, Belgium.
Manuscript received ???, 2008; revised ??, 2008.

and mutation operators. Starting from an initially random population, the population evolves by applying these operators on the most "it" models. The cost of the variables is integrated in the GA by defining the fitness of a model based on its predictive performance but penalized by the total cost of the participating variables. Optimizing the set of variables used as input for Bayesian network modeling in this manner also provides opportunities to concurrently optimize variable order and variable selection. The order of the variables is important since it imposes restrictions on the search space of possible Bayesian networks and therefore reduces the computational cost for learning the structure of a Bayesian network. Moreover variable selection also directly contributes to optimize variable cost since removing variables from the model will obviously decrease cost.

We illustrate our method on the IOTA data set by predicting the malignancy of ovarian masses. Our results show that our method significantly decreases the cost while keeping the decrease in predictive performance minimal. We conclude that our approach allows to concurrently optimize variable cost, selection and order using a combination of GA optimization and Bayesian network modeling.

II. Materials and Methods

A. IOTA data

We developed our methodology to facilitate the assessment of ovarian masses through mathematical modeling. The IOTA data set consists of 68 variables that were gathered during phase 1 of the IOTA study. These variables include medical history (e.g. personal and family history of breast and ovarian cancer, age, etc.), serum tumor marker CA125, Color Doppler imaging and blood flow indices, ultrasonography, morphology and echogenicity [1]. Based on results from previous studies and with the help of a medical expert 35 variables were selected that were most relevant to the domain. These 35 variables are the starting point of the Bayesian network learning algorithms. We used for patients with masses in two ovaries the dominant tumor. Furthermore patients who were pregnant and patients with missing values were excluded from the analysis. This resulted in a data set of 812 patients which was split up in a training set (568 patients) and test set (244 patients). Subsequently, we asked an expert to score each variable according to subjectivity, financial cost and time cost. A low score means an easy-to-measure, objective or cheap variable while a high score means a hard-to-measure, subjective or expensive variable. An example of the former is age while an example of the latter is the CA125 biomarker as explained in the introduction.

B. Bayesian networks

1) Definition: A Bayesian network consists of two parts: a network structure and local probability models. The network structure is a directed acyclic graph where the nodes in the graph represent the variables and the edges between nodes represent (in)dependencies between the variables. Typically each variable x_i is linked with only a few other variables and therefore has only a limited number of parents (i.e. Pa_i). A Bayesian network thus is a sparse way of representing the joint probability distribution. The local probability models specify how the Pa_i influence x_i. There exist different flavors of local probability models depending on the nature of the variables. Variables can be discrete, continuous or a mixture of both and each type demands different local probability models. In this paper we focus on discrete-valued Bayesian networks since a significant portion of the variables is discrete. Moreover discrete-valued Bayesian networks allow to represent arbitrary discrete probability distributions with a limited number of parameters. The remaining continuous variables are discretised into bins manually according to an expert. The local probability models are then represented by Conditional Probability Tables (CPTs). Such a table specifies the probability that a variable takes a certain value given the value of its parents. Each Bayesian network structure implicitly constrains the ordering of the variables. A directed edge is only allowed from x_i to x_j if x_i is before x_j in the ordering. Therefore based on the chain rule of probability, a Bayesian network can be represented as:

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

2) Model building: Previously we mentioned that a Bayesian network consists of two parts. Consequently, there are two steps to be performed during model building: structure learning and learning the parameters of the CPTs.

The structure is learned by initializing a search procedure which generates new structures that are evaluated by a scoring metric. In the case of discrete valued Bayesian networks a Bayesian scoring metric is the so-called Bayesian Dirichlet metric (BD)[10]:

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

with N_{ijk} the number of cases in the data set D having variable i in state k associated with the j-th instantiation of its parents in current structure S. n is the total number of variables. Next, N_{ij} is calculated by summing over all states of a variable:

$$N_{ij} = \sum_{k=1}^{r_i} N_{ijk} \tag{3}$$

N'_{ijk} and N'_{ij} have similar meanings but refer to prior knowledge for the parameters. When no knowledge is available they are estimated using $N'_{ijk} = N/(r_i q_i)$ [10] with N the equivalent sample size, r_i the number of states of variable i and q_i the number of instantiations of the parents of variable i. $\Gamma(.)$ corresponds to the gamma distribution. Finally $p(S)$ is the prior probability of the structure which is defined un-informatively in the absence of prior information on the Bayesian network structure.

