건강검진 데이터 기반 흡연자 분류를 위한 모형별 성능 분석

윤지선, 유헌창
고려대학교 컴퓨터정보통신대학원
e-mail : {2930sun, yuhc}@korea.ac.kr

Performance Evaluation between Models for Smoker Classification Based on Health Examination Data

Jisun Yun, Heonchang Yu
Graduate School of Computer & Information Technology, Korea University

요 약

흡연여부를 감별하는 지표가 있지만 반감기 등 여러 가지 요인에 따라 결과가 변한다는 단점이 있다. 그렇기 때문에 흡연여부 감별 시 외부요인에 영향을 덜 받는 지표가 필요하게 되었다. 그래서 흡연여부 감별하는데 적합한 모형을 찾아 외부요인에 영향이 적은 지표를 개발에 도움이 될 것을 기대하며 연구를 진행하였다. 실험은 국민건강보험공단에서 제공한 건강검진정보데이터를 기반으로, SVM, Logistic Regression, KNN 등의 머신러닝 모델을 이용하여 흡연 여부를 감별하는 것을 진행한다. 이 실험은 속성에 따른 모형의 성능변화와 학습데이터 수에 따른 모형의 성능변화에 대한 2가지 측면에서 모델의 성능을 측정하였다. 모델의 평가는 정확도(accuracy), 정밀도(precision), 재현율(recall), 조화 평균(f1-score)으로 진행하였으며, 약 70퍼센트 정도의 정확도와, 60퍼센트 대의 재현율을 보인다.

실험 결과, SVM이 속성에 따른 모형의 성능 변화 실험에서는 63%의 재현율, 학습데이터 수에 따른 성능 변화 실험에서는 68%의 재현율을 보여, 흡연자 판별에 가장 좋은 성능을 보였다. 또한 재현율을 기준으로 실험 차수별로 가장 좋은 성능을 보인 모델과 가장 저조한 성능을 보인 모델의 차이를 비교한 결과, '속성에 따른 모형의 성능 변화 실험'에서는 최고 36%의 차이를 보였으며, '학습데이터 수에 따른 성능 변화 실험'에서 최고 42%의 차이를 보여 주었다. 이에 판별을 위한 속성도 중요하지만, 적합한 모형 선택 또한 중요하다는 것을 확인하였다.

1. 서론

사람들이 건강에 관한 관심이 증가됨에 따라 금연에 대한 관심도 증가하였다. 이에 정부에서도 국민건강 증진을 위해 금연사업을 진행하고 있다. 그렇기 때문에 흡연여부 감별을 위한 여러 가지 지표가 개발되었다. 이러한 지표 중 일산화탄소는 반감기가 4시간으로 짧고 매연, 운동 등이 예측 결과에 영향을 줄 수 있다. 또 다른 지표인 니코틴은 담배에 특징적인 물질로서 가장 적합한 지표이지만, 반감기가 1~2시간으로 일산화탄소보다 더 짧다는 단점이 있다. 그렇기 때문에 흡연여부 감별 시 외부요인에 영향을 덜 받는 지표가 금연사업에 필요하게 되었다. 그래서 국민건강보험공단에서 제공하는 건강검진데이터에 항상성을 가진 속성이 존재하다면, 이 데이터를 이용하여 흡연 여부 감별하는데 적합한 모형을 찾아 외부요인에 영향이 적은 지표를 개발하는데 도움이 되고자 연구를 진행하였다.

모형별 성능 평가를 위해 SVM, Logistic Regression, KNN, Decision tree, RandomForest, GradientBoosting, MLP 모형을 이용하였다. 실험은 속성에 따른 모형의 성능과 학습데이터 수에 따른 모형의 성능을 측정하였다. 또한 각 모형별 성능평가는 내부평가 방법인 정확도(accuracy), 정밀도(precision), 재현율(recall), 조화 평균(f1-score)값을 확인하여 모형의 성능을 확인하였다.

2. 관련연구

[1]은 유전자 발현 정보를 정량적인 수치로 제공하는 마이크로어레이(microarray) 데이터 중, 백혈병에 대한 데이터를 정규화하였다. 이렇게 정규화된 데이터에서 Information Gain, Gini Index, One-dimensional Support Vector Machine, T-statistic 방법을 이용하여 특징을 추출하여 학습데이터로 이용하였다. 준비된 학습데이터를 이용하여 Naive Bayes, KNN, Decision Tree, Support Vector Machine, Neural Network 알고리즘을 적용하여 종양 분류 모델을 구축하고 성능평가를 하였다. 평가결과를 통해 종양 분류를 위한 특징추출은 Information Gain을 사용하고 분류기법으로 SVM 알고리즘을 제안하였다[1]. [2]는 간암 진단을 위한 최적의 분류모형을 제안하여 간암 초기진단에 도움이 되고자 연구를 진행하였다.

