Looking for the Optimal Machine Learning Algorithm for the Ovarian Cancer Screening

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

[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
[3]Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon, South Korea
[2]chokood@nate.com, { hjsong, kimjd, cypark, yskim01}@hallym.ac.kr

Abstract

Ovarian cancer is very malignant tumor because it doesn't have any striking symptoms in its early stages. That's why the early screening is really necessary in its clinics. We try to look for the optimal methodology to find out biomarker combination making its classification performance better than other cases. We evaluate 9 machine learning algorithms, those are Random Forest, Logistic, Multilayer Perceptron, Bagging, Classification Via Regression, LogitBoost, MultiClassifer, Simple Logistic, and Logistic Regression. The Area Under the Curve (AUC) of each algorithm is compared. We firstly select 15 biomarkers which are widely spread in the ovarian cancer diagnosis and find the best three combinations which composed of two, three and four biomarkers by using Logistic Regression which is well known for its reliable performance. Than we re-evaluate the best combinations with nine algorithms including Logistic Regression to find the optimal machine learning algorithm. In this research, we can find possibility to use another machine learning algorithm rather than Logistic Regression.

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]. The rate of the cancer patients is increasing in a rapid pace by year, as shown in Figure 1.

* Corresponding author

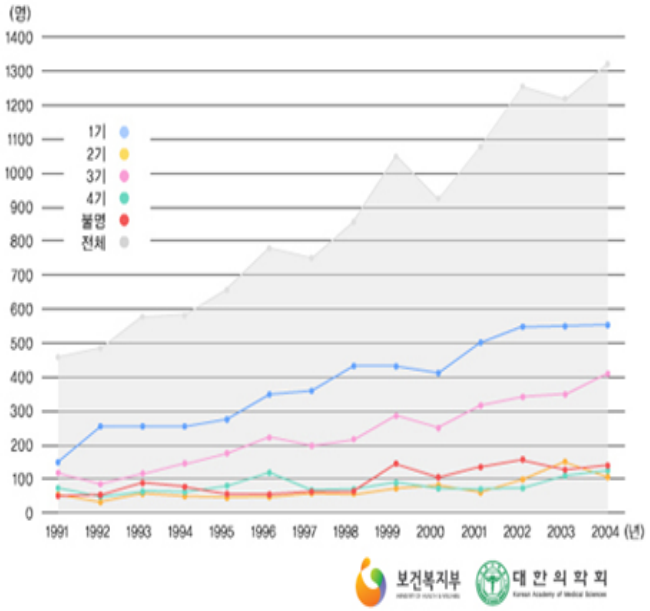

Figure 1. Occurrence of ovarian cancer diagnosis by year

Epithelial ovarian cancer, ranked as 90% of the ovarian cancer, is usually detected after it has developed past the tertiary period. As a result, the survival rate for 5 years after the diagnosis is less than 40%. It is evident that the development of a biomarker for early detection of the ovarian cancer has become paramount [2, 3].

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

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. Researches tend to focus especially on improving the sensitivity and specificity in order to increase the accuracy of the diagnosis. 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 sensitivity and specificity has not yet reached a satisfactory level [4].

In this paper, the concentration value of the biomarkers was obtained using Luminex [5]. Luminex follows the panel reactive antibody (PRA) method, a solid phase-based method of Luminex Corporation [5]. Luminex-PRA reacts the human leucocyte antigen (HLA) marker attached on Luminex-bead to the HLA antibody in urine, and detects the florescence of the antibody from the bead by utilizing its exclusive equipment and software [6].

This paper aims to determine the optimum marker combination from 15 biomarkers using Random Forest [7], Logistic [8], Multilayer Perceptron [9], Bagging [10], Classification Via Regression [11], LogitBoost [12], MultiClassifier [13], Simple Logistic, and Logistic Regression [14]. 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.

Methods of collecting the data are illustrated in chapter 2, and the experimental details are demonstrated in chapter 3. The results of the marker combinations and its classification performance are discussed in chapter 4, and chapter 5 presents the conclusion and possible future researches

2. Data Set

For this experiment, 176 (benign tumor 121, cancer 55) urine samples of Koreans were provided by two hospitals. Figure 2 shows the information of the clinical samples collected for this research.