Next, the search procedure that was used is the K2 algorithm which greedily searches the model space [10], [11]. The disadvantage of K2 is that it needs a prior ordering of the variables before structure learning can start. K2 uses this ordering to put an extra constraint on the structure learning procedure and reduce the search space. A variable in the order can only have another variable as parent if it precedes the variable in the ordering. However, as already mentioned in the introduction, here the genetic algorithm will additionally optimize the variable order while also optimizing variable selection and variable cost. This is explained in more detail in section II-C.

The second step, parameter learning, consists of estimating the parameters of the local probability models corresponding with the dependency structure. We already 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:

$$p(\theta_{ij}|S) = Dir(\theta_{ij}|N'_{ij1}, ..., N'_{ijk}, ..., N'_{ijr_i}) \qquad (4)$$

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 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:

$$p(\theta_{ij}|D, S) = Dir(\theta_{ij}|N'_{ij1} + N_{ij1}, ..., N'_{ijk} + N_{ijk}, ..., N'_{ijr_i} + N_{ijr_i}) \qquad (5)$$

with N_{ijk} defined as before. We summarized this posterior by taking the Maximum A Posteriori (MAP) parametrization of the Dirichlet distribution and used these values to fill in the corresponding CPTs for every variable. Using MCMC could improve our current set-up because this technique allows devising the complete posterior distribution however at a higher computational cost [12].

3) Classification: After learning a Bayesian network according to the two-steps described previously, this model can used for prediction of blinded variables. In the context of Bayesian networks this is called inference. We used the probability propagation in tree of cliques algorithm [13] to predict the state of the outcome variable (i.e. the diagnosis of ovarian masses). Subsequently, the models were then evaluated using the Area Under the ROC curve (AUC) for predicting the outcome which will be used to calculate the fitness of an ordered set of variables as will be explained in more detail in the next section.

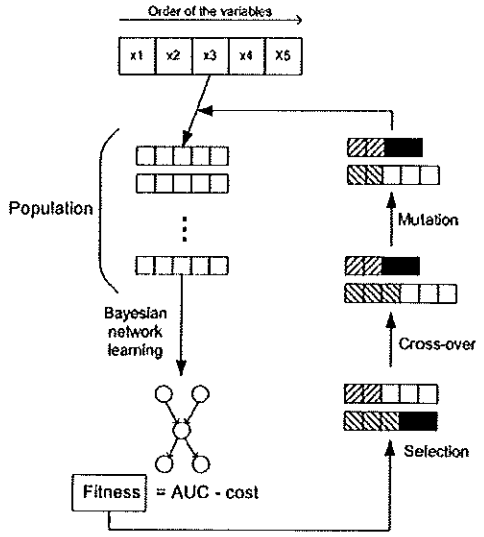


Fig. 1. Overview of the genetic algorithm. AUC stands for Area Under the ROC curve.

C. Genetic algorithm

Genetic algorithms (GAs) are optimization algorithms from the class of evolutionary computation that are inspired by evolutionary biology. GAs mimic natural evolution using a genetic representation of a mathematical model for the purpose of solving an optimization problem. After thorough repetition of the process of evolution through cross-over and mutation on an initial randomly assembled population, the GA evolves to more fit models [14].

A genetic algorithm requires two definitions: a representation of the problem domain in genetic form and a function estimating the fitness of each individual. In our setting, the model consisted of an ordered set of variables which was represented in genetic form by an ordered list. The fitness of each model was determined by calculating the predictive performance of this ordered set of variables, penalized by the sum of the cost of each variable that was present in this model. The predictive performance was estimated by calculating the AUC of the Bayesian network built on such an ordered set of variables as explained earlier in section II-B3.

The cross-over operator was implemented in a straightforward manner by preferentially selecting the best models (i.e. with highest fitness) and by randomly choosing a single crossover point to generate offspring models. The mutation operator was defined by its ability to remove or add variables, and it was allowed to randomly switch two variables in the ordering. Figure 1 shows an overview of the GA and both operators. To summarize, the total cost of all variables participating in the model was optimized through the definition of the fitness, while the variable order and variable selection were optimized by the GA operators.

