Advanced Science and Technology Letters
Vol.120 (AIA 2015), pp.789-792
http://dx.doi.org/10.14257/astl.2015.120.155

Find Alternative Biomarker via Word Embedding

Kyeong-Min Nam[1,2], Hey-Jung Song[1,2], Jong-Dae Kim[1,2], Can-Young Park[1,2],
Yu-Seop Kim[1,21]

[1] Department of Convergence Software, Hallym University 1 Hallymdaehak-gil,
Chuncheon,Gangwon-do, 200-702 Korea

[2] Bio-IT Research Center, Hallym University, 1 Hallymdaehak-gil, Chuncheon, Gangwon-do, 200-702 Korea
{jkre4030, hjsong, kimjd, cypark, yskim01}@hallym.ac.kr

Abstract. We use word embedding to find an alternative biomarker for the early diagnosis of ovarian cancer from the biomedical corpus. First, we derived a low dimensional representation of each biomarker embedding induced from CCA. Second, we found a similar pair of biomarkers in the literature by using cosine similarity of biomarker embedding. In order to determine the clinical similarity between the pair of biomarkers, we used the AUC of the combination of 2 biomarkers used previously. In the experiment, we confirmed that correlation between the high similarity biomarker pair, was highly correlated as the average 0.698 of the actual AUC correlation of the top 10 of the pair.

Keywords: Biomarker, Serum, Ovarian Cancer, Canonical Correlation Analysis, Embedding, AUC, Cosine similarity.

1 Introduction

Ovarian cancer is a malignant tumor that is common in women between the ages of 50–70. It was reported that about 1,000–1,200 new ovarian cancer patients were diagnosed in 2002. Additionally, ovarian cancer is the second most common gynecological cancer after cervical cancer [5]. The 5-year survival rate associated with ovarian cancer is 50~95% when diagnosed in the early stages (I, II), and it is below 25% when diagnosed at the later stages (III, IV) [1]. Therefore, the identification of diagnostic biomarkers for early detection of ovarian cancer is paramount.

In previous studies, research papers on the relevant cancer were collected and then analyzed. After that, they collected biomarkers that were reported to have a correlation with the relevant cancer. After analyzing the correlation between the concentration of each biomarker and cancer patients, biomarker candidates with values above a certain level were selected. Moreover, measurement of the concentration of selected biomarkers is expensive.

[1] Corresponding author

ISSN: 2287-1233 ASTL

In this paper, we use word embedding to find an alternative biomarker for the early diagnosis of ovarian cancer from the biomedical corpus. Word embedding is a word vector representation with previously proven efficiency in the biomedical domain [2].

First, we derived a low dimensional representation of each biomarker embedding induced from Canonical Correlation Analysis (CCA) [4], which is a powerful and flexible statistical technique for dimensionality reduction. Second, we found a similar pair of biomarkers in the literature by using cosine similarity of biomarker embedding. In order to determine the clinical similarity between the pair of biomarkers, we used the area under the curve (AUC) of the combination of 2 biomarkers used previously [3]. In the experiment, we confirmed that correlation between the high similarity biomarker pair, was highly correlated as the average 0.698 of the actual AUC correlation of the top 10 of the pair.

2 Data and Methods

In this study, Canonical Correlation Analysis (CCA) was used to find 22 biomarker embeddings. Cosine similarity of all the combinations of the two biomarkers was calculated using the embedding values, and it was identified whether the top 10 of the biomarker pair was clinically alternative.

2.1 Data

We established a new bio-corpus by using titles and abstracts of 136,725 biomedical papers in PubMed. Biomarker embedding was made by using the established bio-corpus. Additionally, we used the AUC value of the combination of 2 biomarkers of serum samples obtained from 254 Korean women to compare the AUC of the biomarker pair; note that these biomarkers were used in a previous study [3].

2.2 Canonical Correlation Analysis (CCA)

CCA is a powerful technique for inducing general representations that operates on a pair of multi-dimensional variables. CCA seeks k dimensions in which these random variables X and Y are maximally correlated. When they are defined as one-hot encodings, the CCA computation reduces to performing an support vector machine (SVD) of the matrix Ω where each entry is:

$$\Omega_{w,c} = \frac{\text{count}(w,c)}{\sqrt{\text{count}(w)\text{count}(c)}} \tag{1}$$

Where **count(*w, c*)** denotes co-occurrence matrix count of word *w* and context *c* in the given corpus, **count(*w*)** and **count(*c*)** denotes co-occurrence matrix count of word or context.

