암 예측의 정확성을 위한 특성 합성 방법

신승연*, 김현진*, 박상현*
*연세대학교 컴퓨터과학과
e-mail : guguing@yonsei.ac.kr

A Method for Synthesizing Features for the Accuracy of Predicting Cancer

SeungYeon Shin*, Hyunjin Kim*, Sanghyun Park*
*Dept. of Computer Science, Yonsei University

요 약

machine learning 기법 중 하나인 logistic regression 을 이용하여 benign sample 과 breast cancer sample 을 구분할 수 있는데, 이 연구를 통해 classification 의 정확도를 높이고 false positive 와 false negative 의 비율을 줄이려고 했다. 그래서 logistic regression 의 parameter 값을 바탕으로 regression function 에 영향을 많이 주는 feature 들을 선택하고, 영향력 있는 feature 들을 더한 새로운 feature 를 추가했다. 그 결과 정확도와 F-score 가 증가했으며, false positive, false negative 의 비율이 감소했다.

1. 서론

방대한 양의 생물학적 데이터로부터 의미를 찾아내려고 하는 시도가 최근 몇 년 동안 증가하고 있고, 컴퓨터 과학이나 통계학의 영역이었던 데이터 분석 기술이 바이오 분야에 도입되고 있다. 바이오 데이터 분석의 목적은 질병을 정확하게 진단하고 치료하는 것인데, 이 연구에서는 유전자 발현 데이터에 logistic regression 을 적용해서 sample 의 breast cancer 여부를 예측해볼 것이다. 그리고 logistic regression 을 적용한 원래 결과의 정확성을 향상시키기 위해 새로운 feature 를 추가하는 방법에 대해 논의할 것이다.

2. 관련 연구

machine learning 의 한 분야인 supervised learning 으로 해결할 수 있는 문제에는 regression problem 과 classification problem 이 있다. regression 은 주어진 데이터의 경향성을 분석해서 새로운 데이터의 값을 예측하는 방법이고, classification 은 데이터와 각각의 데이터들이 속한 group 이 주어졌을 때 새로운 데이터가 어떤 group 에 포함될지 예측하는 것이다. 이 중에서 classification problem 을 해결하기 위한 학습 모델에는 logistic regression, neural network 와 같은 여러 가지가 있다. decision tree 나 SVM 등 여러 가지 machine learning 기법을 이용해서 cancer classification 을 시도한 논문들[1][2]도 있었지만, 이 연구에서는 우선 직접 구현하기 쉬운 logistic regression 을 사용했다.

※ 이 논문은 2015 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2015R1A2A1A05001845).

일반적인 polynomial regression 과 같이, logistic regression 에도 데이터의 경향성을 나타내는 hypothesis function(regression function)이 존재한다. 하지만 logistic regression 에서는 test 데이터에 sigmoid function 을 적용한 결과값이 1 에 가까우면 positive class, 0 에 가까우면 negative class 로 분류한다.

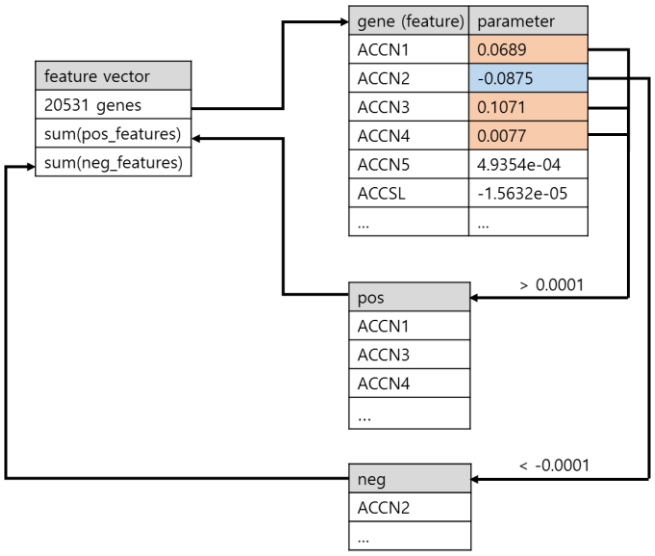
hypothesis function $h_\theta(x) = g(\theta^T x)$
sigmoid function $g(z) = \frac{1}{1+e^{-z}}$

위의 수식에서 theta 는 learning model 의 parameter vector 를 나타내고, x 는 feature 의 vector 를 나타낸다. 이 연구에서 logistic regression 의 모든 과정은 Matlab 을 이용해 직접 구현하였으며, parameter learning 에서는 gradient descent 알고리즘을 사용했다.

