35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Comparison of robustness against missing values of alternative decision tree and multiple logistic regression for predicting clinical data in primary breast cancer

Masahiro Sugimoto, Masahiro Takada, Masakazu Toi

Abstract— **Nomogram based on multiple logistic regression (MLR) is a standard technique for predicting diagnostic and treatment outcomes in medical fields. However, the applicability of MLR to data mining of clinical information is limited. To overcome these issues, we have developed prediction models using ensembles of alternative decision trees (ADTree). Here, we compare the performance of MLR and ADTree models in terms of robustness against missing values. As a case study, we employ datasets including pathological complete response (pCR) of neoadjuvant therapy, one of the most important decision-making factors in the diagnosis and treatment of primary breast cancer. Ensembled ADTree models are more robust against missing values than MLR. Sufficient robustness is attained at low boosting and ensemble number, and is compromised as these numbers increase.**

I. Introduction

Decision-making in the diagnosis and treatment of breast cancer is becoming increasingly complex as medical examination technologies advance, and multiple adjuvant therapies become available. Prior to surgery, neoadjuvant chemotherapy (NAC) is administrated to reduce the tumor size (thereby preserving breast tissue) and to assess chemosensitivity. The latter assists the design of post-operative adjuvant therapy [1]. The NAC response must be evaluated by low-invasive techniques such as imaging. This goal has inspired the development of many mathematical prediction models.

Predicting the pathological complete response (pCR) of NAC is regarded as a binary classification problem whose outcomes comprise pCR or non-pCR. The input variables are clinical features such as tumor size, age, and estrogen receptor expression status. The associations among these variables are commonly quantified by multiple logistic regression (MLR) models. Nomogram is a visualized representation of a prediction model. In the clinical breast cancer setting, many MLR nomograms are available, which not only facilitate MLR calculations of MLR but also establish quantitative relationships for such factors as chemotherapy sensitivity [2-4], non-sentinel metastasis in patients with sentinel-positive status [5-7], prognosis-specific to triple-negative subtype [8], and risk of arm lymphedema after axillary lymph node (AxLN) resection [9] in patients with primary breast cancer. However, for generality and to prevent over-fitting, the MLR incorporates only a few independent variables as predictive features, which limits its prediction accuracy.

Clinical problems can also be solved by machine learning techniques, but these approaches require careful consideration of problem-specific data. For example, missing values are common in datasets, especially in retrospectively collected clinical data. An if-then type decision tree, usually built by the ID3 or C4.5 algorithm, is preferred because it enables clinical or biological validation of the model as well as its statistical validation. However, such models are unsuitable for data with missing values because they cannot predict the probability of belonging to a class in many cases. Therefore, these models cannot inherently predict the outcome of an incomplete dataset, which is a serious disadvantage compared with MLR. The same disadvantage occurs in classification and regression tree (CART) models [10].

A common technique applied to binary problems is Bayesian network (BN), which maintains interpretability throughout the model development. Other popular methods are artificial neural networks (ANN), and support vector machines (SVM). These methods are frequently more accurate than BN because they are generalizable to non-linear problems, but are less interpretable. Parametric features of these methods are also problematic. Model development requires optimization of many parameters. To overcome these drawbacks, we have developed alternative prediction models to help decision-making in breast cancer treatment and diagnosis [11, 12]. The models, based on the alternative decision tree (ADtree) model [13], have been validated in predictions of axillary lymph node (AxLN) metastasis [12] and pCA after NAC in patients with primary breast cancer [11]. The tree is an improved-accuracy epigone of the if-then type decision tree [13] and processes more variables [11, 12]. In this study, we compare the robustness of ADtree and MLR to missing values in datasets.

*This study was supported by research grants from the Ministry of Health, Labour and Welfare ("A study on the construction of an algorithm for multimodal therapy with biomarkers for primary breast cancer by formulation of a decision-making process" led by MT, No. H18-3JIGAN-IPPAN-007, "Reduction and lowering of recurrence risk, toxicity and pharmacoeconomic cost by prediction of efficacy for anticancer agents in breast cancer patients" led by MT; No. H22-GANRINSHO-IPPAN-039), and JSPS KAKENHI Grant Number 2471025 led by MS.

M. Sugimoto is with the Institute for Advanced Biosciences, Keio University, 246-2, Mizukami, Kakuganji, Tsuruoka, Yamagata 997-0052, Japan (e-mail: msugi@sfc.keio.ac.jp).

