Algorithm Discovery of the Multiple Biomarkers in Urine for Early Diagnosis of Ovarian Cancer

Hye-Jeong Song[1,3], Seung-Kyun Ko[2,3], Jong-Dae Kim[1,3], Chan-Young Park[1,3], Yu-Seop Kim[1,31]

[1] Dept. of Ubiquitous Computing, Hallym University, 1 Hallymdaehak-gil, Chuncheon,Gangwon-do, 200-702 Korea
[2]Dept. of Ubiquitous Game Engineering, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702, Korea chokood@nate.com
[3]Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon, South Korea
{ hjsong, kimjd, cypark, yskim01}@hallym.ac.kr

Abstract. This research evaluates the classification performance from benign tumor to cancer under Luminex environment of the optimum biomarker combinations that were selected by Random Forest, Logistic, Multilayer Percepron, Bagging, Classification Via Regression, LogitBoost, MultiClassifier, Simple Logistic, Logistic Regression. The Area Under the Curve (AUC) of each marker combination selected was compared. For the experimental data, total of 178 urine samples (benign tumor 121, cancer 57) were provided from two hospitals, and the concentrations of the 15 biomarkers were extracted using Luminex-PRA. In the experiment, we firstly select the best three marker combinations using logistic regression and then apply nine classification algorithms on the combinations to get higher AUC values. This paper shows that even marker combination selected by logistic regression improves its classification performance after applying other algorithms.
Keywords: Biomarker, Urine, Ovarian Cancer, Logistic Regression, Random Forest, Bagging, LogitBoost, Early Diagnosis

1 Introduction

Ovarian cancer is a malignant tumor frequently arising in the age between 50~70. According to the statistical results in 2002, about 1,000 to 1,200 new ovarian cancer patients are diagnosed ranked as the second most frequently occurring cancer in gynecology following cervical cancer [1]. It is evident that the development of a biomarker for early detection of the ovarian cancer has become paramount [2, 3]. The early stages of research had focused on a single biomarker for cancer diagnosis. Recent researches, however, focus on combining multiple biomarkers to diagnose cancer more efficiently. A new technology to find the right biomarker combinations is required, since the sensitivity and specificity has not yet reached a satisfactory level [4].

[1] Corresponding author

This paper aims to determine the optimum marker combination from 15 biomarkers using Random Forest [5], Logistic [6], Multilayer Perceptron [7], Bagging [8], Classification Via Regression [9], LogitBoost [10], MultiClassifier [11], Simple Logistic, and Logistic Regression [12]. The AUCs of the selected combinations were compared. We firstly find the best three combinations showing the highest AUC values by using Logistic Regression which is the most widely spread. Then we apply other classification algorithms to improve the accuracy. By doing this, we try to find possibility to apply another algorithm instead of the logistic regression.

2 Experiment

For this experiment, 178 (benign tumor 121, cancer 57) urine samples of Koreans were provided by two hospitals. The 15 biomarkers used in this paper are commonly discussed biomarkers in the ovarian cancer researches [13, 14]. This research aims to find another algorithm instead of logistic regression in determining the optimum marker combination from the detected biomarkers in the urine.

Three biomarker combinations were selected using logistic regression from fifteen biomarkers. The performance of the selected combination was compared with that of Random Forest, Logistic, Multilayer Perception, Bagging, Classification Via Regression, LogitBoost, MultiClassifier, Simple Logistic, and Logistic Regression by cross-validation. For the cross-validation, the 5-fold cross validation was used. By applying other algorithms on the combinations linear regression selects, we try to show the possibility of other algorithms.

3 Results

In the experiment, the AUCs [15] of the multi-biomarker combinations consisting of 2~4 biomarkers selected by logistic regression were obtained using the nine algorithms. In measuring the performance, the AUC of each algorithm classifying the benign and cancer was compared.

The markers that ought to be combined were limited to four, because the high cost to combine more than 4 markers will make it difficult to realize and commercialize the use of multi-biomarkers. Also to avoid the infringement of patent, the names of the markers are concealed.

Table 1 shows the best algorithms with their AUC values. The three rows mean the best three combinations given by Logistic Regression and three columns mean the number of biomarkers to be used. Numbers inside the parenthesis are AUC values. Interestingly, the three marker combinations of two markers chosen by logistic regression show the highest performance when applied to Bagging and Classification Via Regression algorithm. And in case of three markers combination, Random Forest shows the highest AUC value. From these results, we infer that another algorithm other than the Linear Regression can be utilized for the better performance.