Characteristics	No. of patients
No. of patients studied	176
Ovarian Cyst	121
Ovarian Cancer	55
Age (Mean ± S.D.)	44.5±13.42
(Range)	21-80
FIGO stage	
I	24 (43.6%)
II	1 (1.8%)
III	21 (38.1%)
IV	9 (16.36%)
I	24 (43.6%)

Figure 2. Data of the Clinical Samples

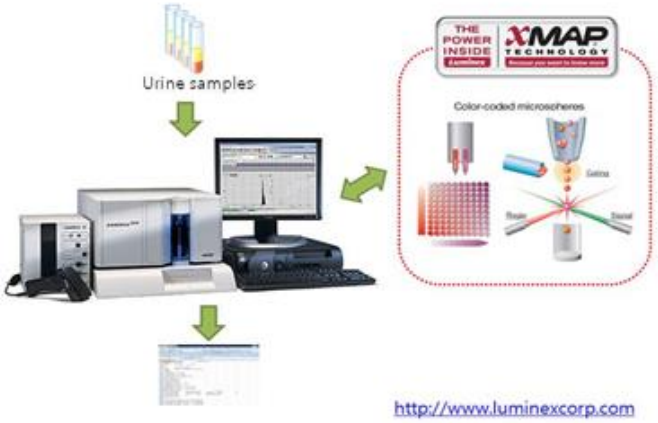

Figure 3. Lumixex facility

The urine samples were reacted with a Luminex-bead shown of biomarkers attached, and the fluorescence signal from the antibody of each bead was measured by using Luminex. Each fluorescence value of each biomarker was normalized to 0~1 according to the maximum and minimum values in order to standardize the range of the fluorescence value from each biomarker. Figure 3 shows the Luminex facility.

The 15 biomarkers used in this paper are commonly discussed biomarkers in the ovarian cancer researches [15, 16]. This research aims to find another algorithm instead of logistic regression in determining the optimum marker combination from the detected biomarkers in the urine.

3. Experiment

Three biomarker combinations were selected using logistic regression from fifteen biomarkers. These were selected first since the stability of Logistic Regression was proven in several previous researches [17, 18]. 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.

4. Results

In the experiment, the AUCs [19] of the multi-biomarker combinations consisting of 2~4 biomarkers selected by logistic regression were obtained using the nine algorithms mentioned in Chapter 3. 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 lists the AUC values when the top three combinations of two biomarkers from logistic regression are applied to nine algorithms including Logistic Regression. It can be seen that each algorithm shows different AUC values for the same combination. However, excluding the Logistic Regression, the rank among the three combinations did not change. From the fact that the M5 marker is in all three combinations, it can be said that M5 from urine samples plays a critical role in the early diagnosis of ovarian cancer.

Interestingly, the three marker combinations chosen by logistic regression show the highest performance when applied to Bagging and Classification Via Regression algorithm. Although the difference of the AUC values can be said to be statistically meaningless, this indicates that finding the marker combination with a different algorithm other than Logistic Regression might show better performance.

Table 1. Results for the combination of two biomarkers

Marker Combination / Algorithms	M5, M15	M3, M5	M5, M12
Bagging	0.862	0.813	0.817
Classification Via Regression	0.862	0.831	0.808
Logistic	0.844	0.831	0.81
LogisticBoost	0.859	0.848	0.787
Logistic Regression	0.86	0.836	0.85
MultiClass Classifier	0.844	0.831	0.81
Multilayer Perception	0.858	0.82	0.804
Random Forest	0.847	0.833	0.787
SimpleLogistic	0.837	0.82	0.817

Table 2 shows the optimum marker combinations and their performance when all the possible marker combinations consisting of three markers are used to classify benign and cancer by cross-validation. As in table 1, another algorithm, Random Forest, shows the best performance. Also, M5 was included in all three combinations, confirming the importance of the marker M5.

Table 2. Results for the combination of three biomarkers

Marker Combination / Algorithms	M3, M5, M15	M3, M5, M12	M2, M5, M12
Bagging	0.849	0.818	0.829
Classification Via Regression	0.867	0.83	0.855
Logistic	0.853	0.844	0.838
LogisticBoost	0.872	0.824	0.821
Logistic Regression	0.875	0.867	0.861
MultiClass Classifier	0.853	0.844	0.838
Multilayer Perceptron	0.861	0.821	0.821
Random Forest	0.891	0.815	0.801
SimpleLogistic	0.863	0.841	0.836

Table 3. Results for the combination of four biomarkers

Marker Combination Algorithms	M4, M5, M12, M15	M3, M5, M12, M15	M3, M5, M14, M15
Bagging	0.846	0.843	0.839
Classification Via Regression	0.865	0.862	0.865
Logistic	0.847	0.852	0.847
LogisticBoost	0.847	0.867	0.867
Logistic Regression	0.885	0.884	0.874
MultiClass Classifier	0.867	0.852	0.847
Multilayer Perceptron	0.867	0.863	0.87
Random Forest	0.856	0.873	0.862
SimpleLogistic	0.846	0.862	0.851

Table 3 shows the optimum marker combinations and their performance when all the possible marker combination consisting four markers are used. In this case, Logistic Regression shows the best performance. Also compared to the combinations consisting two and three biomarkers, when four biomarkers are combined, it had the highest AUC values. In all four combinations, markers M5 and M15 were included, which is the two markers that showed the best performance amongst the combination of two biomarkers when combined.