예측모형으로 Logistic Regression, CART, Neural Network를 사용하였다. 여기에 단일 모형의 성능을 높이기

위해 앙상블 알고리즘을 적용하였다. 또한 실험인자로 앙상블 기법 그리고 분류기의 개수를 적용하고 반응변수로는 분류 정확도, 민감도, 그리고 특이도를 값으로 하여 삼원배치법 실험을 실시하였다. 실험한 결과를 분산 분석과 던칸(Duncan) 검정을 이용하여 분석한 결과를 소개하였다[2]. [3]은 심혈관이나 관상 동맥 심장질환이 없는 환자의 위험 인자로 부터 위험도를 평가하고, 향후 10년 내 당뇨병 및 심장질환이 발생할 위험도를 예측하였다. 여기에 예측의 성능을 높이기 위하여 SVM을 사용하였으며, 이를 검증하기 위하여 SVM을 사용한 회귀방법과 사용하지 않은 회귀 방법의 성능을 실험한 결과를 소개하였다[3]. [4]는 한국인 유방암 환자의 예후 인자 분석과 보조적 화학치료를 위한 환자군 선별을 위한 모형을 Decision tree를 이용하여 구현하였다. 이 모형의 성능 평가는 서울대병원 외과에서 수술 받은 유방암 환자 중 림프절 전이가 없는 참윤성 관암 환자의 임상병리학적 결과와 추적관찰 중 원격전이 유무를 관찰하여 평가한 결과를 소개하였다[4].

3. 실험환경

국민건강보험공단에서 제공한 2016년 건강검진정보데이터의 속성 34개를 이용하였다. 속성은 <표 1>에 정리하였다.

<표 1> 제공된 속성정보 34개

기본속성(22개)	성별코드, 연령대코드(5세단위), 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력(좌),시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당(공복혈당), 총콜레스테롤, 트리글리세라이드, HDL콜레스테롤, LDL콜레스테롤, 혈색소, 요단백, 혈청크레아티닌, (혈청지오티)AST, (혈청지오티)ALT, 감마지티피, 흡연상태
실험에 제외한 속성(11개)	기준년도, 시도코드, 결손치유무, 치아마모증유무, 제3대구치(사랑니) 이상, 치석, 음주여부, 데이터 공개일자, 가입자일련번호, 구강검진 수검여부, 치아우식증유무,

결측치 값을 가진 속성, 단일값을 가진 속성 그리고 흡연 여부 판별에 불필요한 속성을 학습데이터에서 제외하였다. 총 22개 속성을 학습데이터로 이용하였으며 이중 흡연상태를 타겟 속성으로 사용하였다. 또한 실험에는 기계학습 모형인 SVM, Logistic Regression, KNN, Decision tree, 앙상블 모형 RandomForest, GradientBoosting 그리고 퍼셉트론 기반의 모형인 MLP 등 7개의 모형을 실험에 사용하였다.

흡연 여부 감별에 적합한 모형을 학인하기 위해 속성에 따른 모형의 성능과 학습데이터 수에 따른 성능을 확인하기 위해 실험을 진행하였다. 속성에 따른 모델의 성능을 측정하기 위해 학습데이터의 속성을 추가 또는 제거를 하면서 총 6번 실험을 수행하였다. 또한 학습데이터 수에 따른 성능을 측정하기 위해 학습데이터의 수를 늘려가면서 총 5번의 실험을 진행하였다.

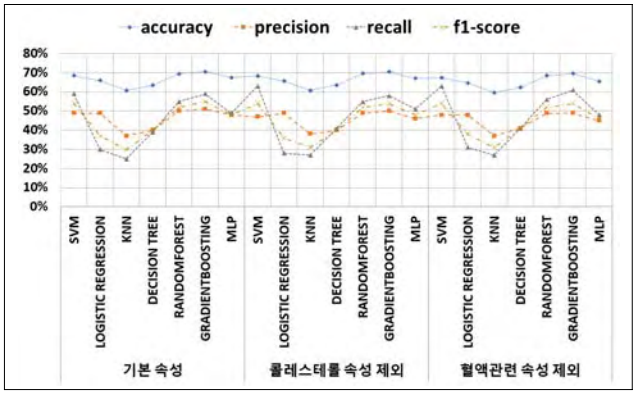
4. 속성에 따른 모형의 성능 변화 실험 결과

속성에 따른 모형의 성능 변과 실험은 <표 2>에 따라 속성을 변화시키며 실험을 진행하였다.

