한국빅데이터학회지
제2권 제2호, 2017, pp. 115-128

COPD 코호트 자료에서의 Machine Learning 방법론 비교

Comparison of Machine Learning Methodology in COPD Cohort Data

정현명[1] · 박헌진[1†] · 이진국[2] · 이종민[2]

인하대학교 통계학과[1], 서울성모병원[2]

요 약

최근 머신러닝 방법은 높은 예측력과 함께 널리 이용되지만 머신러닝을 제대로 활용하기 위해서 데이터가 가진 한계를 통계적 기법으로 해결한다면 기존보다 더 높은 예측력을 이끌어 낼 수 있다. 본 연구에서는 Longitudinal and Imbalanced Data에서 SMOTE 방법을 활용하여 불균형 문제를 해결한 결과 예측력이 증가하는 것을 확인할 수 있었다. 추가적으로 만성폐쇄성폐질환 급성악화 관련 연구가 활발히 이루어지고 있지만 급성악화와 관련 있는 요인을 찾는 연구만 이루어지고 있어 여러 요인들에 대한 복합적인 관철과 예측모형을 통한 급성악화 예측 연구는 이루어지지 않는다. 본 연구에서는 여러 요인을 같이 살펴봤을 때 어떤 요인들이 만성폐쇄성폐질환 급성악화와 관련이 있는지 확인하고 개인 맞춤형 특정 질환 예측 모형을 구축하였다.

■ 중심어 : 머신러닝, 만성폐쇄성폐질환, 질환예측

Abstract

Recently, Machine Learning Methods are widely used with high prediction performance. But if the limit of the data is solved by the statistical technique, It can, lead to higher prediction performance than the existing one. In this study, the SMOTE method is used to solve the imbalance problem in the longitudinal and imbalanced data. As a result, It, was confirmed that the prediction performance increases. Additionally, Although, studies on COPD have been actively conducted, only studies that are related to acute exacerbation have been conducted. So there are no studies on the prediction of acute exacerbation through multiple perspectives and predictive models for various factors. In this study, We examined the factors related to acute exacerbation of COPD and constructed a personalized specific disease prediction model.

■ Keyword : Machine Learning, Chronic Obstructive Pulmonary Disease(COPD), Specific Disease Prediction

Ⅰ. 서 론

Chronic Obstructive Pulmonary Disease(COPD)란 만성 폐쇄성 폐질환이라고 일컬으며 비가역적인 기류제한을 특징으로 하는 폐질환[2, 21]으로 만성염증을 동반하여 폐실질 손상을 일컫는다. COPD는 현재 세계에서 사망률 3위[2, 22], 우리나라에서는 사망률 7위에 이르고 특히 80세 이상에서는 전체 사망원인 중 5위를 차지[23]하고 있는 매우 위험한 질병이며 유병률 또한 매우

2017년 11월 27일 접수; 2017년 12월 18일 수정본 접수; 2017년 12월 26일 게재 확정
† 교신저자 hjpark@inha.ac.kr

높아 2012년 기준 40세 이상에서는 14.6%의 유병률을 보이며 남성의 유병률은 23.4%, 여성의 경우에는 7.9%로 특히 남성에게 많이 발생하는 질병이다[23]. 또한 70세 이상의 남성 환자의 경우 유병률이 38.4%[2]로 매우 높은 것을 알 수 있다. 하지만 이런 상황에도 불구하고 우리나라에서는 아직 질환 자체에 대한 인지도가 매우 낮아 실제 환자 중 2.9%만이 질환을 인지[2]하고 있는 실정이다. 이에 따라 대한결핵 및 호흡기학회에서 2014년 COPD 치료를 위한 Guideline을 제시하며 COPD 질병 치료에 대한 중요성을 나타내고 있다. 특히 COPD 질병에서 가장 조심해야 하는 것은 바로 급성악화이다[4, 11]. 급성악화의 종류는 Mild, Moderate, Severe 세 가지로 구분되며 COPD 질병 사망의 주 요인[4]인데 그 이유는 Aucte Exacerbation(급성악화) 발생 시 급격히 폐 기능이 떨어지기 때문에 호흡곤란으로 인해 사망할 수 있다[4, 11]. 만약 이런 특징을 지닌 급성악화를 미리 예견하고 치료를 준비하거나 대비책을 마련 한다면 급성악화로 인한 사망을 줄일 수 있을 것이라 기대할 수 있다. 위와 같은 급성악화에 대한 사망률을 줄이기 위해 본 연구에서는 COPD 급성악화 종류 중 Moderate, Severe 두 급성악화 정의에 대한 예측 모델링을 개발 하였다.

본 연구 전 COPD 급성악화 연구는 대부분 급성악화에 영향을 주는 여러 요인들에 대해 파악할 뿐 사실상 예측 모델을 통해 급성악화를 미리 예견하는 연구는 이루어지지 않았다.

이런 COPD 급성악화 예측 모델링에 대한 연구가 이루어지지 않은 이유는 급성악화 예측 모델을 만들 수 있을 만큼의 충분한 데이터의 부재와 함께 COPD 질병이 만성이라는 성질을 지니고 있어 분석에 사용하는 대부분의 데이터는 환자들을 오랜 기간 Follow up 한 Cohort 데이터로 작성 되어 있어 반복측정으로 인한 데이터간 독립성 위배 문제로 인해 통계적인 모델링 구축이 어려운 점 때문 이였다. 특히 본 연구에서 추가적으로 발견한 문제는 Imbalance 문제로 COPD 유병 중 급성악화 발생률이 낮아 COPD로 인해 병원에 내원할 경우 불과 3% 정도만 급성악화로 인한 내원으로 확인되어 진다. 그렇기 때문에 Longitudinal 속성과 Imbalance 속성 두 가지를 해결하지 않은 이상 예측 모델을 만들기 어려움은 쉬이 예견할 수 있다.

