Accurate prediction of neoadjuvant chemotherapy pathological complete remission (pCR) for the four sub-types of breast cancer

Xin Feng[1], Lelian Song[2], Shaofei Wang[3], Haoqiu Song[1,4], Hang Chen[3], Yuxuan Liu[1,3], Chenwei Lou[3], Jian Zhao[3], Quewang Liu[3], Yang Liu[5], Ruixue Zhao[3], Kai Xing[3], Sijie Li[2], Yunhe Yu[2], Zhenyu Liu[2], Chengyang Yin[2] , Bing Han[2] , Ye Du[2], Ruihao Xin[6], Lan Huang[3], Zhimin Fan[2], Fengfeng Zhou[3], Senior Member, IEEE.

[1] Cancer Systems Biology Center, The China-Japan Union Hospital, Jilin University, Changchun 130033, China.

[2] Department of Breast Surgery, The First Hospital of Jilin University, Changchun, Jilin, China, 130012.

[3] BioKnow Health Informatics Lab, College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012.

[4] School of Computer Science, Hubei University of Technology, China

[5] BioKnow Health Informatics Lab, College of Software, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012.

[6] College of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun, Jilin, China, 130012.

Corresponding author: Fengfeng Zhou (e-mail: FengfengZhou@gmail.com or ffzhou@jlu.edu.cn). Additional correspondence may also be addressed to Zhimin Fan (e-mail: fanzhimn@163.com).

Problem settings

Pathological complete remission (pCR) is a good measurement to describe whether a patient achieves favorable outcomes after a few chemotherapeutic treatments [1]. Chemotherapy itself has serious side effects including vomiting [2], seizure [3], and psychological distress [4], *etc*. So it’s essential to evaluate the prognosis before the chemotherapeutic treatments. The subtypes LBP and LBN have 17 and 15 positive samples (pCR=1), respectively. Another subtype HER2 has 23 positive samples (pCR=1). And the triple negative breast cancer (TNBC) has 24 positive samples.

This study tries to predict the measurement pCR using only the data of the first three chemotherapeutic treatments. Two problem settings were utilized for this binary classification problem, because the clinicians usually determined two values 0 or 1 for pCR. Firstly, a binary classification model was optimized using the breast cancer nodal sizes of the first three chemotherapeutic treatments. Secondly, the nodal sizes of the next three treatments were predicted using the regression algorithms and then a binary classification model of pCR was optimized using the real data of the first three treatments and the predicted data of the next three treatments.

Binary classification algorithms and performance measurements

Six representative classification algorithms were chosen to build the binary classification model of pCR, *i.e.*, SVM (Support Vector Machine), KNN (k nearest neighbors), NBayes (Naïve Bayes), DTree (Decision Tree), RF (Random Forest) and XGB (extreme gradient boosting).

SVM (Support Vector Machine) searches for a separating plane between two groups of samples in the space of the chosen features and is specially optimized for a binary classification problem [5]. KNN (k nearest neighbors) assigns a sample to the class with the largest number of neighbors among the top k nearest ones [6]. NBayes (Naïve Bayes) calculates the probability of a sample in each class and the final decision is determined by the highest probability [7].

Tree based classifiers have the inherent property of easy interpretabilities and DTree (Decision Tree) is the simplest tree based classifier [8]. RF (Random Forest) summarizes the decisions of multiple weak tree classifiers and usually generates a very good classification performance [9].

XGB (extreme gradient boosting) is a recently developed ensemble algorithm that intensively summarizes the decisions of multiple weak tree classifiers [10]. XGB achieved very high ranks in many recent machine learning competitions [11].

