2010 Ninth International Conference on Machine Learning and Applications

Predicting local failure in lung cancer using Bayesian networks

Jung Hun Oh, Jeffrey Craft, Rawan Al-Lozi, Manushka Vaidya,
Yifan Meng, Joseph O Deasy, Jeffrey D Bradley, and Issam El Naqa
Department of Radiation Oncology
Washington University School of Medicine
MO 63110, USA
Email: {joh, ielnaqa}@radonc.wustl.edu

***Abstract*—Despite various efforts to develop new predictive models for early detection of tumor local failure in locally advanced non-small cell lung cancer (NSCLC), many patients still suffer from a high local failure rate after radiotherapy. Based on recent studies of biomarker proteins' role in predicting tumor response following radiotherapy, we hypothesize that incorporation of physical and biological factors with a suitable framework could improve the overall prediction. To this end, we propose a graphical Bayesian network framework for predicting local failure in lung cancer. The proposed approach was tested using a dataset of locally advanced NSCLC patients treated with radiotherapy. This dataset was collected prospectively, which consisted of physical variables and blood-based biomarkers. Our experimental results demonstrate that the proposed method can be used as an efficient method to develop predictive models of local failure in these patients and to interpret relationships among the different variables. The combined model of physical and biological factors outperformed individual physical and biological models, achieving an accuracy (acc) of 87.78%, Matthew's correlation coefficient (r) of 0.74, and Spearman's rank correlation coefficient (rs) of 0.75 on leave-one-out cross-validation analysis.**

I. Introduction

Lung cancer is a leading cause of cancer death worldwide [1]. Of all lung cancer cases, non-small cell lung cancer (NSCLC) accounts approximately for 80%. For treating patients with advanced and inoperable stage, a combination of chemotherapy and radiotherapy is mainly used instead of surgical resection [1]. However, patients with locally advanced NSCLC following radiotherapy suffer from a high local failure rate [2]. Despite many efforts to improve treatment outcomes, a low two-year local control rate as low as 27% in these patients requires innovative diagnostic and prognostic models to improve early detection of tumor local failure [3].

In our previous work [4], we used various approaches to evaluate linear and nonlinear models for prediction of tumor local control in lung cancer. In this work, we propose a novel method for modeling local failure of patients with locally advanced NSCLC using Bayesian networks. A Bayesian network has been used as a useful tool to create individualized predictive models due to its several attractive characteristics, which have led to various studies successfully in the field of oncology. Recently, Jayasurya *et al.* proposed a Bayesian network model for survival prediction in lung cancer patients [5]. They also showed that the Bayesian network can be efficiently used when handling missing data compared with other machine learning techniques. Velikova *et al.* designed a multi-view mammographic analysis system using a Bayesian network framework to detect breast cancer and demonstrated the potential of the system for selecting the most suspicious cases [6]. Chen *et al.* proposed an effective Bayesian structure learning method based on the mutual information and K2 algorithm to reconstruct reliable gene networks [7]. van Gerven *et al.* demonstrated the development of a prognostic model for carcinoid patients using dynamic Bayesian networks [8]. Armañanzas *et al.* used a hierarchical Bayesian structure learning method to detect gene interactions [9]. Smith *et al.* developed a prognostic model for prostate cancer with intensity modulated radiation therapy (IMRT) plans and calculated a quality-adjusted life expectancy for each plan using Bayesian networks [10].

The aim of this study is to develop an efficient modeling method to predict local failure in lung cancer post-radiotherapy treatment using Bayesian networks. Through our experiments, we show that incorporating physical and biological factors into the Bayesian network can further improve the predictive power. It is our expectation that the proposed model will help physicians better predict early recurrence in lung cancer and lead to more individualized radiotherapy prescriptions.

II. Bayesian network

A Bayesian network is a probabilistic graphical model that encodes a joint probability distribution among variables of interest [11], [12], [13], [14]. A Bayesian network forms a directed-acyclic graph (DAG) by a set of nodes and a set of directed edges. Given n variables, $X = \{X_1, X_2, \cdots, X_n\}$, the joint probability distribution can be decomposed into a product form of conditional probability distributions:

$$P(X) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \qquad (1)$$

where $Pa(X_i)$ indicates the set of parents of X_i in the Bayesian network.

Learning a Bayesian network structure is a task to find a DAG that best represents the dataset. In Bayesian structure learning, a challenging problem is to identify an optimal

978-0-7695-4300-0/10 $26.00 © 2010 IEEE
DOI 10.1109/ICMLA.2010.112

Bayesian network structure among all possible network structures. Since the search space increases super-exponentially as the number of nodes increases, heuristic search strategies, including hill climbing and K2 are alternatively utilized. The K2 algorithm employs a greedy search strategy, which dramatically reduces the computational complexity in learning the Bayesian network structure by using a prior ordering of all the nodes [15]. Therefore, using a K2 strategy, a correct node ordering is very important for successful Bayesian structure learning.