결론적으로 본 연구에서 COPD의 급성악화를 예측하기 위해 Longitudinal 이면서 Imbalanced 데이터의 예측 모델링을 생성을 목적으로 하고 Imbalance 문제를 해결하기 위해 Under sampling과 SMOTE 방법+Under sampling을 실시하여 Imbalance 문제를 해결하고자 하고 Longitudinal 의 경우 데이터간의 Correlated 문제를 해결 할 수 있는 GEE(Generalized Estimating Equation)모델과 머신러닝 방법 중에는 별다른 가정이 필요 없고 예측력이 높은 랜덤포레스트와 DNN(Deep Neural Network) 방법을 이용하여 예측 모델을 생성 한 후 모델의 성능을 비교하였다.

Ⅱ. Data

2.1 Data Source

본 연구에서 사용한 데이터는 총 5가지로 건강보험심사평가원 데이터를 활용하여 성별, 나이, 과거 약제 사용이력, 과거 동반질환 이력 등을 나타내는 변수를 생성하였고, COPD 질병과 연관이 깊은 흡연력, 폐 기능 검사, 삶의 질 검사 등의 자료를 이용하기 위해 KOCOSS 데이터[1]를 활용하였다. 추가적으로 COPD 급성악화와 관련된 여러 요인들을 확인하고 예측 모델에 활용하기 위해 한국 기상청의 기상자료[26], 에어코리아에서 수집한 미세먼지 자료[27]와 질병관리본부에서 제공하는 주간 질병감시정보 중 하나인 인플루엔자 및 호흡기바이러스 주별

발생정보[25]를 활용하였다.

2.2 Data 활용

2.2.1 KOCOSS 환자 코호트

본 연구에서는 COPD 환자를 Follow up 한 KOCOSS 환자 코호트 자료를 활용하였다. 특히 KOCOSS 환자 코호트 자료에서 환자들의 폐 기능 검사 및 삶의 질 검사, 흡연력 등에 대한 정보를 획득 할 수 있었다. 이 중 폐 기능 검사에 대한 변수는 FEV1%값이 존재하는데 FEV1%란 FEV1 값을 가지고 reference(키, 나이, 인종에 따른 정상치)와 비교해서 몇 %인지 보는 수치로 기존 의학계에선 흔히 폐 기능 FEV1의 감소는 COPD 급성악화와 관련이 있다[6, 8, 17]고 알려져 있으며 삶의 질을 나타내는 CAT 점수의 경우 CAT 점수를 지속적으로 확인할 경우 COPD 급성악화의 위험성을 확인할 수 있다[20]는 것이 일반적인 의학계의 공통적인 의견이다.

이에 더하여 COPD와 가장 관련이 높은 요인 중 하나를 흡연력으로 꼽을 수 있다. COPD와 관련된 수많은 연구에서 COPD 질병과 급성악화에 가장 관련이 높은 요인이 흡연력[4]이라고 일컬으며 금연하는 것만으로도 COPD 급성악화 위험률을 줄일 수 있다[5]는 연구도 의학계에서 이견이 없는 연구 결과이다.

여러 문헌과 전문가의 조언에 따라 KOCOSS 연구 자료 중 폐 기능검사 자료인 FEV1%와 삶의 질 조사 CAT 점수, 흡연력 등을 각 환자들의 Base Line으로 설정하였다.

2.2.2 HIRA(건강보험심사평가원) 데이터

KOCOSS 환자 Cohort 자료와 더불어 본 연구에서 가장 중요한 데이터 중 하나는 바로 건강보험심사평가원의 명세서 데이터 이다. HIRA 데이터는 기존에 사용하는 표본데이터가 아닌 KOCOSS 코호트 자료의 대상자의 데이터를 추출하여 KOCOSS 데이터와 연계할 수 있으며 HIRA 데이터 또한 5년간 코호트 자료로 만들어 실질적으로 Follow-up 할 수 있는 데이터로 설정하였다.

본 연구에서 사용한 HIRA 데이터 테이블은 명세서일반 테이블, 진료내역 테이블, 수진자 상병 테이블, 처방전교부상세 테이블 등을 이용하였다. 명세서일반 테이블에서는 수진자들의 성별, 나이, 입원여부 및 해당 진료의 시점 등을 확인하였고 수진자 상병 테이블에서 해당 수진자가 어떠한 질병들로 병원에 내원했는지, 진료내역 테이블과 처방전교부상세 테이블을 통해서는 각 약 처방에 대해 알 수 있었고 이를 이용하여 본 연구에서 가장 중요한 요인인 과거 약제 사용이력과 과거 동반질환 유병 이력을 알 수 있었다.

추가적으로 건강보험심사평가원 데이터를 활용 할 경우 본 연구에서 예측하고자 하는 COPD 급성악화는 질병코드에서 나타나지 않으므로 질병 처방내역을 통해 유추할 수 있는 조작적 정의가 필요하였고 COPD 급성악화 발생 시 치료방법에 대한 정보를 바탕으로 COPD관련 질병코드로 인해 병원에 내원하여 systemic steroid를 처방받으면서 입원하는 경우나 응급실 내원하는 경우 또한 단순 병원 방문하는 경우를 모두 COPD 급성악화라고 정의하였다.

2.2.3 공공데이터

본 연구에서 환자들의 과거 약제 사용이력 및 동반질환 이력 등 다음으로 중요한 데이터는 바로 기타 공공데이터였다. KOCOSS 코호트 자료를 통해 얻을 수 있는 폐 기능검사나 삶의 질 검사, 흡연력 등에 대한 연구는 이미 많이 이루어져 있고 현재 COPD 질병에 대한 관심의 증가와 함께 최근에는 이런 연구데이터 이상의 효과를 살피기 위해 환경적 요인들을 중점적으로 살펴보는 연구가 나타나고 있다. 이런 트렌드에 맞춰 본 연구에서는 기상청의 기상데이터[26], 에

어코리아의 미세먼지데이터[27], 질병관리 본부의 주간 질병감시 정보 중 하나인 인플루엔자 및 호흡기 바이러스 주별 발생 정보[25] 등을 활용하여 현재 이루어지고 있는 여러 환경적 요인들에 대한 효과를 함께 살펴보았다.