D. Performance evaluation

Before starting model building, the data set (812 samples) was split into a training set used to develop a model and a test

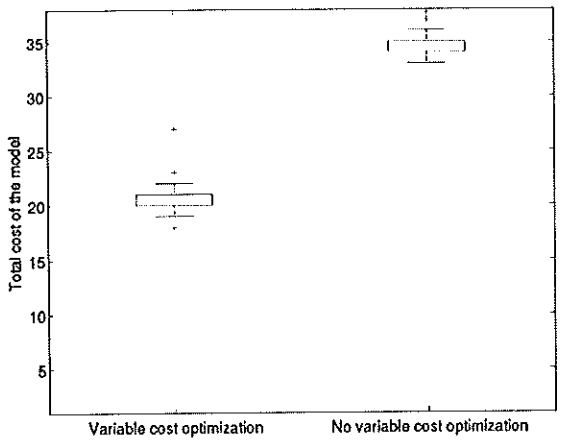


Fig. 2. Box plot of the total cost of models that are built with and without cost optimization.

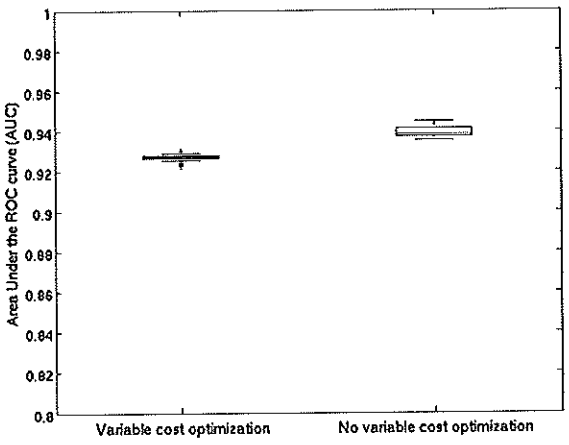


Fig. 3. Box plot of the total cost of models that are built with and without cost optimization.

set to estimate the generalization performance. The training set contained the first 70% of the samples (568 samples of which 179 malignant samples or 31.5%) and the test set had the remaining 30% of the samples (244 samples of which 65 malignant samples or 26.6%). The genetic algorithm combined with Bayesian network learning was then run with an initial population of 100 randomly drawn ordered lists of variables with and without cost optimization. The genetic algorithm is allowed to run as long as no significant improvement is made compared to the previous population.

III. Results

THE model building strategy was repeated 100 times on the IOTA data set with and without cost optimization. This allowed us to estimate the ability to reduce cost by comparing the best models from all runs with and without cost optimization. Figure 2 shows a boxplot of the cost of the 100 models that were built with and without cost optimization. The cost of models when no distinction is made between high or low cost variables is on average 34 compared to 20 when cost optimization is performed. Next, we investigated the predictive performance of both settings. The average training set AUC with and without cost optimization was 0.927 and 0.939 resp. Figure 3 shows a boxplot of the training set performance of both models. Since overfitting can be an issue, we also investigated the test set performance. Subsequently the average AUC of both settings was calculated on the independent test set. This resulted in an average test set AUC of 0.958 with cost optimization and 0.966 without cost optimization (see Figure 4).

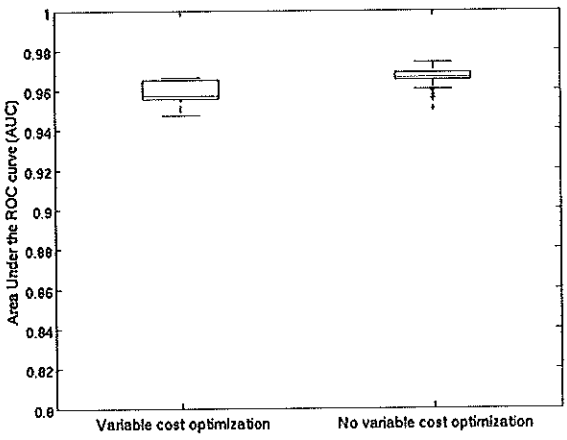


Fig. 4. Box plot of the total cost of models that are built with and without cost optimization.

