워드 임베딩을 이용한 COPD와 암 관련 바이오마커의 상관관계 분석

윤병훈○, 김유섭

한림대학교, 융합소프트웨어 학과
yqudgns1222@gmail.com, yskim01@hallym.ac.kr

Correlation Analysis of Cancer Biomarkers and COPD Using the Word Embedding

Byeong-Hun Yoon○, Yu-Seop Kim
Department of Convergence Software, Hallym University

요 약

본 연구에서는 COPD와 기존에 연관이 있는 것으로 알려진 바이오마커 이외의 새로운 바이오마커를 찾고자 한다. Pubmed Data에서 선정한 암 관련 바이오마커를 추출하여 COPD와 암 관련 바이오마커의 관계를 파악하는 데이터로 사용한다. 그리고 워드 임베딩 모델 중 Word2vec을 사용하여 워드 임베딩 한다. 워드 임베딩한 K차원의 COPD와 암 관련 바이오마커를 t-SNE를 사용하여 시각화한다. 또한 코사인 유사도를 이용하여 COPD와 암 관련 바이오마커의 유사도를 측정한다. 그리고 코사인 유사도와 t-SNE 결과를 이용하여 COPD와 암 관련 바이오마커와의 상관관계를 파악할 수 있으며, 암 관련 바이오마커와 COPD 관련 바이오마커를 비교 하여 기존의 COPD와 연관이 있다고 알려진 바이오마커 이외의 새로운 바이오마커를 찾을 수 있다.

주제어: Word-embedding, COPD, Biomarker, Word2vec

1. 서론

COPD(Chronic Obstructive Pulmonary Disease, 만성 폐쇄성 폐질환)는 천식과 비슷하게 호흡곤란, 기침, 가래 등의 기도 질환 증상을 나타내다가 폐 기능을 악화시켜 사망에 이르게 하는 질병이다[1]. 이와 관련하여 건강 상태를 확인하는데 지표가 되는 바이오마커(Biomarker)[2]를 이용하여 여러 질병을 예측하고, 몸 안의 변화를 알아내는 연구가 진행되고 있다.[3]

바이오마커란 질병을 발견, 모니터링하거나 치료하는 데에 사용되는 신체의 변화를 감지할 수 있는 척도가 되는 생체지표를 가리킨다.

이와 같이, 질병과 바이오마커에 대한 연구가 증가하면서 특정 질병과 관련 있는 바이오마커로 다른 질병의 바이오마커를 찾기 위한 다양한 연구들이 진행되고 있다[4][5].

본 논문에서는 COPD와 관련이 있는 바이오마커 이외에 새로운 바이오마커를 찾기 위하여 발병률이 높은 암(Cancer) 관련 바이오마커를 선정하여 상관관계를 파악한다. Pubmed Data[1]의 전체 문서를 Word2vec[6]을 이용하여 워드 임베딩(Word-Embedding)하고, COPD와 암 관련 바이오마커 38개[2]를 추출한다. 또한, 워드 임베딩한 K차원의 COPD와 암 관련 바이오마커 데이터를 t-SNE(t-distributed Stochastic Neighbor Embedding)[7]를 이용하여 2차원으로 매핑하고 결과를 시각화한다. 그리고, COPD와 암 관련 바이오마커를 코사인 유사도(Cosine Similarity)를 이용하여 COPD와 암 관련 바이오마커 간의 유사도를 측정한다. 마지막으로, 코사인 유사도와 t-SNE 결과를 이용하여 COPD와 암 관련 바이오마커와의 상관관계를 파악한다. 그리고, 암 관련 바이오마커와 COPD 관련 바이오마커를 비교하여, COPD와 밀접한 연관성을 띄는 대체 바이오마커를 추정하고자 하는 것이다.

2. 방법론

본 논문에서는 전체 807,821개의 Pubmed Data의 문서를 Word2vec을 이용하여 워드 임베딩하고, COPD와 바이오마커 38개를 추출하여 실험 데이터로 사용한다. 코사인 유사도 및 t-SNE 결과를 통하여 COPD와 바이오마커의 상관관계를 파악하고, 암 관련 바이오마커와 COPD 관련 바이오마커를 비교한다.