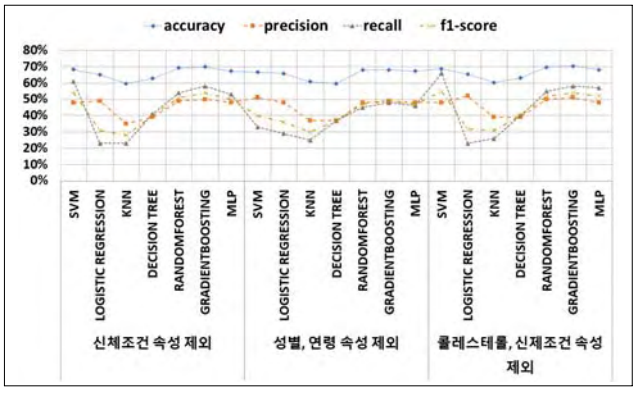
<표 2> 속성에 따른 모형의 성능 변화 실험 정보

실험차수	실험내용
1번째 실험 (기본속성)	실험조건에 따라 11개의 속성을 제거하여 22개 속성변수와 1개의 타겟 변수로 진행
2번째 실험 (콜레스테롤 제외)	기본속성에 콜레스테롤 속성(3개)을 제거하여 19개의 속성변수와 1개의 타겟변수로 진행
3번째 실험 (혈액관련 속성제외)	기본속성에 혈액검사결과 속성(5개)을 제거하여 17개 속성변수와 1개의 타겟변수로 진행
4번째 실험 (신체조건 속성 제외)	기본속성에 신체조건 속성(6개)을 제거하여 16개의 속성변수, 1개의 타켓변수로 진행
5번째 실험 (성별, 연령 속성 제외)	기본속성에 성별 및 연령 속성(2개)을 제거하여 20개 속성변수와 1개의 타겟변수로 진행
6번째 실험 (콜레스테롤과 신체조건 속성 제외)	콜레스테롤과 신체조건 속성(9개)를 제거하여 13개의 속성 변수와 1개의 타겟변수로 진행

(그림 1)과 (그림 2)는 속성에 따른 각 모델별 정확도, 정밀도, 재현율, 조화 평균 값을 통해 각 모델 별 성능을 측정한 결과를 나타내는 그래프이다.



(그림 1) 속성에 따른 모형의 성능 변화(1)



(그림 2) 속성에 따른 모형의 성능 변화(2)

속성에 따른 모형의 성능 변화 실험에 사용한 7개 모형별 최고치 성능을 비교하였다. 그 결과 정확도는 최저 61%(KNN)에서 최고 70%(dandomforest, gradientboosting)로 9%차이를 보였다. 정밀도는 최저 39%(KNN)에서 최고 52%(Logistic Regression)로 13%차이를 보여 주었다. 또한 재현율은 최저 27%(KNN)에서 최고63%(SVM)로 36% 차이를 보여주고 있다. 마지막으로 조화 평균은 최저 31%(KNN)에서 최고 55%(SVM, gradientboosting)로 24%의 차이를 보여주고 있음을 실험을 통해 확인하였다. <표 3>은 속성에 따른 모형별 최고치를 <표 4>는 속성에 따른 모형별 성능의 최저치를 표로 나타냈다.

<표 3> 속성에 따른 모형별 최고치 성능

모형	정확도	정밀도	재현율	조화 평균
SVM	69%	51%	**63%**	**55%**
Logistic Regression	66%	**52%**	31%	38%
KNN	**61%**	**39%**	**27%**	**31%**
decision tree	64%	41%	41%	41%
randomforest	**70%**	50%	56%	52%
gradientboosting	**70%**	51%	61%	**55%**
MLP	68%	48%	57%	52%

<표 4> 속성에 따른 모형별의 성능 최저치

모형	정확도	정밀도	재현율	조화 평균
SVM	67%	47%	33%	40%
Logistic Regression	65%	48%	23%	31%
KNN	60%	35%	23%	28%
decision tree	60%	37%	37%	37%
randomforest	68%	48%	45%	47%
gradientboosting	68%	49%	48%	49%
MLP	65%	45%	46%	46%

<표 5>는 속성에 따른 모형의 성능의 최저치와 최고치의 차이를 보여 준다.