M. Takada and M. Toi are with the Department of Breast Surgery, Graduate School of Medicine, Kyoto University, 54 Kawaracho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan (e-mail: toi@kuhp.kyoto-u.ac.jp and masahiro@kuhp.kyoto-u.ac.jp).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

II. Data description

The previously published datasets used in our study are provided in [11] and are merely summarized here. Data from 150 patients were collected from multiple institutes, including Tokyo Metropolitan Cancer and Infectious Diseases Centre at Komagome Hospital, Osaka National Hospital and Tsukuba University Hospital (here named data 1). These data were used to train the prediction model. We also collected 173 patient data from the Organisation for Oncology and Translational Research OOTR N003 trial (Niigata Cancer Centre Hospital, National Kyushu Cancer Centre and Aichi Cancer Centre) as independent validation datasets (named data 2). The number of missing values depended on the patient data; see [11] for details.

All patients received the same neoadjuvant; four courses of FEC (5-fluorouracil 500 mg/m^2, epirubicin 100 mg/m^2 and cyclophosphamide 500 mg/m^2, i.v., every 3 weeks) followed by four courses of docetaxel (75 mg/m^2, i.v., every 3 weeks) with or without capecitabine (1,650 mg/m^2/day, oral administration, for 14 days every 3 weeks).

Patients whose tumor had reached 5 cm and who had completed 75% of the planned NAC courses were eligible for the study. The data included 28 clinicopathological variables, such as histological type, estrogen receptor (ER) status, HER2 status, histological/nuclear grade, and ultrasound imaging findings. The study protocol was approved by the institutional review board of Kyoto University Hospital.

III. Methodology

A. Comparison of ADTree and MLR for missing value robustness

The ADTree-based prediction model used in the robustness test had been trained using data 1 [11]. Here, the model building procedures are summarized. The model parameters were optimized based on 10-fold cross-validations (CVs) using a typical training/test ratio (90%/10%). Prediction performance (ability of the model to correctly discriminate pCR from non-pCR) was evaluated by area under the receiver operating characteristics curve (AUC). Following training, the model was validated on test data using a parameter set. This procedure was repeated 10 times, such that all patients had been included in the training dataset at least once. The CV was repeated 200 times and the parameters evaluated from the averaged AUC. The above procedure was repeated for all possible parameter sets. The parameters yielding the best AUC values in CV tests were adopted in subsequent analysis. The parameters, which include number of nodes in the ADTree model (boosting number), the number of trees, and a random seed to generate multiple cohorts for the ensemble, are described in [11].

The model was constructed by ensemble techniques used for combining ADTree models and 19 ADTrees, allocating three nodes per tree. Number of variables was 15. The AUC values were 0.766 ($P < 0.0001$) using data 1 and 0.787 ($P < 0.0001$) when validated with data 2.

MLR models were also developed using data 1 and yielded 0.754 ($P = 0.00019$) after validation with data 2. Four features for MLR were selected by stepwise forward selection ($P < 0.2$ as each new feature was added). After eliminating data without missing values, data 1 and data 2 comprised 121 and 172 elements, respectively.

To evaluate robustness against missing values, artificial datasets were prepared by replacing each value with a random value. Numerical features were assigned a randomly generated value within the actual data range. For ordinary scale and nominal scale features, randomly generated possible scales were assigned. In this way, 200 sets of both training and validation datasets were generated. Ready-developed ADtree-based and MLR models were validated on these generated datasets.

B. Effect of ensemble and boosting of prediction model on missing value robustness

To gauge the effect of ensemble and boosting number on robustness to missing values, we changed the numbers of trees and nodes, and repeated the analyses described in Section *A*.

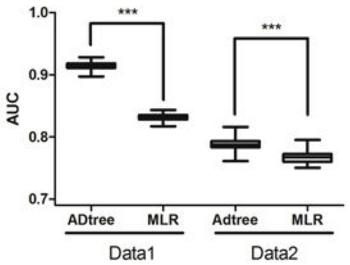


Figure 1 AUC values of ADtree and MLR models using data 1 and data 2. Each box-whisker plot includes 200 AUC values. The horizontal bars indicate (from top to bottom) the maximum, quartile, median, third-quartile, and minimum. The asterisks indicate $P < 0.0001$.

IV. Results and discussion

Figure 1 shows the distribution of AUC values yielded by the ADTree-based model (here condensed to ADTree) and by MLR. For both datasets, the AUC values of ADtree are significantly higher than those of MLR ($P < 0.0001$; Student's *t*-test). For the validation datasets (data 2), ADTree and MLR yielded 0.783 (95% CI: 0.786 – 0.789) and 0.767 (95% CI: 0.765 – 0.770), respectively. The upper and lower 95% CI differs by 0.003 for ADTree and 0.05 for MLR; therefore, the AUC values of ADTree are slightly more consistent than those of MLR.