2.1 데이터

본 논문에서는 실험에 사용할 데이터인 암 관련 바이오마커를 선정하기 위하여 국가정보암센터[3] 사이트의 통계자료를 바탕으로 발병률이 높은 암을 선정하여 아래 [표 1]과 같이 암 관련 바이오마커 38개 선정하였다.

1) ftp://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_package/

2) https://en.wikipedia.org/wiki/Tumor_marker
http://www.bea.hi-ho.ne.jp/~ahcc/ahcc18.htm

3) http://www.cancer.go.kr/

[표 1] 암 관련 바이오 마커 38개

바이오 마커	P53, EGFR, ACT, I1-6, Brca1, Brca2, Kras, PSA, Braf, I1-8, CRP, AFP, CD117, MIF, Cytokeratin, CA125, FSH, HER-2/NEU, CD20, CA19-9, TTR, MMP-7, ALK, SOD, SP1, Cortisol, PAP, OPN, HCG, Prolactin, TPA, PTHRP, PDGFR, IAP, HE4, TK, ugt1a1

예를 들어, P53[8]은 암 억제 단백질로 알려져 있으며, 유전자의 돌연변이가 나타나면 암 발병 확률이 증가하게 된다. EGFR[9]은 표피성장인자 수용체로 변이를 통한 과다 발현이나 과다반응은 폐암, 항문암, 등의 암 발생과 관련있다고 알려져 있다. 이외에도, PSA, PAP는 전립선암, HE4는 난소암, SCFR은 위장관 간질종양, CRP는 폐암, Cortisol은 유방암, IAP는 면역력에 중요한 조절인자로 알려져 있다.

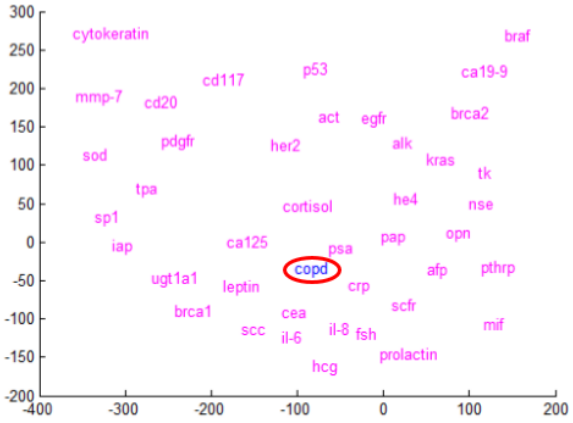
2.2 Word2vec

워드 임베딩은 문서 내에 있는 모든 단어들에 대해 벡터 값을 부여하여 벡터 표현을 학습하는 기술이다.

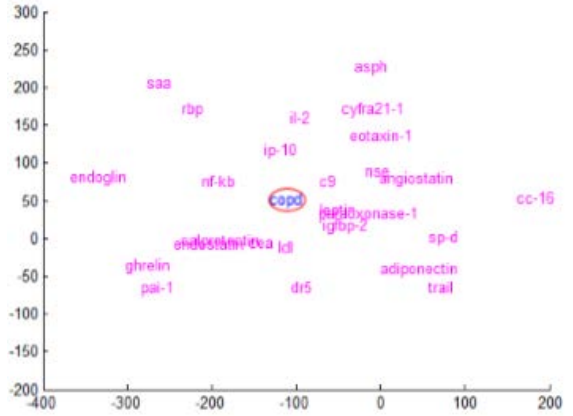
본 논문에서는 워드 임베딩 모델 중 Word2vec을 사용한다. Word2vec은 2013년 구글에서 발표된 연구 모델로 같은 맥락을 지니고 있는 단어는 서로 가까운 의미를 지니고 있다는 전제를 가지고 있다. 또한, 가장 흔하게 텍스트로 된 문장을 이해할 때에 사용된다. word2vec 모델의 학습방법에는 Skip-gram방식과 CBOW(Continous Bag Of Words)방식이 있다[10]. 하지만, 대규모 데이터 셋에서는 Skip-gram이 더 정확한 것으로 알려져있다. 따라서, 본 논문에서는 Skip-gram 방식을 사용하여 5차원, 10차원, 15차원, 20차원, 100차원으로 워드 임베딩한다.