기상데이터 관련하여서는 최저기온이 COPD 급성악화 발생 확률을 높인다[19]라는 연구결과가 가장 보편적인 의견을 보이고 있어 이를 확인하고자 기상데이터에서 최저기온 데이터를 활용하였고 급성악화 발생 일로부터 며칠간의 최저기온의 행태 또한 연관이 있을 수 있다고 판단하여 전날 최저기온, 해당 일로부터 3일 전까지의 최저기온 합, 해당 일로부터 5일 전까지의 최저기온 합, 해당 일로부터 7일 전까지의 최저기온 합, 해당 일로부터 3일전과의 최저기온 차, 해당 일로부터 5일 전과의 최저기온 차, 해당 일로부터 7일 전과의 최저기온 차 등의 파생변수를 생성하였고 체감기온과의 연관성을 살펴보기 위해 최저기온과 최대풍속의 곱 Interaction 효과를 추가하였다. 추가적으로 기상의 경향성을 나타내는 평균습도와 일사량 등의 변수까지 추가하여 전체적인 기상의 효과를 살펴보았다.

미세먼지의 경우에도 미세먼지의 농도와 COPD 급성악화의 발생 건수와 연관이 있다[18]는 연구결과 또한 부각이 되고 있고 특히 최근 중국 발 미세먼지 및 황사의 심각성과 함께 호흡기 질환의 적신호가 켜져 미세먼지에 대한 관심이 증가하고 있어 이러한 미세먼지의 영향을 살펴보기 위해 기상과 마찬가지로 해당 일로부터 전일 미세먼지 농도, 해당 일로부터 3일 전까지 미세먼지 농도 합, 해당 일로부터 5일 전까지 미세먼지 농도 합, 해당 일로부터 7일 전까지 미세먼지 농도 합에 대한 파생변수를 생성하였다.

마지막으로 호흡기 바이러스에 대한 노출이 COPD 급성악화나 증상과 연관이 있다[16]는 연구 결과를 활용하기 위해 질병관리 본부에서 제공하는 주간 질병감시 정보인 인플루엔자 및 호흡기 바이러스 주별 발생 정보를 활용하였다. 활용한 바이러스는 Korean National Institute of Health (KNIH), Activities of influenza adenovirus (ADV), parainfluenza virus(PIV), respiratory syncytial virus(RSV), influenza virus(IFV), human coronavirus(hCoV), human coronavirus(hRV), human bocavirus(hBoV) 그리고 human enterovirus (hEV) 까지 활용할 수 있는 모든 바이러스 정류에 대해 점검 해 보았다.

바이러스 관련 정보가 다른 공공데이터와 다른 점으로는 기상 데이터 및 미세먼지데이터의 경우 실시간으로 Open API 나 Web에서 실시간으로 현재 정보와 예보를 알 수 있는 것에 반해 바이러스 정보의 경우 주간 질병 감시 정보 보고서가 작성 되어야 알 수 있어 급성악화 발생 일로부터 멀게는 2주 정도의 지연이 파생변수 생성 시 이를 고려하였다.

기본적인 바이러스 검출률에 대한 정보는 2주 전에 대한 질병 정보이고 이를 이용하여 파생변수 생성 시 해당 일로부터 2주 전과 3주 전 바이러스 검출률의 합, 해당 일로부터 2주 전, 3주전, 4주 전, 5주 전의 바이러스 검출률의 합에 대한 파생변수를 생성하여 예측 시에도 실질적으로 활용할 수 있는 데이터를 구성하였다.

2.3 Data 특징

본 연구에서 사용된 주 데이터는 건강보험심사평가원 데이터로 건강보험심사평가원데이터를 기준으로 해당 수진자들의 KOCOSS 코호트 자료 및 기상, 환경, 바이러스 데이터를 Mapping하여 이용하였다. 이런 상황 속에서의 한계점은 바로 분석의 시점이 모두 건강보험심사평가원 데이터가 존재하는 즉 병원에 내원했을 경우이다. 하지만 본 연구에서 최종적인 목표는 병원에 내원을 하지 않은 시점에서도 예측을 실시하고자 했고 이를 위해서 건강보험심사평가원데이터를 좀 더 가공하여 병원에 내원하지 않은

시점까지 모든 데이터를 다시 생성하였다. 이 때 병원에 내원하지 않은 시점까지 고려하다보니 한 수진자당 5년치 데이터인 약 1825일의 데이터를 생성하였고 이 중 병원에 입원 해 있는 날은 병원에 내원한 이벤트가 지속되기 때문에 해당 날 만큼 제거하였다.

위와 같은 데이터 핸들링을 걸치면서 생긴 가장 큰 문제는 바로 Imbalance 문제였는데 본래 급성악화의 발생비율이 낮아 모든 시점으로 데이터를 확장 전에도 급성악화의 비율은 약 4%에 불과했지만 데이터를 모든 시점으로 확장한 결과 급성악화의 비율은 0.28%로 매우 낮아 심각한 Imbalance 문제를 발생시켰고 이를 해결하기 위해 Sampling 방법에서 Under sapling과 SMOTE 방법을 이용하였다.

추가적으로 본 연구에서 사용하는 데이터는 한 수진자의 매일매일 Follow-up하여 한 수진자당 5년치 데이터가 존재하여 데이터의 독립성이 위배되는 Longitudinal Data의 형태를 띠게 돼 이런 데이터 간의 독립성 위배를 해결하기 위해 통계적 모델인 GEE를 이용하였고 머신러닝 중에서도 독립성 가정이 필요 없는 의사결정나무 기반의 랜덤포레스트와 DNN 방법을 활용하여 데이터의 문제를 해결하였다.

Ⅲ. 연구 방법론

3.1 Sampling 방법론

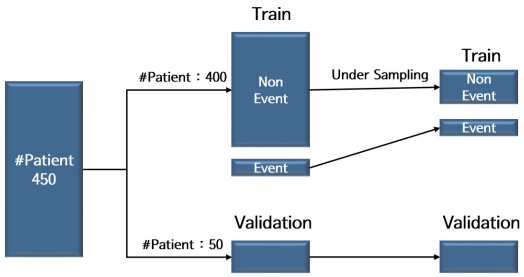
3.1.1 Data 분할

Longitudinal 데이터의 경우 데이터의 반복 문제를 해결하고자 군집을 지정할 수 있다. 그런 이유로 인하여 데이터분석 및 검증 시에도 군집 단위로 분석 및 검증하는 것이 일반적이기 때문에 본 연구에서도 훈련용 데이터와 검증용 데이터 구분 시 군집 단위인 환자 단위로 분석용 환자 셋과 검증용 환자 셋으로 구분지어 분석을 실시하였다. 본 연구에서 활용한 환자의 수는 총 450여 명이였고 400명을 훈련용 환자, 50명을 검증용 환자로 구분하였다.