Table 1. Results from Algorithm Discovery

	Two-Comb	Three-Comb	Four-Comb
1st Comb.	Bagging (0.862) Classification via Regression (0.862)	Random Forest (0.891)	Logistic Regression (0.885)
2nd Comb	Logistic(0.848)	Logistic Regression (0.867)	Logistic Regression (0.884)
3rd Comb	Logistic Regression (0.85)	Logistic Regression (0.861)	Logistic Regression (0.874)

4 Conclusion

This research determines the algorithm that finds the optimum biomarker combination from the multi-biomarkers extracted from urine for early diagnosis of ovarian cancer. Three combinations for each 2~4 biomarkers combined showing the highest performance were selected by logistic regression. From the combinations chosen by logistic regression, some of them showed higher performance when applied on different algorithms. This indicates that other algorithms rather than Logistic Regression can also be adopted in determining the optimum marker combination.

It is encouraged to carry on the same experiment with different algorithms such as Bagging, Classification via Regression, and Logistic Boost in the future to determine the optimum marker combination for early diagnosis of ovarian cancer. Also novel algorithms, apart from the algorithms proposed in this paper, which are found by the Machine Learning research, can be tested on the proposed experiment to evaluate their performance.

Acknowledgments. The research was supported by the Research & Business Development Program through the Ministry of Knowledge Economy, Science and Technology (N0000425) and the Ministry of Knowledge Economy(MKE), Korea Institute for Advancement of Technology(KIAT) and Gangwon Leading Industry Office through the Leading Industry Development for Economic Region.

Reference

1. Seoul National University Hospital, http://www.snuh.org/.
2. Asan Medical Center, http://medical.amc.seoul.kr/.
3. G. Yang, “A Cancer Risk Assessment System and Pedigree Information”, Journal of Korean Institute of Information Technology, 5(1), pp. 27-31, 2007. (in Korean)
4. C. Cho, “Biomarkers for the diagnosis of ovarian cancer and treatment”, Korean Society of Obstetrics and Gynecology, pp.36-40, 2011. (in Korean)
5. A. Liaw and M. Wiener, “Classification and Regression by Random Forest,” R News, 2(3), pp.18-22, 2002.

6. F. E. Harrell Jr., “Regression Modeling Strategies,” Springer, 2001.
7. S. K. Rogers, M. Kabrisky, M.E. Oxley and B.W. Suter, “The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function,” IEEE Transactions on Neural Networks, 1(4), pp.296-298, 1990
8. L. Breiman, “Bagging Predictors,” Machine Learning, 24, pp.123-140, 1996.
9. E. Frank, M. Hall, L. Trigg, G. Holmes and I.H. Witten, “Data Mining in Bioinformatics using Weka,” Bioinformatics, 20(15), pp.2479-2481, 2004.
10. Y.D. Cai, K.Y. Feng, W.C. Lu and K.C. Chou, “Using LogitBoost Classifier to Predict Protein Structural Classes,” Journal of Theoretical Biology, 238(1), pp.172-176, 2006.
11. A. Sicsu, L. Heutte, E. Menu, E. Lecolinet, O. Debon and J.V. Moreau, “A Multi-Classifier Combination Strategy for the Recognition of Handwritten Cursive Words,” In Procs. of the Second International Conference on Document Analysis and Recognition, 1993.
12. D. Freedman, R. Purves, and R. Pisani, “Statistics, 3rd Edition,” W. W. Norton & Company, 1998.
13. B. Nolen, A. Marrangoni, L. Velikokhatnya, D. Prosser, M. Winans, E. Gorelik, and A. Lokshin, “A Serum based Analysis of Ovarian Epithelial Tumorigenesis,” Gynecologic Oncology, 112(1), pp.47-54, 2009
14. S.D. Amonkar, G.P. Bertenshaw, T. Chen, K.J. Bergstrom, J. Zhao, P. Seshaiah, P. Yip and B.C. Mansfield, “Development and Preliminary Evaluation of a Multivariate Index Assay for Ovarian Cancer,” PLoS ONE, 4(2), e4599, 2009.
15. A.P. Baradley, “The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms,” Pattern Recognition, 33(7), pp.1148-1159, 1997.