III. Materials and Methods

In this study, we tested the proposed Bayesian structure learning method with a dataset collected at Mallinckrodt Institute of Radiology. The description of the dataset is followed by our proposed approach.

A. Dataset

This dataset was collected prospectively and was approved by the Human Research Protection Office at our institute. In this protocol, blood sera were drawn at pre-treatment and mid-treatment of radiotherapy NSCLC patients in addition to collecting gross tumor volume (GTV) and percentage volumes receiving x% radiation dose (Vx). A total of 18 patients were evaluable for the current study. This dataset allows for unique integration of physical and biological variables. After post-evaluation, the patients were divided into a local failure group ($n = 8$) and a control group ($n = 10$). With this dataset, the goal is to investigate interactions among heterogeneous physical and biological variables and to evaluate the complementary role of these variables in improving the prediction of tumor local control in NSCLC patients post-radiotherapy treatment.

From the blood sera, four selected biomarker proteins were extracted. The biomarkers were selected because of their potential role in tumor response in lung cancer. The four chosen biomarker proteins are as follows: transforming growth factor β (TGF-β), interleukin-6 (IL-6), angiotensin converting enzyme (ACE), and osteopontin (OPN). These blood-based candidate proteins were selected based on previous reports linking their serum expression to tumor response post-radiotherapy treatment [16], [17], [18].

We measured expressions of these proteins using enzyme-linked immunosorbent assay (ELISA). As physical variables, two variables (V75 and GTV) were chosen from the top model in our previous study [4]. In addition, tumor regression volume on 4D-CT images between pre-treatment and mid-treatment was measured using an active contour tracking algorithm [19].

B. Preprocessing

In general, prior to Bayesian structure learning and evaluation of Bayesian classifiers, variables are discretized into two or three bins. For the determination of discretization boundaries, researchers have used the mean and standard deviation or the equal width. In both cases, however, it is difficult to find the optimal boundaries. Recently, Kuschner *et al.* proposed an efficient binning method based on mutual information using a three-bin strategy, where bin boundaries are determined such that the mutual information of each variable given the treatment response class is maximized [20]. As a result, each variable is discretized into three bins: high, medium, and low. We applied the same approach for discretization to our dataset.

C. Markov Chain Monte Carlo (MCMC) algorithm

To efficiently search the space of Bayesian network structures, a Markov Chain Monte Carlo (MCMC) method based on the Metropolis-Hastings algorithm was applied using the Bayes Net Toolbox (BNT) for Matlab, which rapidly converges to a locally optimal structure [21]. We accumulated the Bayesian structure information of each MCMC run as follows. Let e_{ij} be the arc from variable X_i to variable X_j. Then, we define a_{ijk} as

$$a_{ijk} = \begin{cases} 1 & \text{if } e_{ij} \text{ is present in the } k\text{th MCMC run,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The number of occurrences of an arc e_{ij} over m runs of MCMC is

$$f_{ij} = \sum_{k=1}^{m} a_{ijk}. \tag{3}$$

Therefore, a matrix M that contains the number of occurrences for all edges can be expressed as

$$M = \begin{bmatrix} 0 & f_{12} & \dots & f_{1n} \\ f_{21} & 0 & \dots & f_{2n} \\ & & \vdots & \\ f_{n1} & f_{n2} & \dots & 0 \end{bmatrix}. \tag{4}$$

D. Maximum spanning tree

After several runs of MCMC algorithm, a weighted directed graph represented by the matrix M was populated. We applied a directed maximum spanning tree algorithm based on Chu-Liu/Edmonds's method to the weighted directed graph [22], [23]. As a result, a directed graph was obtained, from which we gained a node ordering that can be used as an input to the K2 algorithm.

E. K2 algorithm

As mentioned earlier, exhaustive search over the space of the DAGs is computationally impractical. Therefore, greedy searches such as K2 are typically used. The K2 algorithm is one of the most frequently used Bayesian structure learning methods [15]. For the successful Bayesian structure learning with K2 algorithm, knowing the node ordering is essential. In this study, we proposed a novel method for finding the node ordering using methods described above.