2.3 t-SNE

본 논문에서 워드 임베딩한 고차원의 암 관련 바이오마커의 벡터를 가지고 코사인 유사도를 계산 하게 되면 정확한 유사도를 구하기 어렵다. 따라서, t-SNE를 사용하여 2차원으로 매핑하고 결과를 시각화 한다. 아래 [그림 1]은 암 관련 바이오마커 100차원의 벡터 값을 2차원으로 맵핑하여 가시화시킨 결과이다. [그림 1]에서 파란색은 COPD를 나타내며, 마젠타색은 암 관련 바이오마커를 나타낸다. [그림 2]는 COPD 관련 바이오마커이며, 파란색은 COPD, 마젠타색은 COPD 관련 바이오마커를 나타낸다. [그림 1 ~ 2]을 보면, COPD 기준으로 바이오마커가 어떻게 분포 되어 있는지를 한눈에 알 수 있다.



[그림 1] 암 관련 바이오마커 t-SNE 결과



[그림 2] COPD 관련 바이오마커 t-SNE 결과

2.4 코사인유사도

본 논문에서는 COPD와 바이오마커 간의 유사도 계산하기 위하여 코사인 유사도를 사용하여 유사도를 측정한다. 코사인 유사도는 내적 공간의 두 벡터간 각도를 코사인 값을 이용하여 측정된 벡터간의 유사도이다. 이는 Vector Space Model에서 가장 많이 사용되는 문서와 질의어 간의 유사도 계산법이다. 벡터 A와 벡터 B가 주어졌을 때, 내적과 벡터의 크기 등을 이용하여 표현된다. 계산된 유사도는 -1에서 1사이의 값을 가진다.

$$\text{Similarity} = \text{Cos}(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

3. 실험결과

본 논문에서는 코사인 유사도를 사용하여 COPD와 암 관련 바이오마커의 유사도를 계산한다. 아래 [표 2]는 각각 5차원, 10차원, 15차원, 20차원으로 유사도 값를 계산한 결과의 상위 5개씩을 정리한 것이다.

[표 2] 암 관련 바이오마커 유사도 상위 5개

차원	바이오마커	유사도	차원	바이오마커	유사도
5	PSA	0.879	15	PSA	0.795
	PAP	0.870		CRP	0.775
	HE4	0.810		Cortisol	0.753
	SCFR	0.792		SCFR	0.734
	CA125	0.762		PAP	0.714
10	PSA	0.782	20	CRP	0.805
	SCFR	0.762		PSA	0.775
	CRP	0.744		Cortisol	0.744
	Cortisol	0.737		SCFR	0.719
	IAP	0.726		MIF	0.692

그리고, 암 관련 바이오마커와 COPD 관련 바이오마커를 비교하기 위하여 COPD 관련 바이오마커의 각각 5차원 10차원, 15차원, 20차원 유사도 결과를 아래 [표 3]과 같이 정리 하였다.

[표 3] CODP 관련 바이오마커 유사도 상위 5개

차원	바이오마커	유사도	차원	바이오마커	유사도
5	CC-16	0.935	15	CC-16	0.799
	Calprotectin	0.838		Calprotetin	0.768
	Cyfra21-1	0.800		SAA	0.749
	Endoglin	0.770		leptin	0.717
	Leptin	0.763		PON-1	0.699
10	CC-16	0.892	20	CC-16	0.773
	Calprotectin	0.825		Calprotectin	0.769
	PON-1	0.795		SAA	0.752
	Leptin	0.768		Leptin	0.706
	SAA	0.759		PON-1	0.687

[표 3]을 기준으로 [표 2]와 비교하여 암 관련 바이오마커 중 COPD 관련 바이오마커 보다 높은 유사도 계산 결과를 보이는 바이오마커를 추출한다. [표 2] 5차원의 경우, PSA, PAP, HE4, SCFR가 [표 3] 5차원의 CC-16을 제외한 유사도 상위 4개 보다 높게 나왔다. [표 2] 10차원의 경우, PSA, SCFR이 [표 3] SAA보다 유사도 값이 높게 나왔다. [표 2] 15차원의 경우, PSA, CRP, Cortisol, SCFR이 [표 3] Leptin, PON-1보다 유사도 값이 높게 나왔다. 마지막으로, [표 2] 20차원의 경우, CRP, PSA, Cortisol, SCFR, MIF가 [표 3] PON-1보다 유사도 값이 높게 나온 것을 알 수 있다.

