Advanced Science and Technology Letters
Vol.120 (AIA 2015), pp.803-806
http://dx.doi.org/10.14257/astl.2015.120.158

Best Combination Looking for Ovarian Cancer

Eun-Suk Yang[1], Jong-Dae Kim[1],
Chan-Young Park[1], Hye-Jeong Song[1] and Yu-Seop Kim[1],

[1] Department of convergence software, Hallym University
{miriam219, kimjd, cypark, hjsong, yskim01}@hallym.ac.kr

Abstract. In this paper, we find the alternative combination which can be replace CA-125. We compare several algorithms for diagnosis of ovarian cancer. The serum samples contain 101 cancer and 92 healthy samples. We perform two major tasks: Marker selection and Classification. For optimal marker selection, we use Genetic Algorithm, Random Forest, T-test and Logistic Regression. For classification, we compare Linear Discriminative Analysis, K-Nearest Neighbor and Logistic Regression. As a result, we find HE4-ELISA, PDGF-AA, Prolactin, TTR can be alternative combination for detecting ovarian cancer.

Keywords: Marker selection, Classification, Ovarian Cancer, Logistic Regression

1 Introduction

Ovarian cancer is the eighth most common cancer and has the fifth fatality-to-case ratio in United States. According to a statistics of Centers for Disease Control and Prevention (CDC) in 2012, about 20 thousands women in United States were diagnosed with ovarian cancer, and about 75% died from it. In addition, when ovarian cancer is found in its early stage, the probability of 5-year survival yields up to 92%. However the early detection rate is only 19%. It is clarify that the early detection of ovarian cancer improves the clinical output [1,2].

For early diagnosis, many researches have been performed as follows: finding multiple biomarkers [3], early detection using menopausal information [4], and finding optimal combination using machine learning algorithms [5,6]. Specifically, many of them have been developed for distinguishing between benign and cancer [3-8]. Unlikely them, we focus on distinguishing cancer sample from normal samples.

One of the most popular screening test for ovarian cancer is CA-125 blood test. The level of CA-125 is high from many patients with ovarian cancer. However, using CA-125 for screening test is that common conditions other than cancer can also cause the level of CA-125 [9,10].

Finding related biomarkers for specific disease is required a huge amount of clinical experiment. If there are a lot of candidate set of biomarker, it is time-consuming and expensive, corresponding the number of biomarkers and the number of combination. The underlying motivation of our paper is that machine learning can

ISSN: 2287-1233 ASTL

decrease the cost and time for experiments, by suggesting the best optimal combination set.

In this paper, we perform two parts for finding alternative marker combination for detecting cancer. First, we identify the best marker combinations of 16 biomarkers. Second, we compare the classification method of optimal combination, distinguishing normal and cancer samples. Several machine learning algorithms are used in both parts.

2 Methods

Serum samples were obtained from 101 patients with ovarian cancer and from 92 healthy women provided through two hospital, Hallym University Chuncheon Sacred Heart Hospital and Asan Medical Center. We do not care about cancer stage, and the state of menopause which is important factor associated with the risk of malignancy [4,9]. The 16 serum biomarkers, which is commonly discussed for ovarian cancer researches, are used to our experiment [2,10,11].

To select optimal marker combination which can diagnose cancer and normal data, we use four algorithms: Random Forest (RF), Genetic Algorithm (GA), T-test and Logistic Regression (LR). The size of combination is set from 2 to 4 for reducing a time consuming. The top marker combinations for each algorithm are computed to 5-fold cross-validation. We repeated it 1000 times in order to decrease the deviation of result. The final best marker combinations are selected to average Receiver Operating Characteristic (ROC) Area Under the Curve (AUC).

With the selected optimal combinations, we perform the three classification method: Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN) and Logistic Regression (LR). We compare the accuracy for classification between normal and cancer data.

3 Results

Table 1 shows AUC value for marker selection and accuracy for classification. The GA yields 0.86 AUC value as the lowest value among four algorithms, and the RF performs 0.98 AUC. T-Test and LR shows the highest score, 0.99 AUC value. Except RF, the rest of algorithm shows the best AUC value to 4 combinations. However, there are no major differentiation of AUC value between 3 and 4 marker combinations. Intuitively, we find that it is not necessary to use 4 marker combination. We compare three different classification method. In marker sets selected by GA, 2 combination shows the best performance of 0.88 by using KNN. In marker sets chosen by RF, 3 and 4 combination yield almost same accuracy by using LR. Marker sets chosen by T-Test shows totally same performance of 0.95 by using LR. In optimal combinations by LR, a classifier trained by LR yields 0.95 accuracy.

Not surprisingly, the GA algorithm which has a lowest AUC value for marker selection, performs the lowest accuracy of 0.81 for 3 combination. All marker selection algorithm except GA, shows better performance for the 3 and 4

combinations rather than 2 combinations. The classifier using Logistic Regression shows the outstanding performance in over 70% of marker sets. In the result, we find HE4-ELISA, PDGF-AA, Prolactin, TTR is the best combination for detecting ovarian cancer.