F. Classification

A Bayesian network can also be used for evaluation of classification performance. Suppose that for a test sample, the values of all nodes but the class label node c are known. Then, the Bayesian classifier assigns the test sample to a class with the highest posterior probability that is mathematically determined according to Bayes' theorem [24]:

$$c^* = \text{argmax}_c \; p(c|X_1, X_2, \cdots, X_n). \tag{5}$$

G. High-confidence Bayesian network

For unbiased evaluation of the proposed method, we performed the proposed algorithm q times using r-fold cross validation and the results were averaged. As a result, we attained $q \times r$ Bayesian network structures. To find a high-confidence Bayesian network, we employed a hierarchical structure learning method proposed by Armañanzas and his colleagues [9]. Similar to Eqs. (2) and (3), we define the following equations:

$$b_{ijk} = \begin{cases} 1 & \text{if } e_{ij} \text{ is present in the } k\text{th-induced graph,} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and

$$g_{ij} = \sum_{k=1}^{q \times r} b_{ijk}. \tag{7}$$

As the $g_{ij}(\leq q \times r)$ is larger, the relationship between variable X_i and variable X_j is more likely to be reliable. Let t be the confidence threshold and let L_t denote the set of edges that meet the following condition:

$$L_t = \{e_{ij} | g_{ij} \geq t\}. \tag{8}$$

Starting from the maximum confidence threshold, $t_{\max} = \max\{g_{ij}\}$ for $i, j \in \{1, \ldots, n\}$, we attempt to build a Bayesian network structure. Decreasing the confidence threshold by 1, we keep building the Bayesian network connecting the corresponding edges while avoiding cyclic pitfalls. This process is continued until a predefined confidence threshold is reached.

IV. Results

A. Experimental results of the proposed method

As performance evaluation metrics, in addition to the Spearman's rank correlation (rs) coefficient that was used in our previous work [4], we also employed accuracy (acc) and Matthew's correlation coefficient (r) that are respectively calculated as follows:

$$acc = \frac{TP + TN}{TP + FN + TN + FP}, \tag{9}$$

$$r = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

where TP and TN are the number of patients correctly classified in the local failure and control group, and FN and FP are the number of patients falsely classified in the local failure and control group, respectively.

We conducted experiments with the dataset described in Section III-A. The dataset consisted of physical variables and four biomarker proteins. For physical variables we used two variables (V75 and GTV) selected from our top logistic regression analysis in addition to volume regression between pre- and mid-treatment. To test the importance of each type of variables in prediction of local failure, we generated three different sub-datasets each of which consisted of physical variables, biomarker proteins, and physical variables+biomarker proteins. However, the small number of samples ($n = 18$) in this dataset might lead to biased estimates. As an alternative to overcome this small sample size problem, we applied the following resampling scheme. First, leave-one-out cross-validation (LOO-CV) was performed in which at each iteration, one out of the 18 available samples was reserved for testing, while using resampling with replacement from the remaining 17 samples, $2 \times (n - 1)$ bootstrap samples ($b = 34$) were randomly generated for training. Then, with these resulting testing and training sets, 5-fold cross validation was repeated. Finally, we iterated this procedure 20 times and averaged the results.

As can be seen in Fig. 1, when both physical variables and biological proteins were used, a slightly better performance (with a classification accuracy of 87.78%, $r = 0.7396$, and $rs = 0.7512$) was obtained compared to an accuracy of 86.11%, $r = 0.7042$, and $rs = 0.7168$ and an accuracy of 85.00%, $r = 0.6933$, and $rs = 0.6946$ with biomarker proteins and physical variables alone, respectively. From these results, it is implied that the contribution of biomarker proteins in classification of this dataset is considerable. Figure 2 illustrates a Bayesian network structure with the high-confidence dependency to predict the local failure in lung cancer using heterogeneous variables. From this figure, it is observed that overall the physical variables affect protein expressions as would be expected.

B. Evaluation of strategy effects on Bayesian structure learning algorithm

Our Bayesian structure learning algorithm was designed based on MCMC and K2 algorithms, in which the MCMC algorithm is used to obtain a node ordering that is fed as an input into the K2 algorithm. To test the role of K2 algorithm, we excluded it from the process and the results were summarized in Fig. 3. That is, the figure shows the results when the DAG obtained by the directed maximum spanning tree algorithm was used only. For all cases, the performance was degraded considerably (particulary, when biomarker proteins were used) compared to that of the proposed method (an accuracy of 87.78%, r=0.7396, and rs=0.7512 using all variables).

In addition, to evaluate the effect of using the collective information from MCMC simulation, we tested the proposed method using the last model (Bayesian structure) after the completion of MCMC simulation. As can be seen in Fig. 4, it was observed that the performance in all cases was somewhat degraded against the proposed method, but better than that shown in Fig. 3.