[표 2]와 [표 3]을 비교한 결과, 암 관련 바이오마커 중 PSA, PAP, HE4, SCFR, Cortisol, CRP가 COPD 관련 바이오마커와 유사도 값이 비슷하거나 높은 것을 알 수 있었다. 또한, 암 관련 바이오마커 PSA, PAP, HE4, SCFR, Cortisol, CRP 6개가 COPD와 상관관계가 높을 것으로 추정할 수 있다.

4. 결론

본 논문에서는 Pubmed Data의 문서를 Word2vec을 이용하여 워드 임베딩하고, 코사인 유사도 및 t-SNE 결과를 통하여 COPD와 바이오마커의 상관관계를 파악한다. 그리고, 암 관련 바이오마커와 COPD 관련 바이오마커를 비교하여 COPD와 암 관련 바이오마커 사이의 관계를 파악한다.

특정 질병과 유의한 관계를 갖는 바이오마커를 찾는 것은 의료계에서 매우 의미 있는 일이다. 바이오마커는 질병의 사전 진단이나 진행 정도를 모니터링하는데 있어서 매우 중요하다. 또한 특정 질병과 관련이 있을 것으로 추정되는 바이오마커에 대하여 섣불리 임상연구를 진행 할 수 없다는 점을 고려할 때, 문서의 사전 정보에서 관련성을 입증하는 것은 아주 큰 의미를 가진다. 실험결과 상관관계가 높지만 현재 연구되고 있지 않은 COPD 바이오마커 쌍에 대해서는 임상연구를 통하여 보다 정확한 관련성을 입증하는 것이 좋을 것이다. 향후 연구에서 유사도에 대한 분명한 검증을 하여 그 유용성을 입증할 것이고, 더 나아가서 현재까지 관련이 있다고 알려지지 않은 특정 질병과 특정 바이오마커의 조합을 찾을 수 있는 방법을 제시하겠다.

참고문헌

[1] Vestbo J, Hurd SS et al, " Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease", American Journal of Respiratory and Critical Care Medicine, 187(4) : 347-65, 2013.

[2] Aronson, Jeffrey, "Biomarkers and surrogate endpoints", British Journal of Clinical Pharmacology, 59 (5): 491-494, 2005.

[3] Craig E. Wheelock, Victoria M. Goss et al, "Application of'omics technologies to biomarker discovery in inflammatory lung diseases", European Respiratory Journal, 42 : 802-825, 2013.

[4] Young-shin Youn, Chan-youngPark, Jong-daeKim, Hye-jeongSong, Yu-seopKim. "New Biomarker Discovery of Specific Disease using Word Embedding", IMETI, 2016.

[5] Byeong-Hun Yoon, Young-shin Youn, Hye-jeongSong, Jong-dae Kim, Chan-young Park, Yu-seopKim. " Correlation Analysis of BioMedical Entities using Word Embedding", ICBEI, 2017.

[6] T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean, "Distributed representations of words and phrases and their compositionality", Advances in neural information processing systems, 2013.

[7] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE.", Journal of Machine Learning Research 9.Nov, 2579-2605, 2008.

[8] Huarte Maite, Guttman Mitchell et al, "A Large Intergenic Noncoding RNA Induced by p53 Mediates Global Gene Repression in the p53 Response", CELL, Volume 142, Issue 3, 2010.

[9] Zhang H, Berezov A, Wang Q, Zhang G, Drebin J,

Murali R, Greene MI, "ErbB receptors: from oncogenes to targeted cancer treatment", The Journal of Clinical Investigation, 117 (8): 2051-8, 2007.

[10] T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean, " Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781, 2013.