Figure 2 plots the AUC value of ADTree models as a function of ensemble number. The boosting number (3) had been previously optimized by CV using data 1. In both cases (Fig. 2A and 2B, displaying results for data 1 and data 2, respectively), the AUC values are minimized when a single tree is used, i.e. without ensemble techniques. Moreover, in both cases, the overall AUC values dramatically increase for small bagging number, indicating that ensembling enhances the robustness against missing values more efficiently when the ensemble number is small. This trend is particularly obvious for data 2 (Fig. 2B); the AUC values dramatically increase at ensemble number 4 and decrease slightly thereafter. Thus, a few ensemble numbers sufficiently enhance the robustness against missing values.

A similar analysis was conducted on boosting number. The number of nodes in a tree was varied while the ensemble number was fixed at 19 (like the fixed boosting number, this number had been previously optimized by CV of data 1). As is evident in Fig. 3, as the boosting number increases, the overall AUC values uniformly increase for data 1, possibly indicating over-fitting. By contrast, for data 2, the highest AUC is attained for three nodes; additional nodes exert little influence on AUC values. In both cases, the AUC variability, i.e. the difference between maximum and minimum AUC values, decreases as boosting number increases, indicating that the model becomes less sensitive to missing values at higher boosting number.

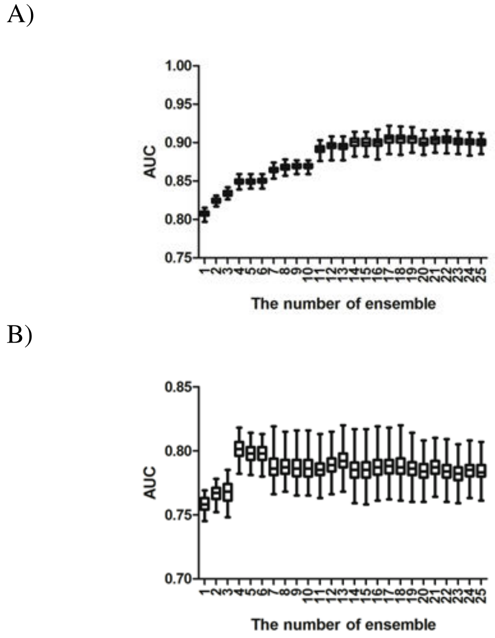
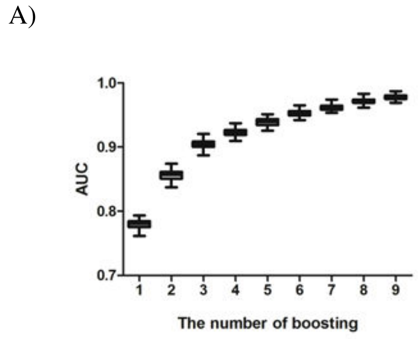


Figure 2 AUC values of ADtree model, obtained by varying the ensemble number at fixed boosting number for **(A)** data 1and **(B)** data 2. The *x*-axis indicates the ensemble number i.e. the number of trees in a model. The AUC values and meaning of the horizontal bars in the box-whisker plots are as described in Figure 1.



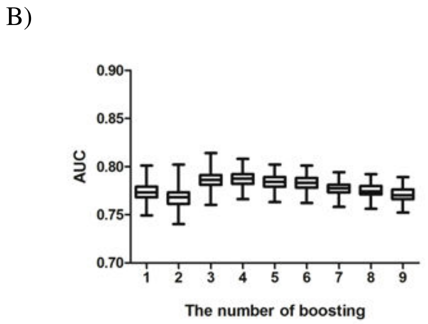


Figure 3 AUC values of ADtree model, obtained by varying the boosting number at fixed ensemble number for (**A**) data 1 and (**B**) data 2. The *x*-axis indicates the boosting number i.e. the number of nodes in a model. The AUC values and meaning of the horizontal bars in the box-whisker plots are as described in Figure 1.

V. Conclusion

In this study, we compared the ensemble ADTree and MLR models in terms of robustness against missing values, using NAC-administrated breast cancer data. Although many studies have compared the accuracy among different data mining methods, insensitivity to missing values is also important, especially in data mining of retrospectively collected medical datasets, which frequently contain missing values. In fact, since the missing values might not be randomly missing, predictive ability could be enhanced by identifying the type of missing data and appropriately estimating their values. However, because of the complex structure of clinical datasets, we instead evaluated the model on randomly replaced values, an approach that is applicable to any missing values. The ensembled ADtree yielded higher overall accuracy (higher AUC values) than MLR. In addition, sufficiently high robustness was attained at low ensemble and boosting number.

References

[1] J.S. Mieog, J.A. van der Hage, and C.J. van de Velde, “Preoperative chemotherapy for women with operable breast cancer,” *Cochrane database of systematic reviews (Online)*, (no. 2), pp. CD005002, 2007.