<표 5> 속성에 따른 모형의 성능별 차이

모형	정확도	정밀도	재현율	조화 평균
SVM	2%	4%	30%	15%
Logistic Regression	1%	4%	8%	7%
KNN	1%	4%	4%	3%
decision tree	4%	4%	4%	4%
dandomforest	2%	2%	11%	5%
gradientboosting	2%	2%	13%	6%
MLP	3%	3%	11%	6%

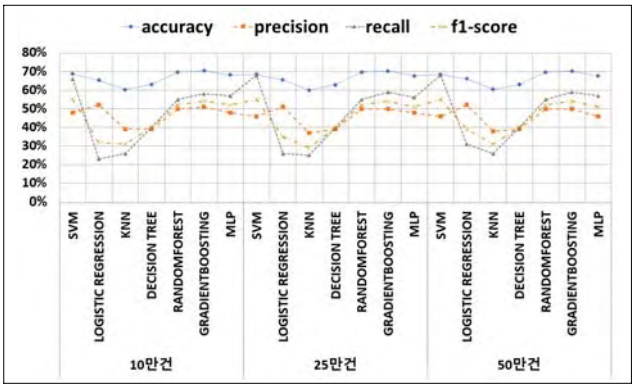
속성에 따른 모형별 성능의 최저치와 최대치의 차를 확인하였다. 그 결과 정확도는 최저 1%(Logistic Regression, KNN)에서 최고4%(decision tree) 차이를 보여주고 있다. 정밀도에서는 최저 2%(randomforest, gradientboosting, MLP)에서 최고 4%(SVM ,logistic regression, KNN, decision tree)차이가 나타나고 있다. 또한 재현율은 최저 4%(KNN, decision tree)에서 최고30%(SVM) 차이를 보이고 있다. 마지막으로 조화 평균은 최저 3%(KNN)에서 최고 15%(SVM)의 차이가 나타나는 것을 실험을 통해서 확인하였다.

모형별 성능을 확인했을 때, 재현율의 경우 SVM이 63%로 가장 좋은 성능을 보이는 것을 확인하였다. 또한 성별과 연령속성을 제외 한 결과 33%까지 재현율이 하락하여 흡연자를 분류하는데 중요한 속성임을 확인하였다. 그리고 콜레스테롤 수치와 신체조건 속성을 학습데이터에서 제외한 경우 SVM의 재현율이 66%까지 향상되어 해당 속성이 흡연자 예측에 불필요한 속성임을 실험을 통해서 확인하였다.

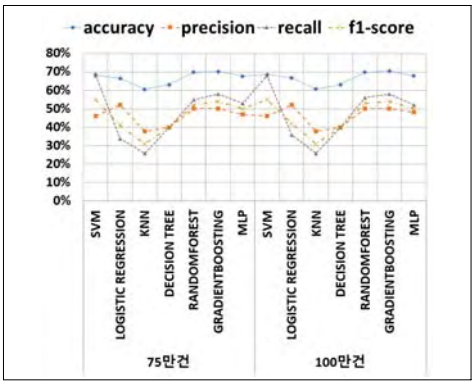
5. 학습데이터 수에 따른 성능 변화 실험결과

속성에 따른 모형의 성능 변화 실험의 6번째 실험 조건(콜레스테롤과 신체조건 속성 제외)을 기준으로 실험을 진행하였다. 다만, 학습데이터 수를 10만, 25만, 50만, 75만, 100만건으로 변경하여 모델별 성능을 측정하였다.

(그림 3)과 (그림 4)는 학습데이터 수에 따른 성능 변화를 실험한 결과를 나타낸 그래프이다.



(그림 3) 학습데이터 수에 따른 성능 변화(1)



(그림 4) 학습데이터 수에 따른 성능 변화(2)

학습데이터 수에 따른 성능 변화 실험에 사용한 7개 모형별 최고치 성능을 비교하였다. 그 결과, 정확도는 최저 61%(KNN)에서 최고 70%(randomforest, gradientboosting)로 9%차이를 보였다. 정밀도는 최저 39%(KNN)에서 최고 52%(Logistic Regression)로 13%차이를 나타내고 있다. 또한 재현율은 최저 26%(KNN)에서 최고68%(SVM)로

42%차이를 보이고 있다. 마지막으로 조화 평균은 최저 31%(KNN)에서 최고 55%(SVM)로 24%의 차이를 보여주고 있음을 실험을 통해 확인하였다.

<표 6>은 학습데이터 수에 따른 모델별 성능 최고치를 정리한 표이다.