3.1.2 Under Sampling

본 연구에서 사용한 데이터는 1의 비율이 매우 낮은 Imbalance 문제를 지니고 있었다. 환자당 5년 동안 입원한 날을 제외하고 모든 날짜에 대한 정보가 존재하고 COPD 급성악화 자체의 발생비율도 낮아 5년여 동안 COPD 급성악화 발생하는 비율이 극히 드물어 Imbalance 문제를 야기 시켰다.

Imbalanced Data가 가지는 문제는 많이 있지만 일반적으로 가장 크게 모델 자체의 변동성 때문에 모델의 성능이 안정적이지 못하는 문제[15]를 지니고 있다. 이를 해결하기 위한 일반적인 방법은 Under Sampling과 Over Sampling이 있다. 우선적으로 본 논문에서 머신러닝에 의한 시간적 문제를 해결하고자 Under sampling을 실시했고 Over sampling의 경우 단순한 Re sampling 방법을 이용하면 Over fitting을 야기 시킬 수 있는 문제[7] 때문에 Re sampling의 방법에 의한 Over sampling은 고려하지 않고 SMOTE 방법을 이용하였다.



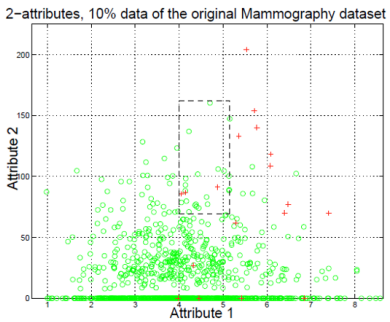
〈그림 1〉 Under-sampling with Cluster

3.1.3 SMOTE Method

SMOTE(Synthetic Minority Over-sampling Technique)란 Over sampling의 방법 중 하나로 기존의 Re sampling에 의해 발생하는 Over fitting을 줄여 예측력을 올리고자 하는 방법론 이다[7]. 이 방법론은 기본적으로 KNN 알고리즘을 이용하여

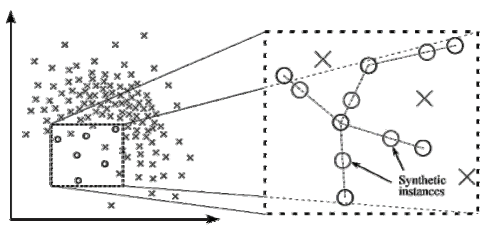
Minority 부분 사이에 새로운 데이터를 인조적으로 생성하여 예측력을 올리고자 하는 방법론이다.

<그림 2>를 봤을 때 만약 두 가지 속성을 통하여 Minority 공간과 Majority 공간을 구불 할 때 네모 칸에 안에 Minority 데이터를 생성 할 경우 좀 더 Grid를 정확히 구분하여 모델의 성능을 향상 시킬 수 있을 것이라는 아이디어에 착안한 방법이다.



〈그림 2〉 SMOTE : Minority Space in Decision Region

<그림 2>와 같이 공간 안에 새로운 Minority Data를 생성하기 위해서는 <그림 3>[24] KNN 방법을 통하여 Minority Data사이에 새로운 인조적인 데이터를 만들어 낼 수 있다.



〈그림 3〉 Generation of Synthetic Instanced with the help of SMOTE

추가적으로 본 연구에서 중요한 점은 바로 Longitudinal Data라는 점이다. 그렇기 때문에 Minority 간에 거리를 구할 때 같은 군집끼리 구하고 새로운 데이터를 생성할 때 또한 같은 군집의 Minority 안에 새로운 데이터를 생성해야 하기 때문에 SMOTE 방법 자체는 군집 단위로 이루어졌다.

3.1.4 Sampling 결과

본 연구에서 Sampling은 단순 Under sampling과 SMOTE 방법과 Under sampling의 결합을 활용하여 Sampling을 실시했고 결과는 <표 1>과 같다.

본 연구에서 SMOTE 방법의 경우 통계프로그램 SAS v9.4의 Modeclus Procedure를 이용하였으며 k = 10으로 설정을 하여 거리를 구한 후 거리 조절을 통해 Over sampling의 비율을 조절 하였기에 SMOTE 방법의 결과 200%, 300%, 400%는 대략적으로 늘린 1의 비율을 의미하고 실제적으로는 같은 군집 내의 거리를 200, 250, 300내의 데이터를 활용하여 새로운 데이터를 생성하였다.

〈표 1〉 Sampling Result

구 분	Target	빈도	백분율 (%)
Train-Data	0	667787	99.73
	1	1831	0.27
Under-sampling	0	40000	95.62
	1	1831	4.38
Under-sampling + SMOTE(200%)	0	40000	90.38
	1	4259	9.62
Under-sampling + SMOTE(300%)	0	40000	87.67
	1	5625	12.33
Under-sampling + SMOTE(400%)	0	40000	84.82
	1	7156	15.18
Validation-Data	0	83082	99.63
	1	310	0.37

3.2 통계적 모델

본 연구에서 사용한 통계적 기법은 GEE(Generalized Estimating Equation), 랜덤포레스트, DNN (Deep Neural Network) 세 가지를 이용하였다. 우선 Parametric Model인 GEE의 경우 Longitudinal

Data의 처리를 위해 사용하였고 머신러닝 방법 중 하나인 랜덤포레스트의 경우 의사결정나무 기반의 앙상블 기법이기 때문에 특별한 가정에 구애받지 않고 사용할 수 있었고 DNN의 경우 또한 마찬가지이다. 이런 연유로 세 가지 모델에 대해 모델링을 실시하고 결과를 비교하였다.