5. 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. For the experiment, urine samples from benign tumor patients and cancer patients were provided from two hospitals, and 15 types of biomarkers were extracted. Three combinations for each 2~4 biomarkers combined showing the highest performance were selected by logistic regression. For these nine combinations, nine classification algorithms were applied and the AUC values were obtained.

Although the AUC values for each marker combination were different according to the applied algorithm, the rank of combinations was similar in all cases. 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 [20-22], can be tested on the proposed experiment to evaluate their performance.

Acknowledgements

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] 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," J. Korean Institute of Information Technology, vol. 5, no. 1, **(2007)**.
[4] C. Cho, "Biomarkers for the Diagnosis of Ovarian Cancer and Treatment," Korean Society of Obstetrics and Gynecology, **(2011)**.
[5] S. Jung, E. Oh, C. Yang, W. Ahn, Y. Kim, Y. Park and K. Han, "Comparative Evaluation of ELISA and Luminex Panel Reactive Antibody Assays for HLA Alloantibody Screening", J Lab Med, vol. 29, no. 5 **(2009)**.
[6] N. El-Awar, J. Lee and P. Terasaki, "HLA Antibody Identification with Single Antigen Beads Compared to Conventional Methods", Hum Imminol, vol. 66, no. 9, **(2005)**.
[7] A. Liaw and M. Wiener, "Classification and Regression by Random Forest", R News, vol. 2, no. 3, **(2002)**.
[8] F. E. Harrell Jr., "Regression Modeling Strategies", Springer, **(2001)**.
[9] 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 Transaction on Neural Networks, vol. 1, no. 4, **(1990)**.
[10] L. Breiman, "Bagging Predictors", Machine Learning, vol. 24, **(1996)**.
[11] E. Frank, M. Hall, L. Trigg, G. Holmes and I. H. Witten, "Data Mining in Bioinformatics using Weka", Bioinformatics, vol. 20, no. 15, **(2004)**.
[12] Y. D. Cai, K. Y. Feng, W. C. Lu and K. C. Chou, "Using LogitBoost Classifier to Predict Protein Structural Classes", J Theoretical Biology, vol. 238, no. 1, **(2006)**.
[13] 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", Proceedings of the Second International Conference on Document Analysis and Recognition, **(1993)**.
[14] D. Freedman, R. Purves and R. Pisani, "Statistics, 3rd Edition", W. W. Norton & Company, **(1998)**.
[15] 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, vol. 112, no. 1, **(2009)**.
[16] S. D. Amnkar, 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, vol. 4, no. 2, **(2009)**.
[17] C. Lombardi, G. F. Tassi, G. Pizzocolo and F. Donato, "Clinical Significance of a Multiple Biomarker Assay in Patients with Lung Cancer. A Study with Logistic Regression Analysis", Chest Journal, vol. 97, no. 3, **(1990)**.
[18] J. Li, Z. Zhang, J. Rosenzweig, Y. Y. Wang and D. W. Chan, "Proteomics and Bioinformatics Approaches for Identification of Serum Biomarkers to Detect Breast Cancer", Clinical Chemistry, vol. 48, no. 8, **(2002)**.
[19] A. P. Baradley, "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms," Pattern Recognition, vol. 33, no. 7, **(1997)**.
[20] H. T. Huynh, J. Kim and Y. Won, "Classification Study on DNA Microarray with Feedforward Neural Network Trained by Singular Value Decomposition", International Journal of Bio-Science and Bio-Technology, vol. 1, no. 1, **(2009)**.
[21] G. Mirceva and D. Davcev, "HMM based Approach for Classifying Protein Structures", International Journal of Bio-Science and Bio-Technology, vol. 1, no. 1, **(2009)**.
[22] R. Pal, P. Garg, R. Chechi, S. Kumar and N. Kumar, "Cancer Growth Prediction via Artificial Neural Networks", International Journal of Bio-Science and Bio-Technology, vol. 2, no. 2, **(2010)**.

Authors

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

Seung-Kyun Ko is now the B.S. student in Ubiquitous Game Engineering of Hallym University. His recent interests focus on mobile game and bioinformatic

Jong-Dae Kim received his M.S. and 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.

Chan-Young Park received his B.S. and 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.

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.