2.3 AUC correlation among the combination of 2 biomarkers

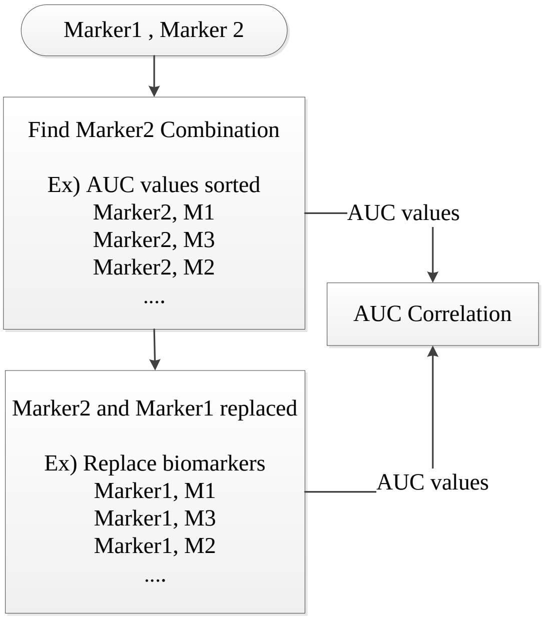

Fig 1. AUC correlation of the 2 replaced biomarkers

The calculation of AUC correlation using the 22 biomarker pairs of the top 10 is shown in Figure 1. As shown in Figure 1, by using the AUC result of the 22 biomarker pairs obtained from [3], first AUC value of mark2 and other biomarker combination was sorted in the descending order. Then, calculate the AUC value of combination of 2 biomarkers when Marker2 is replaced with Marker1, and the correlation between the two AUC values.

3 Results

Table 2 shows the AUC correlation calculated using the embedding similarity of the top 5% of the biomarker pairs and the method in section 3.3. Table 2 indicates that the performance is good regardless of embedding similarity, but the correlation is low when similarity is below 0.6.

Table 1. AUC correlation using embedding similarity of top 10 the biomarker pairs.

Marker	Embedding Cosin similarity	AUC Correlation
M15, M17	**0.962**	**0.955**
M9, M18	0.819	0.750
M14, M18	0.788	0.703
M10, M14	0.744	0.763
M9, M14	0.724	0.9691
M9, M10	0.690	0.758
M13, M17	**0.681**	**0.865**
M8, M10	0.677	0.673
M10, M18	**0.659**	**0.902**
M13, M15	**0.651**	**0.860**

4 Conclusion

In this study, we found an alternative biomarker for the early diagnosis of ovarian cancer from a biomedical corpus by using word embedding. In particular, we determined the AUC correlation to find the similarity between the alternative biomarker pair and others. Most biomarker pairs showed a high correlation regardless of the embedding similarity. The average of the entire AUC correlation is 0.698, and the correlation in the section below 0.6 was 0.611, and 0.766 in the section above 0.6. Thus, the biomarker pairs with biomarker embedding similarity above 0.6 are believed to be clinically alternative.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (2015R1A2A2A01007333) and by the Ministry of Education, Science and Technology (2010-0010612).

References

1. Kozak, K.R., Su, F, Whitelegge, J.P., Faull, K., Reddy, S., and Farias-Eisner, R.: Characterization of serum biomarkers for detection of early stage ovarian cancer. Proteomics. 5, 4589—4596 (2005)
2. Muneeb, T.H., Sahu, S.K., and Anand, A.: Evaluating distributed word representations for capturing semantics of biomedical concepts. ACL-IJCNLP. (2015)
3. Jang, M.K., Kim, Y.S., Park, C.Y., Song, H.J., and 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, 215—222 (2013)
4. Dhillon, P.S., Foster, D.P., and Ungar, L.: Multi-view learning of word embeddings via cca. Advances in Neural Information Processing Systems 199—207 (2011)
5. Seoul National University Hospital, http://www.snuh.org.