3.2.1 GEE(Generalized Estimating Equation)

의학계에선 흔히 지속적인 검진을 통해 시간의 흐름에 따라 환자의 상태가 어떻게 변화하는지에 관심이 많아 Longitudinal Data가 많이 다뤄지고 있다[13]. 이런 Longitudinal Data의 경우 Subject에 대한 반복이 측정되며 데이터 간의 Correlate 문제를 야기[13, 22]하기 때문에 이 문제를 해결해야 한다.

위와 같은 문제가 나타나는 데이터를 Repeated Data 혹인 Longitudinal Data라고 일컫는데 Repeated Data의 경우는 동일한 실험의 반복에 의해 생기는 경우를 일반적으로 나타내며 Longitudinal Data의 경우 시간의 경과에 따른 동일한 Subject에 여러 데이터가 생성되는 경우 일컫는다.

GLM에서 Mixed Effect를 통한 모델링은 주로 Repeated Data에서 사용하지만 Longitudinal Data에서도 이용이 가능하다. 하지만 GLM의 경우에는 Stochastic process, Mixed distribution에 따라 환자군 별 시간에 따라 변화한다고 가정[22]하기 때문에 본 연구에서는 환자군 별로 heterogeneity를 지니는 것에 집중하기 위하여 GEE 방법을 이용하였다.

GEE 모델의 일반적인 데이터의 구성을 먼저 살펴보면 Response Variable는 $Y_{i1}, Y_{i2}, \cdots, Y_{ij}, \cdots, Y_{in_i}$로 이루어져 본 연구에서는 급성악화 발병 여부에 해당되는 1, 0 Binary Variable이고 이 때 Covariate vector는 $X_{i1}, X_{i2}, \cdots, X_{ij}, \cdots, X_{in_i}$로 여러 Parameter 요인들이 된다. 이때 $i \in [1, N]$인 Cluster 내의 Index로 Subject에 해당되는 환자별 ID가 되며 $j \in [1, n_i]$는 Cluster내 측정을 위한 Index로 환자 별 명세서에 해당된다.

위 정의와 같은 데이터를 이용하여 heterogeneity 가정에 의해 Population-Average를 식 (1)을 이용하여 구할 수 있다.

$$E[Y_{ij}|X_{ij}] = \mu_{ij} \tag{1}$$

또한 본 연구에서 예측하고자 하는 Target은 급성악화 발생 여부로 Binary 변수이기 때문에 GEE 모델을 활용하기위해서는 Link Function이 필요하고 이는 가장 일반적으로 쓰이는 식 (2)의 Logit Link function을 이용하였다.

$$Logit[u_{ij}] = \log\left(\frac{P[Y_{ij}=1|X_{ij}}{1-P[Y_{ij}=1|X_{ij}}\right) \tag{2}$$
$$= X_{ij}^T\beta = \beta_0 + \beta_1 X_{ij,1} + \cdots + \beta_p X_{ij,p}$$

본 연구에서 Liang and Zeger[22] Gee Notation 식 (3) 이용하여 Beta를 추정하였다.

$$U(\beta) = \sum_{i=1}^{N} D_i^T V_i^{-1}(Y_i - \mu_i(\beta)) = 0 \tag{3}$$

이 때 D_i는 식 (4)를 이용하여 구할 수 있다.

$$D_i = \frac{\partial \mu_i}{\partial \beta} \tag{4}$$

식 (3)에서 V_i는 식 (5)를 통해 구할 수 있고 이 때 $var(Y_{ij}|X_i) = V_i$이고 $S_i(\mu_i) = diag(V_{ij})$로 구할 수 있으며 $R_i(\alpha)$는 Correlation Matrix이다.

Correlation Matrix의 경우 상황에 맞춰 Independent, Exchangeable, Unstructured 등 다양한 옵션을 선택할 수 있지만 본 분석에서는 Independent Matrix를 이용하였다.

$$V_i = S_i(\mu_i)^{\frac{1}{2}} R_i(\alpha) S_i(\mu_i)^{\frac{1}{2}} \tag{5}$$

3.2.2 Random Forest

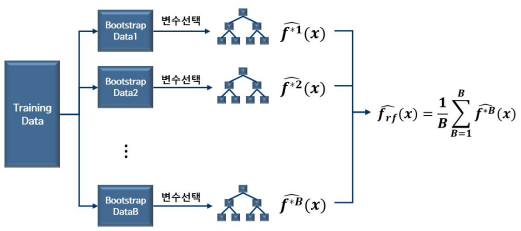
최근 배깅, 랜덤포레스트, 부스팅, SVM 등 다

양한 머신러닝 등이 개발되면서 일반적으로 모수적 모델링 기법보다 예측력이 높다고 알려져 있으며 본 연구에서는 대체적으로 가장 높은 예측력을 지녔다고 알려진 랜덤포레스트 기법을 활용하였다.

랜덤포레스트는 의사결정나무 기반의 앙상블 기법 중 하나로 Bootstrap을 통한 모델링 단계와 앙상블을 통한 결과 통합단계로 나뉜다. Bootstrap을 통한 모델링 단계에서는 우선적으로 훈련용 데이터를 Bootstrap을 통해 데이터를 만들어 낸 후 모형을 구축한다.

모형 구축 단계 전에는 배깅과 랜덤포레스트는 동일한 단계를 거치는데 배깅방법과 달리 랜덤포레스트에서 각각 의사결정나무 모델을 만들 때 변수를 선택하게 된다. 배깅에 비해 흔히 랜덤포레스트가 예측력이 높다고 알려져 있는데 이는 변수 선택 단계를 거치면서 배깅에 비해 다양한 의사결정나무 모델을 만들어 Over fitting 문제를 해결하기 때문이다.

랜덤포레스트 방법에서 대체적으로 총 변수의 3분의 1이나 변수의 제곱근 개수를 선택하며 일련의 과정을 모두 거친 후 마지막에 통합과정을 통해 예측 결과를 살필 수 있다.



〈그림 4〉 Random Forest Working

3.2.3 DNN(Deep Neural Networks)

심층 신경망 구조는 인공신경망(ANN, Artificial Neural Networks)에 기반하여 설계된 개념으로 인공신경망의 입력 층(input layer)과 출력 층(output layer) 사이에 여러 은닉 층(hidden layer)들로 이루어 져 있어 일반적인 인공신경망에 비해 좀 더 복잡한 모형 설계가 가능하다. 심층 신경망은 표준 오류역전파 알고리즘으로 학습될 수 있어 가중치(Weight)들을 식 (6)과 같이 이용한 확률적 경사 하강법을 통하여 갱신한다.