V. Conclusions

We proposed a novel Bayesian structure learning method for building predictive models of tumor local failure post-radiotherapy in locally advanced NSCLC patients. We demonstrated that the proposed method has the potential to improve the prediction power of tumor local failure. We also showed that the Bayesian network can be used as a useful tool to identify interactions among heterogeneous variables. Through our experimental results, it was observed that the integration

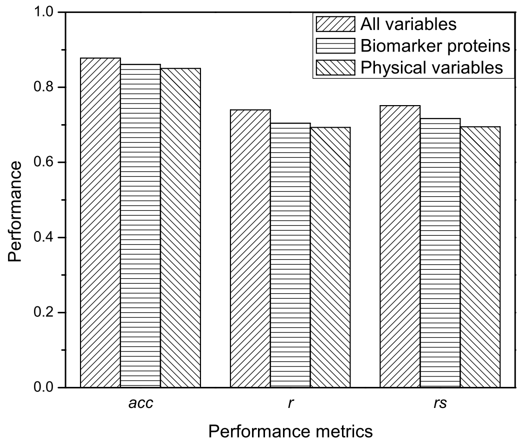

Fig. 1. Performance measurements obtained using different kinds of variables.

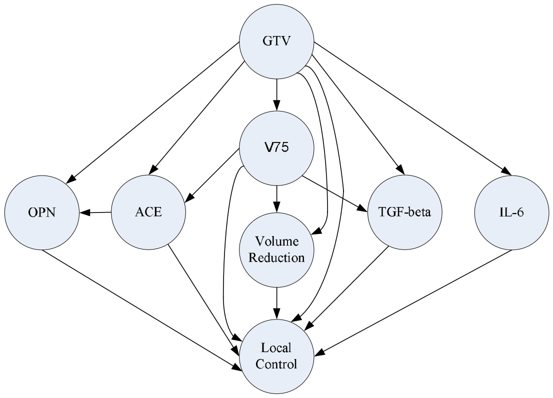

Fig. 2. A Bayesian network constructed using combined biomarker proteins and physical variables.

of heterogeneous variables could provide a better performance than using physical variables or biomarker proteins only. However, evaluation on larger datasets would be required to elucidate these observations.

Acknowledgment

This work was supported by NIH grant K25CA128809.

References

[1] American Cancer Society: Cancer Facts and Figures. Atlanta, GA: American Cancer Society, 2008.

[2] J.G. Armstrong, M.J. Zelefsky, S.A. Leibel, C. Burman, C. Han, L.B. Harrison, G.J. Kutcher, and Z.Y. Fuks, "Strategy for dose escalation using 3-dimensional conformal radiation therapy for lung cancer," *Ann. Oncol.*, vol. 6, pp. 693-697, 1995.

[3] A. Abramyuk, S. Tokalov, K. Zöphel, A. Koch, K. Szluha Lazanyi, C. Gillham, T. Herrmann, and N. Abolmaali, "Is pre-therapeutical FDG-PET/CT capable to detect high risk tumor subvolumes responsible for local failure in non-small cell lung cancer?," *Radiother. Oncol.*, vol. 91, pp. 399-404, 2009.

[4] I. El Naqa, J.O. Deasy, Y. Mu, E. Huang, A.J. Hope, P.E. Lindsay, A. Apte, J. Alaly, and J.D. Bradley, "Datamining approaches for modeling tumor control probability," *Acta Oncol.*, 2010.

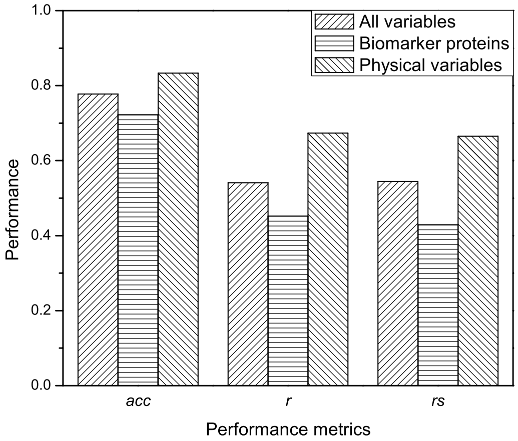

Fig. 3. Performance measurements obtained without using K2 algorithm.