<표 6> 학습데이터 수에 따른 모델별 성능 최고치

모형	정확도	정밀도	재현율	조화 평균
SVM	68%	48%	**68%**	**55%**
Logistic Regression	66%	**52%**	36%	42%
KNN	**61%**	**39%**	**26%**	**31%**
decision tree	63%	40%	40%	40%
randomforest	**70%**	50%	56%	53%
gradientboosting	**70%**	51%	59%	54%
MLP	68%	48%	57%	52%

<표 7>은 학습데이터 수에 따른 모델별 성능 최저치를 정리한 표이다.

<표 7> 학습데이터 수에 따른 모델별 성능 최저치

모형	정확도	정밀도	재현율	조화 평균
SVM	68%	46%	66%	55%
Logistic Regression	65%	51%	23%	32%
KNN	60%	37%	25%	29%
decision tree	63%	39%	40%	40%
randomforest	70%	50%	55%	52%
gradientboosting	70%	50%	58%	54%
MLP	67%	46%	52%	50%

<표 8>은 학습데이터 수에 따른 모델별 성능 최고치와 최저치의 차를 정리한 표이다.

<표 8> 학습데이터 수에 따른 모델별 성능 차이

모형	정확도	정밀도	재현율	조화 평균
SVM	0%	2%	2%	0%
Logistic Regression	1%	1%	13%	10%
KNN	1%	2%	1%	2%
decision tree	0%	1%	0%	0%
randomforest	0%	0%	1%	1%
gradientboosting	0%	1%	1%	0%
MLP	1%	2%	5%	2%

학습데이터 수에 따른 모형별 성능의 최저치와 최대치를 차이를 확인하였다. 그 결과, 정확도는 모형 별 최저 0%(SVM, decision tree, randomforest, gradientboosting)에서 최고1% (Logistic Regression, KNN, MLP) 차이를 보이고 있다. 정밀도는 최저 0%(randomforest)에서 최고 2%(SVM, MLP) 차이를 나타내고 있다. 또한 재현율은 최저 0%(decision tree)에서 최고13%(Logistic Regression) 차이를 보이고 있다. 마지막으로 조화 평균은 최저 0%(SVM, decision tree, gradientboosting)에서 최고 10%의 차이를 보여주고 있음을 실험을 통해 확인하였다. 관련 내용은 <표 6>, <표 7>그리고 <표 8>에 정리하였다.

재현율의 경우 SVM이 68%로 가장 좋은 성능을 보였으나, 학습데이터 수에 따라 각 모델별 성능의 차이는 크지 않은 것으로 이번 실험을 통해 확인하였다.

그리고 모델 중 logisc Regression이 최저 23%에서 최고 36%로 13%의 편차를 보여 데이터 수에 따라 예측결과에 영향을 가장 많이 받는 모형임을 확인하였다.

6. 결론

속성에 따른 모형의 성능 변화 실험에서 SVM은 63%의 재현율을 보였다. 또한 학습데이터 수에 따른 성능 변화 실험에서도 SVM 모형이 68%의 재현율을 보였다. 이 실험 결과를 통해 7개 모형 중 SVM이 실제 흡연자를 판정하는 성능이 가장 좋은 성능을 보이는 것을 확인하였다.

2개 실험의 차이를 재현율 기준으로 보면, '속성에 따른 모형의 성능 변화 실험'에서는 최고 36%의 차이를 보이는 것을 확인했다. 그리고 '학습데이터 수에 따른 성능 변화 실험'에서는 최고 42%의 차이를 확인하였다. 이 실험 결과, 타겟을 판정하기 위한 속성도 중요하지만 적합한 모형 선택이 더 중요하다는 것을 확인하였다.

실험에 사용한 실험데이터가 공공데이터 개방 정책에 따라 제공된 데이터를 이용하였다. 그래서 타겟 분류를 위한 속성이 부족하였다. 또한 의학에 대한 지식이 없기 때문에 속성 선택에 한계점이 존재하였다. 이런 내용을 보완하여 추가 연구를 진행하다면 외부요인에 영향이 적은 흡연여부 감별 지표 개발 시 활용이 가능할 것으로 예측된다.

참고문헌

[1] 박윤정, "종양 분류를 위한 특징 추출과 분류기법의 성능분석", 2005.,이화여자대학교 과학기술대학원

[2] 이우선, "Ensemble algorithm을 이용한 간암진단의 분류분석", 2002., 연세대학교 대학원

[3] 오명환,정용규, "심장질환 및 당뇨병의 재발 위험요인의 예측을 위한 SVM 기반 회귀분석의 성능비교", 한국IT마케팅학회 학술대회/2014(1), 2014., 65-67, 한국IT마케팅학회

[4]정소연, "유방암의 임상병리학적 muiti-marker와 Decision tree를 이용한 원격 전이의 예측 모델 개발", 2007., 서울대학교 대학원