특히 본 연구에서는 R에서 제공하는 H2O 패키지를 사용 하였고 이는 텐서플로우 처럼 RNN과 CNN를 병합하여 학습하는 방법 중 하나이다.

$$\Delta\omega_{jk}(t+1) = \Delta\omega_{jk}(t) + \eta\frac{\partial L(W|j)}{\partial\omega_{jk}} \qquad (6)$$

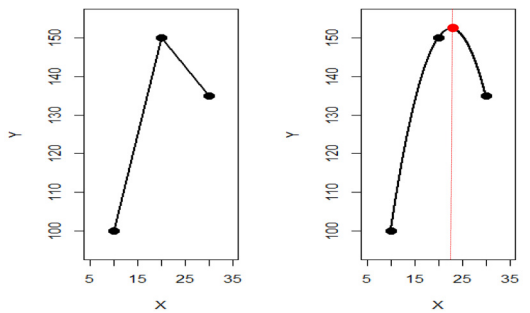
이런 심층 신경망에서도 문제점이 있는데 기존의 인공신경망과 마찬가지로 과적합(Over-fitting)문제를 발생하고 높은 시간 복잡도를 가지게 된다. 그 중 과적합 문제를 해결하기 위해 Hinton의 Drop-out과 L1-정규화, L2-정규화[10]를 이용하여 해결한다.

추가적으로 심층 신경망에서 사용할 수 있는 매개변수는 매우 다양하고 이런 다양성의 문제로 인해 최적의 매개변수를 찾을 때 시간이 매우 오래 걸리는 문제를 지니고 있다. 본 연구에서 사용한 매개변수는 은닉 층의 수, 뉴런의 개수, Epoch, 활성함수, 학습률, 초기 가중치 분포, Drop-out, L1-정규화를 사용 하였다. 이런 다양한 매개변수의 최적 옵션을 찾기 위한 과도한 시간소요를 해결하기 위해서 본 연구에서는 실험계획법의 Response surface 방법을 활용하여 매개변수를 찾으면 Random search나 Grid Search, Golden Section 등의 방법보다 시간이 절약된다[3]는 연구결과를 토대로 마찬가지로 Response surface 방법을 활용하였다.

Response surface 방법은 다양한 요인들의 교호작용 효과가 반응에 미치는 영향을 분석하는 방법[12]으로 실험계획에서 최소한의 실험을 통해 최적의 결과를 얻는 방법과 매우 일치하는 실험계획으로 Box와 Wilson이 고안해 낸 방법이다.

<그림 5>[3]와 같이 일반적인 실험계획을 통해 찾은 최적의 조건과 Response surface 방법을

이용해 찾은 최적의 조건은 다를 수 있다. 특히 모든 조건에서 최적 조건을 찾아야 하는 기존의 방식보다 적은 조합으로 최적의 조건을 찾을 수 있기 때문에 Response surface 방법은 본 연구에서의 시간적인 문제를 해결할 수 있었다.



〈그림 5〉 일반적 최적조건(좌)과 반응표면을 이용해 찾은 최적 조건(우)

〈표 2〉 Response Surface 결과

Variable	Under-sampling			Under-sampling + SMOTE 200%		
Selection	No	0.1	0.05	No	0.1	0.05
Node	120	120	120	120	120	40
Epoch	30	30	30	30	30	90
L1	0	0	0	0.0002	0.0002	0.0002
Drop-out	0.2	0.2	0.2	0.2	0.2	0.2
Rate	0.001	0.001	0.001	0.009	0.009	-
Layer	4	4	4	3	3	3
Weight	1	1	1	2	2	1
Activation	1	1	1	2	2	1
Variable	Under-sampling + SMOTE 300%			Under-sampling + SMOTE 400%		
Selection	No	0.1	0.05	No	0.1	0.05
Node	40	120	-	120	120	-
Epoch	90	90	90	76.06	63.33	63.33
L1	0	0.0001	0.0002	0.0001	0	0
Drop-out	0.2	0.2	0.2	0.1010	0.0929	0.0929
Rate	0.009	-	-	0.009	0.0008	-
Layer	3	3	3	2	2	2
Weight	1	1	1	2	2	2
Activation	3	3	3	1	1	1

※ Weight: 1-Normal, 2-UniformAdaptive.
※ Activation: 1-Tanh, 2-Maxout, 3-Rectifier.

Ⅳ. 분석 결과

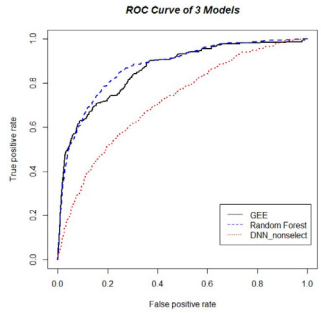
4.1 Response Surface 결과

본 연구에서는 DNN을 이용하여 예측 모델을 생성했으며 이 때 Response surface방법으로 최적화 옵션을 찾아내 분석 시간을 줄일 수 있었다.

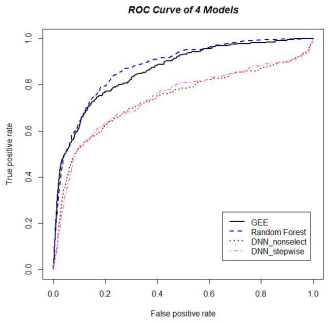
Response surface 분석 결과는 <표 2>와 같이 나타났다. 분석 은 Minitab의 반응표본 설계와 분석을 이용하였고 분석 시 모든 조합을 고려하여 Selection을 할 때와 하지 않을 때를 구분하여 각 데이터에서 최적의 옵션조합을 찾아 낼 수 있었다.