[2] R. Rouzier, L. Pusztai, J.R. Garbay, S. Delaloge, K.K. Hunt, G.N. Hortobagyi, D. Berry, and H.M. Kuerer, “Development and validation of nomograms for predicting residual tumor size and the probability of successful conservative surgery with neoadjuvant chemotherapy for breast cancer,” *Cancer*, vol. 107, (no. 7), pp. 1459-66, Oct 1 2006.

[3] R. Rouzier, L. Pusztai, S. Delaloge, A.M. Gonzalez-Angulo, F. Andre, K.R. Hess, A.U. Buzdar, J.R. Garbay, M. Spielmann, M.C. Mathieu, W.F. Symmans, P. Wagner, D. Atallah, V. Valero, D.A. Berry, and G.N. Hortobagyi, “Nomograms to predict pathologic complete response and metastasis-free survival after preoperative chemotherapy for breast cancer,” *J Clin Oncol*, vol. 23, (no. 33), pp. 8331-9, Nov 20 2005.

[4] A. Frati, E. Chereau, C. Coutant, C. Bezu, M. Antoine, J. Chopier, E. Darai, S. Uzan, J. Gligorov, and R. Rouzier, “Comparison of two nomograms to predict pathologic complete responses to neoadjuvant chemotherapy for breast cancer: evidence that HER2-positive tumors need specific predictors,” *Breast Cancer Res Treat*, vol. 132, (no. 2), pp. 601-7, Apr 2012.

[5] T. Sasada, T. Kataoka, H. Shigematsu, N. Masumoto, T. Kadoya, M. Okada, and H. Ohdan, “Three models for predicting the risk

of non-sentinel lymph node metastasis in Japanese breast cancer patients," *Breast Cancer*, Jan 10 2013.

[6] K.J. Van Zee, D.M. Manasseh, J.L. Bevilacqua, S.K. Boolbol, J.V. Fey, L.K. Tan, P.I. Borgen, H.S. Cody, 3rd, and M.W. Kattan, "A nomogram for predicting the likelihood of additional nodal metastases in breast cancer patients with a positive sentinel node biopsy," *Ann Surg Oncol*, vol. 10, (no. 10), pp. 1140-51, Dec 2003.

[7] M.C. Specht, M.W. Kattan, M. Gonen, J. Fey, and K.J. Van Zee, "Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: clinicians versus nomogram," *Ann Surg Oncol*, vol. 12, (no. 8), pp. 654-9, Aug 2005.

[8] C. Mazouni, F. Spyratos, S. Romain, F. Fina, P. Bonnier, L.H. Ouafik, and P.M. Martin, "A nomogram to predict individual prognosis in node-negative breast carcinoma," *European journal of cancer*, vol. 48, (no. 16), pp. 2954-61, Nov 2012.

[9] J.L. Bevilacqua, M.W. Kattan, Y. Changhong, S. Koifman, I.E. Mattos, R.J. Koifman, and A. Bergmann, "Nomograms for predicting the risk of arm lymphedema after axillary dissection in breast cancer," *Ann Surg Oncol*, vol. 19, (no. 8), pp. 2580-9, Aug 2012.

[10] H.E. Kohrt, R.A. Olshen, H.R. Bermas, W.H. Goodson, D.J. Wood, S. Henry, R.V. Rouse, L. Bailey, V.J. Philben, F.M. Dirbas, J.J. Dunn, D.L. Johnson, I.L. Wapnir, R.W. Carlson, F.E. Stockdale, N.M. Hansen, and S.S. Jeffrey, "New models and online calculator for predicting non-sentinel lymph node status in sentinel lymph node positive breast cancer patients," *BMC Cancer*, vol. 8, pp. 66, 2008.

[11] M. Takada, M. Sugimoto, S. Ohno, K. Kuroi, N. Sato, H. Bando, N. Masuda, H. Iwata, M. Kondo, H. Sasano, L.W. Chow, T. Inamoto, Y. Naito, M. Tomita, and M. Toi, "Predictions of the pathological response to neoadjuvant chemotherapy in patients with primary breast cancer using a data mining technique," *Breast Cancer Res Treat*, vol. 134, (no. 2), pp. 661-70, Jul 2012.

[12] M. Takada, M. Sugimoto, Y. Naito, H.G. Moon, W. Han, D.Y. Noh, M. Kondo, K. Kuroi, H. Sasano, T. Inamoto, M. Tomita, and M. Toi, "Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model," *BMC Med Inform Decis Mak*, vol. 12, pp. 54, 2012.

[13] Y. Freund and L. Mason, "The Alternating Decision Tree Learning Algorithm," *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 124 - 133 1999.