Exploring Multiple Biomarker Combination by Logistic Regression for Early Screening of Ovarian Cancer

Yu-Seop Kim[1,3], Min-Ki Jang[2,3], Chan-Young Park[1,3], Hye-Jeong Song[1,3] and 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*
**Corresponding Author: yskim01@hallym.ac.kr*

Abstract

The best marker combination for differentiating the ovarian cancer from benign is explored with the logistic regression. The serum samples from 81 patients with ovarian cancer and 216 patients with benign pelvic masses provided by 2 institutes were analyzed using Luminex assay test. The selection performance of the logistic regression was compared with three other methods such as t-test, genetic algorithm, and random forest. The evaluation of the four methods were performed also with three classification methods including logistic regression, linear discriminant analysis, and k-nearest neighbor method. The 4 marker combination from the logistic regression showed the best performance against the other selection methods in terms of the 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 in 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 the 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 change.

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

* Corresponding Author

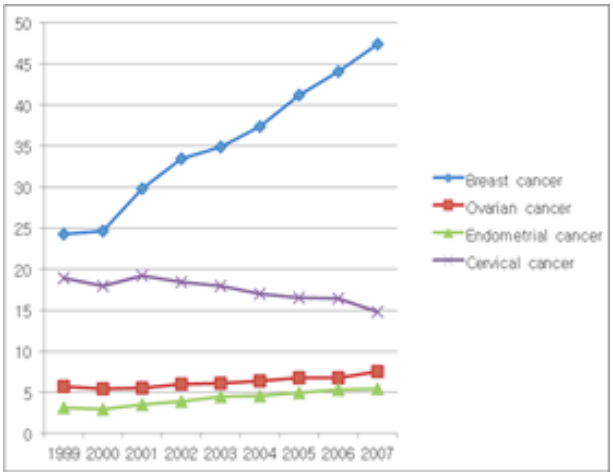

Figure 1. Changes in gynecologic cancer causes in Korea (1999-2007)

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. Researches tend to focus especially on improving the sensitivity and specificity in order to increase the accuracy of the diagnosis, and the commercialization of multi-biomarkers seems to be close at hand. However, a new technology to find the right 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, 8]. 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 [5, 6]. LR, k-nearest neighbor (k-NN) [5], and linear discriminant analysis (LDA) [9] 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 provided by Hallym University Medical Center and ASAN Medical Center were used. The samples were 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.

Logistic Regression is used to systematically combine the identifiers that have different output scales. The output of each biomarker according to the input pattern is ranked for each classification, and the sorted rank is used as the input for the final evaluation [5, 8].

T-Test is a statistical method that evaluates whether the average difference of two groups, mostly small groups consisting 30 samples or less, have statistical value and meaning. Independent samples t-test compares the means of two independent groups. When there are more than three groups, paired samples t-test is used to compare the average of two variables in the same group [5].

Genetic Algorithm is an optimization algorithm based on the principles of natural selection introduced in 1975 by John Holland. It is a search heuristic that is inspired from the mechanisms of natural heredity and evolution. GA is commonly used as a tool for search, optimization, and machine learning [6].

Random Forest produces various decision trees from the randomly extracted sample sets and evaluates the final class from the various classes of the produced decision trees by weighted voting [7].

K-Nearest Neighbor assumes all instances correspond to points in the n-dimensional space $\Re^n$. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance [5].

Linear Discriminant Analysis is not model-based but makes use of data to obtain a specific linear function because "when two or more populations have been measured in several characters, $x_1,\ldots, x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated" [9].

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 algorithms for the marker combination and combinations of evaluation algorithms used in the experiment are shown in Figure 2. The marker selection algorithms investigated all of 4-combinations of 8 biomarkers and selected the most accurate combination.

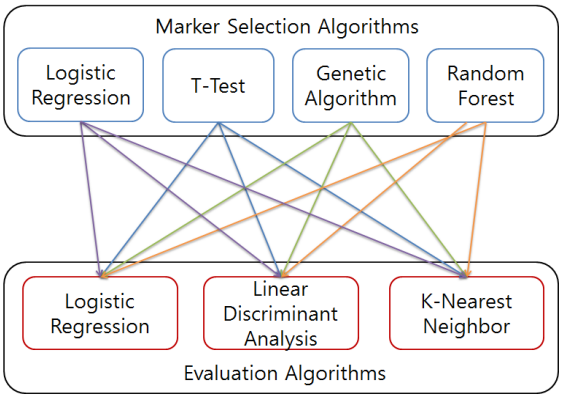

Figure 2. The algorithms for the marker combination and combinations of evaluation algorithms

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 accuracy 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 to combine more than 4 markers will make the realization and commercialization of multi-biomarkers difficult. 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