3. 방법

연구를 수행할 데이터는 TCGA(The Cancer Genome Atlas)에서 얻은 RNA-Seq gene expression data 이다. regression 의 결과값을 계산할 때 parameter 에 feature 의 값을 곱해서 더하는데, 이때 parameter 의 절댓값이 크다면 그 parameter 에 해당하는 feature 는 다른 feature 들보다 결과값에 큰 영향을 주게 된다. 따라서 영향력이 큰 feature 들을 더해서 새로운 feature 를 만든다면 classifier 의 정확성이 더 향상될 것이라고 예상했다. logistic regression training 이 끝난 parameter 중에서 절댓값이 0.001 보다 큰 feature 들을 선택했다. parameter 가 양수인 feature 와 음수인 feature 를 한꺼번에 더한다면 상쇄되어서 영향력이 감소하거나 예상하지 못한 결과가 나올 수 있기 때문에 두 종류의 feature 를 따로 계산했다. 선택한 feature 중에서

parameter 가 양수인 feature 들을 모두 더해서 첫 번째 새로운 feature 를 만들었고, parameter 가 음수인 나머지 feature 들을 모두 더해서 두 번째 새로운 feature 를 만들었다. 2 개의 새로운 feature 들을 기존의 feature vector 에 추가하고 logistic regression training 을 다시 수행하여 parameter 를 재설정했다.



(그림 1) 전체적인 작업의 흐름

4. 결과

<표 1>과 <표 2>에서 두 결과를 비교하였다. <표 1>은 feature 를 추가하지 않고 logistic regression 을 수행한 결과이고, <표 2>는 2 개의 feature 를 추가한 다음에 계산한 logistic regression 결과다. 각각의 표에서 처음 10 개의 행은 10-fold cross validation 으로 10 번 testing 한 결과값이다. 마지막 행은 이 결과값들의 평균을 나타낸다. 6 개의 열은 순서대로 true positive(TP), false positive(FP), true negative(TN), false negative(FN), precision, recall 값이다.

<표 1>feature 를 추가하지 않은 regression 의 결과

TP	FP	TN	FN	Precision	Recall
110	2	9	0	0.9821	1.0000
110	0	11	0	1.0000	1.0000
110	2	9	0	0.9821	1.0000
109	1	10	1	0.9909	0.9909
108	0	11	2	1.0000	0.9818
109	0	11	1	1.0000	0.9909
110	0	11	0	1.0000	1.0000
109	1	10	1	0.9909	0.9909
106	1	10	4	0.9907	0.9636
109	0	11	1	1.0000	0.9909
109	0.7	10.3	1	0.9937	0.9909

gradient descent 알고리즘의 iteration 횟수를 10 번으로 설정했는데도 98%가 넘는 정확도를 보였다. <표 1>에서 정확도는 98.6%, F-score 값은 0.9923 이다.

<표 2> feature 를 추가한 다음의 classification 결과

TP	FP	TN	FN	Precision	Recall
110	1	10	0	0.9910	1.0000
110	2	9	0	0.9821	1.0000
110	0	11	0	1.0000	1.0000
110	0	11	0	1.0000	1.0000
109	0	11	1	1.0000	0.9909
109	0	11	1	1.0000	0.9909
110	0	11	0	1.0000	1.0000
109	0	11	1	1.0000	0.9909
108	0	11	2	1.0000	0.9818
110	0	11	0	1.0000	1.0000
109.5	0.3	10.7	0.5	0.9973	0.9955

<표 2>에서는 새로운 feature 2 개를 더했을 뿐이지만 <표 1>의 결과보다 정확도와 precision, recall 값이 높았다. feature 를 추가했을 때 정확도는 99.3%, F-score 값은 0.9964 이다. 따라서 feature 를 추가하면 cancer 를 benign 으로, benign 을 cancer 로 잘못 예측하는 경우가 줄어들기 때문에 임의의 유전자 데이터가 주어졌을 때도 classification 을 정확하게 수행할 수 있다.

5. 결론

이 연구의 목표는 benign sample 과 cancer sample 을 구분하는 classifier 의 정확성을 향상시키는 것이다. 그래서 영향력이 있는 feature 를 더해 새로운 feature 를 만드는 방법을 제안했고, 10-fold cross validation 으로 testing 한 결과의 평균 precision, recall, accuracy 의 값을 계산했다. <표 1>과 <표 2>의 결과를 보면 feature 2 개를 추가하는 것만으로도 정확도가 향상되었으며, F-score 의 값도 증가했다. 따라서 feature 를 추가하는 방법이 정확성 측면에서 효과가 있다는 것을 알 수 있고, 그렇기 때문에 새로운 sample 이 주어졌을 때 cancer 여부를 더 정확하게 예측할 것이다.

참고문헌

[1] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning, 2002.
[2] Tan, Aik Choon, and David Gilbert. "Ensemble machine learning on gene expression data for cancer classification.", 2003.
[3] Russell, Stuart; Norvig, Peter. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall, 2003.
[4] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2015.
[5] Hosmer Jr, David W., and Stanley Lemeshow. Applied logistic regression. John Wiley & Sons, 2004.
[6] King, Gary, and Langche Zeng. "Logistic regression in rare events data." Political analysis, 2001, 9.2: 137-163.
[7] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.