Table 1. AUC value and accuracy for optimal biomarker combinations. The first column in table 1 indicates algorithm and each describes as follows: GA is Genetic Algorithm, RF is Random Forest and LR is Logistic Regression. The second column is the best combination selected by best AUC value and the third column is its value. The bold shows the highest results in each combination.

		Selection	Classification		
Algo.	Combi.	AUC	LDA	KNN	LR
GA	ApoCIII, TTR	0.86	0.88	**0.88**	0.87
	IL-6, CEA, OPN	0.90	0.67	0.79	**0.81**
	MIF, APoAI, OPN, IL-6	**0.90**	0.74	**0.86**	0.83
RF	CA125, HE4-ELISA	0.92	0.69	**0.83**	0.78
	Prolactin, TTR, HE4-ELISA	**0.98**	0.91	0.93	**0.95**
	TTR, Prolactin, CA125, HE4-ELISA	0.98	0.92	0.91	**0.94**
T-Test	TTR, ApoCIII	0.95	0.88	**0.88**	0.87
	TTR, ApoCIII, Prolactin	0.98	0.93	0.93	**0.95**
	TTR, ApoCIII, Prolactin, OPN	**0.99**	0.93	0.93	**0.95**
LR	Prolactin, TTR	0.98	0.91	0.92	**0.93**
	ApoCIII, HE4-ELISA, Prolactin	0.99	0.93	**0.94**	0.93
	HE4-ELISA, PDGF-AA, Prolactin, TTR	**0.99**	0.92	0.94	**0.95**

4 Conclusion

In this paper, we present the exploration for the marker selection and classification between cancer and normal samples. For marker selection, we find all methods except genetic algorithm, can capture in combining marker sets a marker, which has a high AUC value. For classification, logistic regression also presents the highest accuracy. The experimental results shows that logistic regression is an outstanding algorithm for both marker selection and classification problem.

References

1. Clarke, C.H., Yip, C., Badgwell, D., Fung, E.T., Coombes, K.R., Zhang, Z., Lu, K.H., Bast, R.C.: Proteomic Biomarkers Apolipoprotein A1, Truncated Transthyretin and Connective Tissue Activating Protein III Enhance the Sensitivity of CA125 for Detecting Early Stage Epithelial Ovarian Cancer. Gynecologic oncology. 122(3), 548-553 (2011)
2. Hensley, M.L., Martee, L., Castiel M., Robson, E.: Screenging for Ovarian Cancer: What We Know, What We Need to Know. 14(11), 1601—1607 (2000)
3. Kim, Y.S., Jang, M.K., Park, C.Y., Song, H.J., Kim, J.D.: Exploring multiple biomarker combination by logistic regression for early screening of ovarian cancer. International Journal of Bio-Science and Bio-Technology 5, 67—76 (2013)
4. Jang, M.K., Kim, Y.S., Park, C.Y., Song, H.J., Kim, J.D.: Integration of Menopausal Information into the Multiple Biomarker Diagnosis for Early Diagnosis of Ovarian Cancer. International Journal of Bio-Science and Bio-Technology. 5(4), 215—222 (2013)
5. Song, H.J., Ko, S.K., Kim, J.D., Park, C.Y., Kim, Y.S.: Looking for the Optimal Machine Learning Algorithm for the Ovarian Cancer Screening. International Journal of Bio-Science and Bio-Technology. 5(2), 41—44 (2013)
6. Kim, Y.S., Jang, M.K., Park, C.Y., Song, H.J., Kim, J.D.: Best Biomarker Combination for Ovarian Cancer (2013)
7. Nolen, B., Marks, J., Ta'san, S., Rand, A., Luong, T., Wang, Y., Blackwell, K., Lokshin, A.: Serum biomarker profiles and response to neoadjuvant chemotherapy for locally advanced breast cancer. Breast Cancer Research 10, R45 (2008)
8. Nolen, B., Marrangoni, A., Velikokhatnaya, L., Prosser, D., Winans, M., Gorelik, E., Lokshin, A.: A Serum Based Analysis of Ovarian Epithelial Tumorigenesis. ELSEVIER Gynecologic Oncology. 112, 47-54. (2009)
9. US Department of Health and Human Services. Draft Guidance for Industry, Clinical Laboratories, and Staff: In Vitro Diagnostic Multivariate Index Assays, http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/uc m071455.pdf
10. Kozak, K.R., Su, F., Whitelegge, J.P., Faull, K., Reddy, S., Farias, E.R.: Characterization of Serum Biomarkers for Detection of Early Stage Ovarian Cancer. Proteomics 5(17), 4589--4596 (2005)
11. He, G., Holcroft, C.A., Beauchamp, M.C., Yasmeen, A., Ferenczy, A., Kendall-Dupont, J., Gotlieb, W.H.: Combination of Serum Biomarkers to Differentiate Malignant from Benign Ovarian Tumours, J Obstet Gynaecol Can. 34(6), 567-574 (2012)