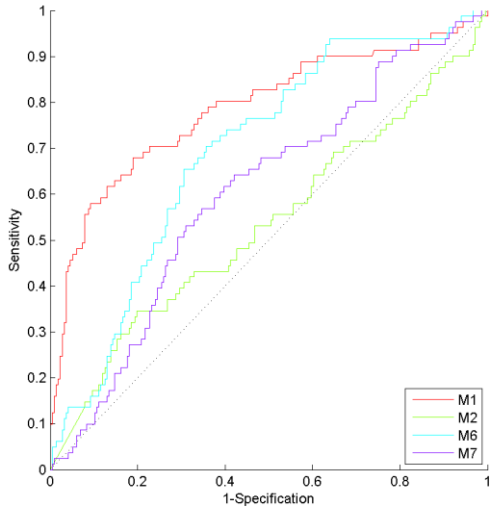

Figure 3. ROC Curves of individual biomarker M1, M2, M6 and M7

Table 1 shows the optimal marker combination obtained through LR and their performance when applying LR, LDA, and k-NN to the combined 4 markers. Figure 3 shows the ROC curves of the individual biomarkers obtained through LR. 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

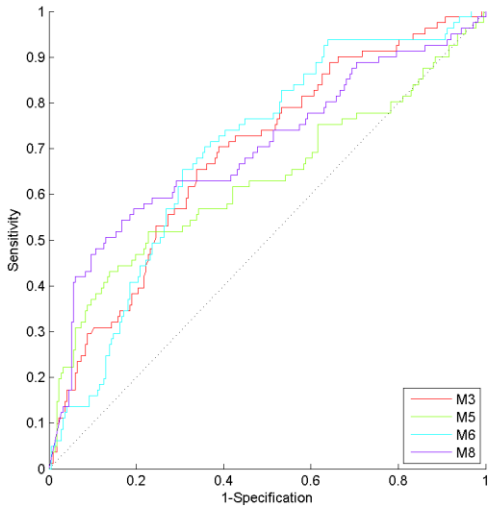

Figure 4. ROC Curves of individual biomarker M3, M5, M6 and, M8

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

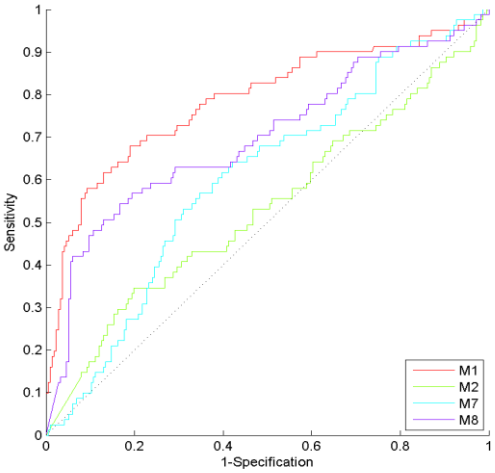

Figure 5. ROC Curves of individual biomarker M1, M2, M7 and, M8

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

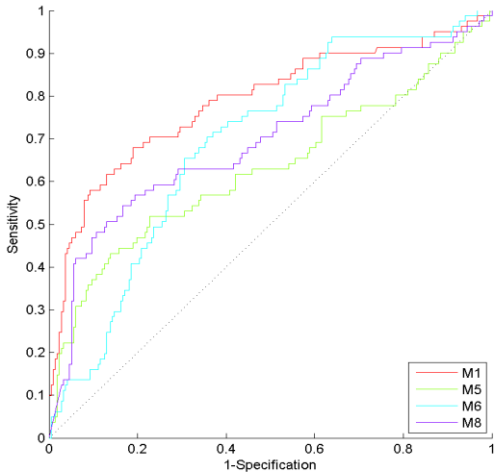

Figure 6. ROC Curves of individual biomarker M1, M5, M6 and M8

Table 2, 3, and 4 shows the optimal marker combination and their performance obtained through t-test, GA, and RF, respectively. The selected combination from each method was also evaluated with LR, LDA, and k-NN. Figures 4, 5, and 6 shows the ROC curves of the individual biomarkers obtained through t-test, GA and, RF, respectively. The tables show that the average accuracy over the three evaluation algorithms was greatest, when LR was employed as the marker selection algorithm.