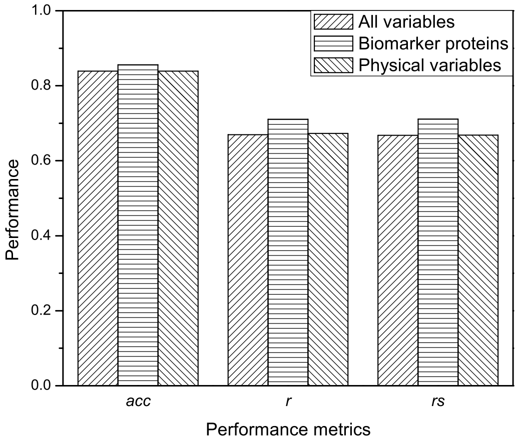

Fig. 4. Performance measurements obtained using the last model after the completion of MCMC simulation.

[5] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, and A.L.A.J. Dekkera, "Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy," *Med. Phys.*, vol. 37, pp. 1401-1407, 2010.

[6] M. Velikova, M. Samulski, P.J.F. Lucas, and N. Karssemeijer, "Improved mammographic CAD performance using multi-view information: a Bayesian network framework," *Phys. Med. Biol.*, vol. 54, pp. 1131-1147, 2009.

[7] X. Chen, G. Anantha, and X. Wang, "An effective structure learning method for constructing gene networks," *Bioinformatics*, vol. 22, pp. 1367-1374, 2006.

[8] M.A. van Gerven, B.G. Taal, and P.J. Lucas, "Dynamic Bayesian networks as prognostic models for clinical patient management," *J. Biomed. Inform.*, vol. 41, pp. 515-529, 2008.

[9] R. Armañanzas, I. Inza, and P. Larrañaga, "Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers," *Comput. Methods Progr. Biomed.*, vol. 91, pp. 110-121, 2008.

[10] W. Smith, J. Doctor, J. Meyer, I. Kalet, and M. Philips, "A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model," *Artif. Intell. Med.*, vol. 46, pp. 119-130, 2009.

[11] S. Sarkar and K.L. Boyer, "Integration, inference, and management of spatial information using Bayesian networks: perceptual organization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, pp. 256-274, 1993.

[12] D. Heckerman and J.S. Breese, "Causal independence for probability

assessment and inference using Bayesian networks," *IEEE Trans. Syst. Man Cybern. A*, vol. 26, pp. 826-831, 1996.

[13] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, pp. 131-164, 1997.

[14] P.J.F. Lucas, "Bayesian network modelling through qualitative pattern," *Artif. Intell.*, vol. 163, pp. 233-263, 2005.

[15] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Mach. Learn.*, vol. 7, pp. 309-347, 1992.

[16] A.S. Varela and J.J.B.L. Saez, "Utility of serum activity of angiotensin-converting enzyme as a tumor marker," *Oncology*, vol. 50, pp. 430-435, 1993.

[17] Q.T. Le, E. Chen, A. Salim, H. Cao, C.S. Kong, R. Whyte, J. Donington, W. Cannon, H. Wakelee, R. Tibshirani, J.D. Mitchell, D. Richardson, K.J. O'Byrne, A.C. Koong, and A.J. Giaccia, "An evaluation of tumor oxygenation and gene expression in patients with early stage non-small cell lung cancers," *Clin. Cancer Res.*, vol. 12, pp. 1507-1514, 2006.

[18] C.E. R*ü*be, J. Palm, M. Erren, J. Fleckenstein, J. K*ö*nig, K. Remberger, and C. R*ü*be, "Cytokine plasma levels: reliable predictors for radiation pneumonitis?," *PLoS One*, vol. 3:e2898, 2008.

[19] I. El Naqa, A. Apte, D. Yang, C. Noel, J. Bradley, and J.O. Deasy, "A robust approach for estimating tumor volume change during radiotherapy of lung cancer," *Med. Phys.*, vol. 35, pp. 2956, 2008.

[20] K.W. Kuschner, D.I. Malyarenko, W.E. Cooke, L.H. Cazares, O.J. Semmes, and E.R. Tracy, "A Bayesian network approach to feature selection in mass spectrometry data," *BMC Bioinformatics*, vol. 11:177, 2010.

[21] K. Murphy, Bayesian Network Toolbox (BNT) http://www.cs.ubc.ca/∼murphyk/Software/BNT/bnt.html, 2007.

[22] Y.J. Chu and T.H. Liu, "On the shortest arborescence of a directed graph," *Science Sinica*, vol. 14, pp. 1396-1400, 1965.

[23] J. Edmonds, "Optimum branchings," *J. Res. Nat. Bur. Standards*, vol. 71B, pp. 233-240, 1967.

[24] F. Pernkopf and P. O'Leary, "Floating search algorithm for structure learning of Bayesian network classifiers," *Pattern Recognit. Lett.*, vol. 24, pp. 2839-2848, 2003.