IV. Conclusion

TO conclude, we have offered a method to take into account financial cost, time cost and objectivity of the variables in a mathematical model which we summarized as "variable cost" In our opinion, few research has been done in this area while reducing the cost of a model, as we defined it, could have a major impact on health care budgets, patient stress and generalization of models to new centers. Our results show that the cost of variables can be efficiently integrated when building models to aid biomedical decision support. Genetic Algorithm optimization and Bayesian networks can be effectively combined to create a powerful modeling strategy for cost-effective medical decision support where cost is interpreted in the broad sense as defined earlier. More specifically when applied to the IOTA data set, our method allows to significantly reduce the cost of the variables in the model with only a small decrease in predictive performance on both the training and test set. Additionally we want to stress that our methodology is sufficiently general to be applied to other clinical data sets.

Acknowledgments

This research is supported by the Institute for the Promotion of Innovation through Science and Technology in

Flanders (IWT-Vlaanderen) and by: Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, IDO (Genetic networks), several PhD/postdoc and fellow grants, Flemish Government: FWO: PhD/postdoc grants, projects 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.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel) research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modelling, 2002-2006) ; EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain

REFERENCES

[1] D. Timmerman, A. C. Testa, T. Bourne, E. Ferrazzi, L. Ameye, M. L. Konstantinovic, B. Van Calster, W. P. Collins, I. Vergote, S. Van Huffel, and L. a. Valentin, "Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group." *J Clin Oncol*, vol. 23, no. 34, pp. 8794–8801, December 2005. [Online]. Available: http://dx.doi.org/10.1200/JCO.2005.01.7632

[2] I. Vergote, J. De Brabanter, A. Fyles, K. Bertelsen, N. Einhorn, P. Sevelda, M. E. Gore, J. Kaern, H. Verrelst, K. Sjövall, D. Timmerman, J. Vandewalle, M. Van Gramberen, and C. G. Tropé, "Prognostic importance of degree of differentiation and cyst rupture in stage i invasive epithelial ovarian carcinoma." *Lancet*, vol. 357, no. 9251, pp. 176–182, January 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11213094

[3] B. Van Calster, D. Timmerman, I. T. Nabney, L. Valentin, C. Van Holsbeke, and S. Van Huffel, "Classifying ovarian tumors using bayesian multi-layer perceptrons and automatic relevance determination: A multi-center study." *Conf Proc IEEE Eng Med Biol Soc*, vol. 1, pp. 5342–5345, 2006. [Online]. Available: http://dx.doi.org/10.1109/IEMBS.2006.260118

[4] B. Van Calster, D. Timmerman, T. Bourne, A. C. Testa, C. Van Holsbeke, E. Domali, D. Jurkovic, P. Neven, S. Van Huffel, and L. Valentin, "Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum ca-125." *J Natl Cancer Inst*, vol. 99, no. 22, pp. 1706–1714, November 2007. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/18000221

[5] M. Stefanelli, "The socio-organizational age of artificial intelligence in medicine." *Artif Intell Med*, vol. 23, no. 1, pp. 25–47, August 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11470215

[6] E. W. Coiera, "Artificial intelligence in medicine: the challenges ahead." *J Am Med Inform Assoc*, vol. 3, no. 6, pp. 363–366, 1996. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/8930853

[7] D. Smet, F., D. Brabanter, J., V. den Bosch, T., Pochet, N., Amant, F., V. Holsbeke, C., Moerman, P., D. Moor, B., Vergote, I., Timmerman, and D., "New models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography," *Ultrasound in Obstetrics and Gynecology*, vol. 27, no. 6, pp. 664–671, June 2006. [Online]. Available: http://dx.doi.org/10.1002/uog.2806

[8] O. Gevaert, F. De Smet, E. Kirk, B. Van Calster, T. Bourne, S. Van Huffel, Y. Moreau, D. Timmerman, B. De Moor, and G. Condous, "Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression," *Human reproduction*, vol. 21, 2006.

[9] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor, "Using literature and data to learn bayesian networks as clinical models of ovarian tumours." *Artif Intell Med*, vol. 30, pp. 257–81, 2004.

[10] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

[11] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, 1992.

[12] R. M. Neal, *Bayesian learning for neural networks*. New York: Springer-Verlag, 1996.

[13] C. Huang and A. Darwiche, "Inference in belief networks: A procedural guide," *International Journal of Approximate Reasoning*, vol. 15, no. 3, pp. 225–263, 1996.

[14] M. Mitchell, *An Introduction to Genetic Algorithms (Complex Adaptive Systems)*. The MIT Press, February 1998.