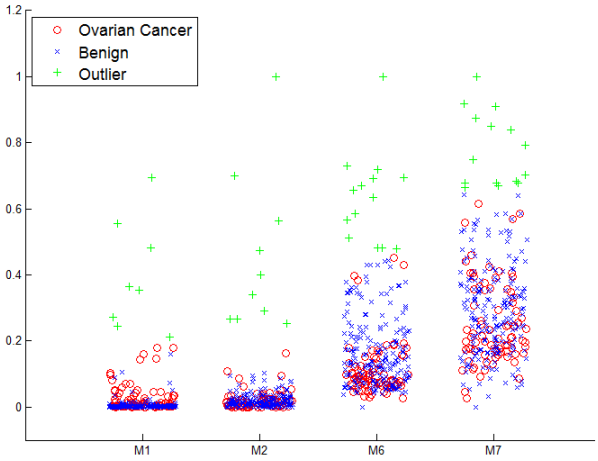

Figure 7. Dot plot of individual markers of the optimum combination

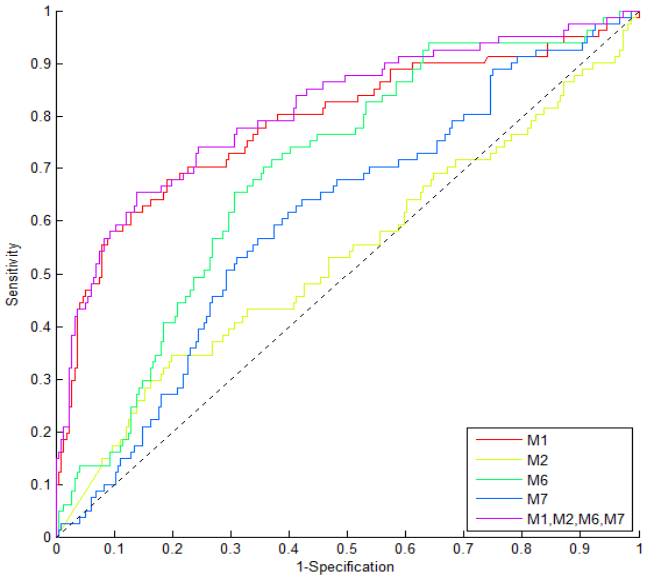

Figure 8. ROC curves of individual markers and the combination of the optimum combination

Figure 7 shows the statistical scatter plots for the individual biomarkers of the most accurate combination. The ROC curves of the individual markers and the logistic regression of the most accurate combination are illustrated in Figure 8.

5. Conclusion

This paper presents the exploration for the biomarker combination using logistic regression that can easily distinguish 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] N. Brian, M. Adele, V. Liudmila., P. Denise, W. Matthew, G. Elesier and L. Anna, "A serum based analysis of ovarian epithelial tumorigenesis", ELSEVIER Gynecologic Oncology, vol. 112, **(2009)**, pp. 47-54.

[3] C. ChiHeum, "Biomarkers related to Diagnosis and Prognosis of Ovarian Cancer", Korean Journal of Obstetrics and Gynecology, vol. 39, **(2008)**, pp. 90-95.
[4] J. SunKyung, O. EunJi, Y. ChulWoo, A. WoongSik, K. YongGu, P. YeonJun and H. KyungJa, "ELISA for the selection of HLA isoantibody and Comparison Evaluation of Luminex Panel Reactive Antibody Test", Journal of Korean Society for Laboratory Medicine, vol. 29, **(2009)**, pp. 473-480.
[5] F. David, P. Roger and P. Robert, "Statistics", 3rd Edition, W. W. Norton & Company, **(1998)**.
[6] M. M. Tom, Machine Learning. The McGraw-Hill", **(1997)**.
[7] B. Leo, Random Forest. Machine Learning, vol. 45, **(2001)**, pp. 5-32.
[8] P. McCullagh and J. A. Nelder, " Generalized Linear Models", New York: Chapman & Hall, **(1990)**.
[9] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems", Annals of Eugenics, vol. 7, no. 2, **(1936)**, pp. 179-188.

Authors

Yu-Seop Kim received his Ph.D. degree in Computer Engineering from Seoul National University. He is currently a Professor in the Department of Ubiquitous Computing at Hallym University, South Korea. His research interests are in the areas of bioinformatics, computational intelligence and natural language processing.

Min-Ki Jang received his B.S. degree in Computer Engineering from Hallym University. He currently studies for a master's degree in Computer Engineering at Hallym University. His recent interests focus on biomedical system and bioinformatics.

Chan-Young Park received his B.S. and the M.S. from Seoul National University and the Ph.D. degree from Korea Advanced Institute of Science and Technology in 1995. From 1991 to 1999, he worked at Samsung Electronics. He is currently a Professor in the Department of Ubiquitous Computing of Hallym University, Korea. His research interests are in Bio-IT convergence, Intelligent Transportation System and sensor networks.

Hye-Jeong Song received her Ph.D. degree in Computer Engineering from Hallym University. She is a Professor in Department of Ubiquitous Computing, Hallym University. Her recent interests focus on biomedical system and bioinformatics

Jong-Dae Kim received his M.S. and the Ph.D. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1984 and 1990, respectively. He worked for Samsung Electronics from 1988 to 2000 as an electrical engineer. He is a Professor in Department of Ubiquitous Computing, Hallym University. His recent interests focus on biomedical system and bioinformatics.