Best Biomarker Combination for Ovarian Cancer

Yu-Seop Kim[1,3], Min-Ki Jang[2,3], Chan-Young Park[1,3], Hye-Jung Song[1,3], Jong-Dae Kim[1,3*]

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

Abstract. Logistic regression is employed to search the optimal combination of multiple markers that can discriminate ovarian cancers from benign by Luminex assay test of the patient sera. To verify the selection performance, three other classification methods were also tested including t-test, genetic algorithm, and random forest. The chosen combinations from each of the four methods were evaluated again with three classification methods: logistic regression, linear discriminant analysis, and k-nearest neighbor. The results show that the combination of 4 markers found by logistic regression had the best average accuracy.

Keywords: Biomarker, Luminex, Ovarian Cancer, Marker, T-Test, Logistic Regression, Genetic Algorithm, Random Forest, Linear Discriminant Analysis

1 Introduction

Ovarian cancer is a malignant tumor frequently arising at the age between 50 and 70. Early diagnosis is associated with a 92% 5-year survival rate, yet only 19% of ovarian cancers are detected in the early stage [1]. Therefore, early detection of ovarian cancer has great promise to improve clinical outcome. It is evident that the development of a biomarker for early detection of ovarian cancer has become paramount [2].

Biomarker consists of molecular information based on the pattern of a single or multiple molecules originating from DNA, metabolite, or protein. Biomarkers are indicators that can detect the physical change of an organism due to the genetic after required genetic change.

Along with the completion of the genome project, various biomarkers are being developed, providing critical clues for cancers and senile disorders.

* Corresponding author

The early stages of research focused on a single biomarker for cancer diagnosis. Recent researches focus on combining multiple biomarkers to diagnose cancer more efficiently. The improvement of the sensitivity and specificity in order to increase the accuracy of the diagnosis is of particular interest, and the commercialization of multi-biomarkers seems to be close at hand. However, a new technology to find the optimum biomarker combinations is required, since the accuracy has not yet reached a satisfactory level [3].

In this research, the relative fluorescence units of the biomarkers were obtained using Luminex [4]. Luminex follows the panel reactive antibody (PRA) and a solid phase-based method of Luminex corp. This paper explores the optimal marker combination for ovarian cancer diagnosis with logistic regression (LR) [5]. To validate the marker combination selection by LR, three other methods including t-test, genetic algorithm (GA), and random forest (RF) are also applied to find the optimal combination [6]-[8]. LR, k-nearest neighbor (k-NN), and linear discriminant analysis (LDA) were used to evaluate the classification accuracy of the optimal combinations to avoid the possible bias when applying only one evaluation method.

The data collection method and the experimental details are demonstrated in chapter 2. The results of the marker combinations and their classification performances are discussed in chapter 3, and chapter 4 presents the conclusion.

2. Method

The serum samples from 81 patients with ovarian cancer and 216 patients with benign pelvic masses were used. They were provided by Hallym University Medical Center and ASAN Medical Center and reacted with Luminex-beads attached with 8 biomarkers, and the florescence from the antibodies on the beads was measured. In order to equalize the range of the biomarker florescence, the florescence values of each biomarker were normalized to 0-1 based on their maximum and minimum values.

This paper conducts two experiments: (1) determination of biomarkers with LR, t-test, GA, and RF, and (2) performance comparison of the selected markers using LR, k-NN, and LDA.

The random tree creation for RF was 50, the k for k-NN was 3, and score threshold value for LR was 0.5. The combination of the biomarkers consisted of 4 markers, and 5-fold cross validation was conducted for evaluation.

The marker selection algorithms investigated all of the combinations consisting 4 biomarkers chosen out of 8 biomarkers and selected the most accurate combination.

3. Results

The experiment compares the difference in performance of the selected 4 multi-biomarkers by LR, T-Test, GA, and RF to that of the optimal combination amongst the total possible combinations of the markers. The sensitivity, specificity, and accu-

racy of the optimal combination from each selection algorithm were measured and evaluated with LR, LDA, and k-NN.

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

Table 1. Classification performance of the optimal marker combination obtained through LR (M1, M2, M6, M7).

Classifier	Sensitivity	Specificity	Accuracy
LR	0.4762	0.9500	0.8272
LDA	0.6286	0.8693	0.8067
k-NN	0.4429	0.9045	0.7844
Average			0.8061

Table 1 shows the optimal marker combination obtained through LR and their performance when applying LR, LDA, and k-NN to the combine the 4 markers. The best accuracy of 82.7% was seen in LR.

Table 2. Classification performance comparison of the optimal marker combination obtained through t-test (M3, M5, M6, M8).

Classifier	Sensitivity	Specificity	Accuracy
LR	0.6500	0.8103	0.7692
LDA	0.5385	0.7868	0.7252
k-NN	0.4462	0.9137	0.7977
Average			0.7640

Table 3. Classification performance comparison of the optimal marker combination obtained through GA (M1, M2, M7, M8).

Classifier	Sensitivity	Specificity	Accuracy
LR	0.2941	0.9167	0.7792
LDA	0.5254	0.8600	0.7838
k-NN	0.3220	0.9050	0.7722
Average			0.7784

Table 4. Classification performance comparison of the optimal marker combination obtained through RF (M1, M5, M6, M8).

Classifier	Sensitivity	Specificity	Accuracy
LR	0.1111	0.9153	0.7273
LDA	0.5085	0.8173	0.7461
k-NN	0.3559	0.9289	0.7969
Average			0.7568

4. Conclusions

This paper presents the exploration for the optimal biomarker combination using logistic regression that can easily distinguish the ovarian cancer to benign using logistic regression. To validate the proposed search method, three common methods were also applied to search the marker combination that delivered the most accurate results. The combinations found by the four methods were evaluated through three existing classification methods. The average accuracy over the classification methods was compared to prove the superiority of logistic regression over the other three search method. The experimental results show that logistic regression produced the greatest average accuracy, recommending logistic regression as the exploration tool for the optimum marker combination.

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.

References

1. American Cancer Society, http://www.cancer.org/Cancer/OvarianCancer (2012)
2. Brian, N., Adele, M., Liudmila. V., Denise, P., Matthew, W., Elesier, G., Anna L.: A serum based analysis of ovarian epithelial tumorigenesis. ELSEVIER Gynecologic Oncology 112, 47—54 (2009)
3. ChiHeum, C.: Biomarkers related to Diagnosis and Prognosis of Ovarian Cancer. Koean Journal of Obstetrics and Gynecology 39, 90—95 (2008)
4. SunKyung, J., EunJi O., ChulWoo, Y., WoongSik, A., YongGu, K., YeonJun, P., KyungJa, H.: ELISA for the selection of HLA isoantibody and Comparison Evaluation of Luminex Panel Reactive Antibody Test. Journal of Korean Society for Laboratory Medicine 29, 473—480 (2009)
5. David Freeman, Roger Purves, and Robert Pisani.: Statistics, 3rd Edition. W. W. Norton & Company (1998)
6. David, F., Roger, P., Robert, P.: Statistics, 3rd Edition. W. W. Norton & Company (1998)
7. Tom, M. M.: Machine Learning. The McGraw-Hill (1997)
8. Leo, B.: Random Forest. Machine Learning 45, 5—32 (2001)