4.2 모델 비교

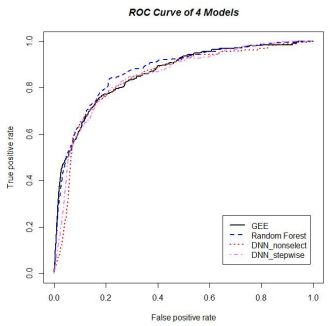
모델 비교는 각 데이터마다 GEE, Random forest, DNN 결과를 나타냈으며 DNN의 경우 Response surface 방법시 selection을 했을 때와 하지 않았을 때로 구분하여 더 세부적인 모델링을 실시했다.



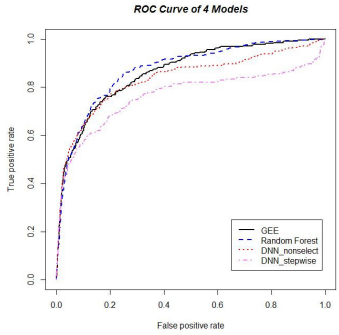
〈그림 6〉 ROC Curve with Under-Sampling Data



〈그림 7〉 ROC Curve with Under-Sampling and SMOTE 200% Data



〈그림 8〉 ROC Curve with Under-Sampling and SMOTE 300% Data



〈그림 9〉 ROC Curve with Under-Sampling and SMOTE 400% Data

분석 결과는 ROC Curve 그림과 AUC, Sensitivity, PPV(Positive Predicted Value)를 활용하였고 ROC Curve 그림은 <그림 6>부터 <그림 9>까지 여러 가지 데이터에서 여러 모델에 대한 결과를 나타냈다.

그림을 살펴보면 GEE와 랜덤포레스트 결과는 네 가지 데이터 모두 비슷한 양상을 띠고 있는 것을 확인할 수 있는 반면에 DNN의 경우 단순한 Under sampling만을 실시했을 때보다 SMOTE 방법을 활용하여 1의 비율이 높아질수록 모델링 결과가 향상하는 것으로 나타났다.

ROC Curve 외의 정확한 모델 비교를 하기위해 AUC, Sensitivity, PPV 등을 나타냈고 이 결과는 <표 3>에서 확인하였다.

<표 3>을 보면 GEE의 경우는 단순히 Under sampling만을 실시했을 때에 비해 SMOTE 방법으로 1의 비율이 증가할수록 AUC가 향상되는 것을 알 수 있었고 랜덤포레스트의 경우와 DNN의 경우 향상되다 다시 떨어지는 양상을 보여 랜덤포레스트의 경우 Under sampling과 SMOTE 방법으로 1의 비율을 200%로 증가 시켰을 때, DNN의 경우에는 300% 증가 시켰을 때 AUC가 가장 높게 나타나는 것을 확인할 수 있었다. AUC 관점으로 봤을 때는 랜덤포레스트가 SMOTE 200%일 때 AUC 0.8755로 가장 높은 예측력을 나타냈다.

〈표 3〉 모델 결과 비교

Variable	Method	AUC	민감도	PPV
Under sampling	GEE	0.8585	0.7097	0.0166
	RF	0.8723	0.7839	0.0158
	DNN noselection	0.7192	0.5742	0.0085
Under sampling + SMOTE 200%	GEE	0.8604	0.7484	0.0163
	RF	0.8755	0.8355	0.0136
	DNN noselection	0.7370	0.5516	0.0171
	DNN selection	0.7402	0.6065	0.0137
Under sampling + SMOTE 300%	GEE	0.8602	0.7581	0.0160
	RF	0.8717	0.8387	0.0144
	DNN noselection	0.8436	0.7419	0.0167
	DNN selection	0.8493	0.8000	0.0134
Under sampling + SMOTE 400%	GEE	0.8617	0.7613	0.0152
	RF	0.8719	0.8613	0.0127
	DNN noselection	0.8347	0.7710	0.0142
	DNN selection	0.7636	0.6032	0.0189

AUC를 통해 전반적인 모델의 예측력을 살펴보는 것 외에도 본 연구에서의 목표는 COPD 급성악화를 찾아내는 것이고 급성악화 질병은 매우 위험한 질병이기 때문에 민감도에 해당하는 Sensitivity를 통해 급성악화에 걸린 사람 중 얼마나 많은 인원을 찾아낼 수 있는지 또한 중요하다. 민감도 기준으로 살펴봤을 때 GEE는 AUC와 마

찬가지로 SMOTE 비율이 높아질수록 좋으며 랜덤포레스트와 DNN은 AUC와 달리 SMOTE 비율이 높아질수록 민감도가 지속적으로 좋아지는 것을 확인할 수 있었고 랜덤포레스트가 SMOTE 400%일 때 0.8613으로 가장 좋아 전체적으로 랜덤포레스트의 예측력이 가장 좋게 나타나는 것을 확인할 수 있었다.

본 연구에서 사용한 데이터가 Imbalanced Data이면서 Longitudinal Data이였는데 Imbalance 문제를 SMOTE 방법으로 조금은 해결을 하긴 했지만 Longitudinal Data의 문제를 해결하지 못해 DNN의 결과가 예상한 것 보다 좋지 않게 나온 것으로 생각 된다.

4.3 GEE 모델링 결과

본 연구의 주 목적은 Imbalance이고 Longitudinal인 Data에서의 모델 비교도 목적에 있지만 추가적으로 이때까지 COPD 급성악화 연구는 COPD 급성악화와 관련 있는 요인이 어떤 것인지 알아보는 단일적인 연구로만 이루어져 있어 여러 요인들을 복합적으로 봤을 때 어떤 요인이 가장 영향력이 크며 어떤 요인은 예측 관점에서 필요 없는지 확인하는 목적을 동시에 지니고 있다.

이를 위해 모델링 결과를 해석할 수 있는 GEE 모델을 통해 이를 확인하고자 한다. GEE 모델의 경우 사실상 AUC와 민감도의 차이가 적기 때문에 변수의 개수가 가장 적게 나타난 Under sampling 결과를 통해 확인하였다.

분석 결과는 <표 4>와 같고 여러 요인 중 Stepwise selection 결과 9가지 정도의 요인이 선택 되었다. 이 때 선택된 요인들에 대해서는 다른 Sampling 옵션에서 선택된 요인들과 비교 했을 때 대체적으로 성별, CAT 점수, FEV1%, 과거 1년간 급성악화 발생 횟수, 과거 1년간 사용한 약제(SYSBRONCH, SABA), 과거 1년간 천식 유병 여부, 바이러스 질병정보(IFV, hCoV)가 선택 되었다.