The binary classification model is usually evaluated by the following performance measurements [12-14]. The correct prediction percentages of positive and negative samples are evaluated by the sensitivity $Sn=TP/(TP+FN)$ and the specificity $Sp=TN/(TN+FP)$, respectively. The accuracy is defined as $Acc=(TP+TN)/(P+N)$. A balanced accuracy $Avc=(Sn+Sp)/2$ is defined to fairly evaluate an unbalanced dataset. The Matthew's Correlation Coefficient $MCC=(TP\times TN-FP\times FN)/sqrt((TP+FP)(TP+FN)(TN+FP)(TN+FN))$ is a measurement with [-1, 1], where *sqrt*() is the function squared root [15, 16]. The measurements *mAvc* and *mMCC* were defined as the maximal values of Avc and MCC among the six classification algorithms.

All the performance measurements were calculated by the stratified 5-fold cross validation strategy, and are better with larger values. The experiments were implemented by the Python version 3.6.1. The running environment is an Inspur Gene Server G100, with 256GB memory, 28 Intel Xeon® CPU cores (2.4 GHz), and 30TB RISC1 disk space.

References

[1] G. von Minckwitz, M. Untch, J.U. Blohmer, S.D. Costa, H. Eidtmann, P.A. Fasching, B. Gerber, W. Eiermann, J. Hilfrich, J. Huober, C. Jackisch, M. Kaufmann, G.E. Konecny, C. Denkert, V. Nekljudova, K. Mehta, S. Loibl, Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes, J Clin Oncol, 30 (2012) 1796-1804.

[2] G. Kvale, K. Hugdahl, A. Asbjørnsen, B. Rosengren, K. Lote, H. Nordby, Anticipatory nausea and vomiting in cancer patients, Journal of consulting and clinical psychology, 59 (1991) 894.

[3] J.A. Koekkoek, M. Kerkhof, L. Dirven, J.J. Heimans, J.C. Reijneveld, M.J. Taphoorn, Seizure outcome after radiotherapy and chemotherapy in low-grade glioma patients: a systematic review, Neuro Oncol, 17 (2015) 924-934.

[4] S.F. Chen, H.H. Wang, H.Y. Yang, U.L. Chung, Effect of Relaxation With Guided Imagery on The Physical and Psychological Symptoms of Breast Cancer Patients Undergoing Chemotherapy, Iran Red Crescent Med J, 17 (2015) e31277.

[5] L. Wang, Support vector machines: theory and applications, Springer Science & Business Media2005.

[6] O. Kramer, K-nearest neighbors, Dimensionality Reduction with Unsupervised Nearest Neighbors, Springer2013, pp. 13-23.

[7] J.K. Asafu-Adjei, R.A. Betensky, A Pairwise Naive Bayes Approach to Bayesian Classification, Intern J Pattern Recognit Artif Intell, 29 (2015).

[8] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE transactions on systems, man, and cybernetics, 21 (1991) 660-674.

[9] M. Pal, Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26 (2005) 217-222.

[10] T. Chen, T. He, M. Benesty, Xgboost: extreme gradient boosting, R package version 0.4-2, (2015) 1-4.

[11] M. Zięba, S.K. Tomczak, J.M. Tomczak, Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction, Expert Systems with

Applications, 58 (2016) 93-101.

[12] C. Xu, J. Liu, W. Yang, Y. Shu, Z. Wei, W. Zheng, X. Feng, F. Zhou, An OMIC biomarker detection algorithm TriVote and its application in methylomic biomarker detection, Epigenomics, 10 (2018) 335-347.

[13] Y. Ye, R. Zhang, W. Zheng, S. Liu, F. Zhou, RIFS: a randomly restarted incremental feature selection algorithm, Sci Rep, 7 (2017) 13013.

[14] M. Zhou, Y. Luo, G. Sun, G. Mai, F. Zhou, Constraint programming based biomarker optimization, Biomed Res Int, 2015 (2015) 910515.

[15] Z. Ju, J.J. He, Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection, Anal Biochem, 550 (2018) 1-7.

[16] S. Khurana, R. Rawi, K. Kunji, G.Y. Chuang, H. Bensmail, R. Mall, DeepSol: A Deep Learning Framework for Sequence-Based Protein Solubility Prediction, Bioinformatics, (2018).