이때 성별의 경우 여자가 남성에 비해 급성악화에 걸릴 확률이 1.42배, CAT 점수의 경우 점수가 높을수록(삶의 질이 안 좋을수록)급성악화 발생의 위험률이 1.0163 증가하고 FEV1%의 경우 폐 기능이 좋을수록 위험률이 0.9944 감소하였다. 또한 과거 1년간 급성악화 발생 횟수는 횟수가 증가할수록 위험률이 1.1931 증가하여 기존에 생각했던 것 보다 위험률 증가가 낮았고 과거 1년간 사용한 약제에 관련해서 SYSBRONCH를 사용한 경우 위험률이 1.2006배 증가하고 SABA를 사용할 경우 1.5856배가 된다. 천식의 경우 1년여 동안 유병이 있던 경우 위험률은 4.2238배로 가장 높은 위험률 증가를 보였고 공공데이터 중에서는 유일하게 바이러스 검출률만 선택되어 IFV 바이러스 검출률의 경우 해당 일로부터 3주 전, 4주 전 검출률의 합이 증가함에 따라 위험률이 1.0039 증가하고 hCoV바이러스 검출률은 3주 전 검출률 증가에 따라 위험률은 1.0290 증가하였다.

〈표 4〉 Under-Sampling에서 GEE 결과

Parameter	오즈비	S.E	P-값
성별	1.4200	0.1961	0.0004
CAT 점수	1.0163	0.0071	<.0001
FEV1%	0.9944	0.0027	0.0008
과거 1년간 급성악화 발생 횟수	1.1931	0.0333	<.0001
과거 1년간 SYSBRONCH 사용 여부	1.2006	0.1209	0.0014
과거 1년간 SABA 사용 여부	1.5856	0.1963	<.0001
과거 1년간 천식 유병 여부	4.2238	0.6658	<.0001
3주, 4주 전 IFV 바이러스 검출률 합	1.0039	0.0009	<.0001
3주 전 hCoV 바이러스 검출률	1.0290	0.0153	0.0071

전체적인 Parameter들의 결과는 기존 연구들에서 나온 결과와 일치하는 결과를 얻었으나 환경적 요인의 경우 기온이나 미세먼지 등의 요인이 뽑히지 않았다. 이 결과는 모델에 따라 다소 차이

는 있지만 비교해 보면 주로 성별, 나이, CAT 점수, FEV1%, 약제, 질병, 바이러스 정보가 뽑혀 기존 연구에서 생각했던 흡연력, 미세먼지, 최저기온 등의 효과가 여러 데이터를 살펴봤을 때 나타나지 않은 것을 확인할 수 있었다. 흡연력의 경우에는 Base-Line으로 활용하여 한 시점의 흡연력만 나타나기 때문에 흡연량의 변화와 급성악화 간의 관계를 추가적으로 살펴볼 필요는 있지만 미세먼지와 최저기온의 경우 다른 요소들에 의해 충분히 설명이 되고 이전에 나타났던 연관성은 다른 요소에 의해 나타난 연관성일 수 있기 때문에 좀 더 정확한 분석을 위해서는 Causality를 확인할 필요가 있음을 알 수 있었다.

Ⅴ. 결 론

본 연구는 Imbalanced and Longitudinal Data의 Parametric Model과 머신러닝의 비교를 하고 머신러닝 방법에서 Imbalanced 문제를 해결 했을 때 예측력의 변화를 살펴보는데 중점을 두었다.

본 연구에서 Imbalanced Data 문제를 해결하기 위해 Under sampling 기법과 SMOTE 방법을 연계하여 사용하였고 그 결과 SMOTE 방법을 통해 Minority의 비율을 높일수록 예측력이 올라가는 것을 확인할 수 있었다.

특히 본 연구 시작 전 생각과 같이 Parametric 기법 보다는 Over fitting 문제가 큰 머신러닝에서 결과치의 향상이 크게 나타나는 것을 확인할 수 있어서 단순히 머신러닝을 통해 예측을 하는 것이 아니라 통계적 기법을 활용하여 데이터의 문제를 해결 했을 경우 더욱 향상된 결과를 얻을 수 있음을 확인할 수 있는 좋은 기회였다.

하지만 생각과 달리 DNN의 결과는 항상 GEE나 랜덤포레스트의 결과에 미치지 못해 Longitudinal Data에서의 학습에 대한 연구가 필요함을 느낄 수 있었다. 현재 Longitudinal Data에서의 머신런닝 기법에 대한 연구는 지속적으로 이루어지고 있으며 Mixed Effect LSSVM[14]과 Adaptive Bayes를 이용한 User's Modeling[9] 등이 있지만 사실상 사용에 대한 제약이 많아 이용하기가 어렵기도 하고 인공신경망 기반의 딥러닝에서는 아직 연구되지 않고 있어 만약 이런 Longitudinal Data에서의 머신러닝기법이 연구된다면 통계적 모델을 통해 개개인의 질병 예측 시대가 도래 할 것이라 전망한다.

참 고 문 헌

[1] 유광하, 정기석, 김영삼, 박용범, 신경철, 윤형규, 이상엽, 이진국, 이진화, "전국적 COPD Cohort 연구 기초 자료(KOCOSS 연구 cohort)", 대한결핵 및 호흡기학회 추계학술발표 초록집, pp.196-196, 2012.

[2] 유지홍, "COPD 진료지침", 대한결핵 및 호흡기학회, 2014.

[3] 이범석, "반응 표면 방법을 이용한 딥러닝 매개변수 최적화 연구", 인하대학교학위논문, 2017.

[4] Andersson, F., S. Borg, S.-A. Jansson, A.-C. Jonnson, A. Erincsson, C. Prutz, E. Ronmark, and B. Lundback, "The costs of exacerbations in chronic obstructive pulmonary disease (COPD)", *Respiratory Medicine*, Vol.96, No.9, pp.700-708, 2002.

[5] Au, D.H., C.L. Bryson, J.W. Chine, H. Sun, E.M. Udris, L.E. Evans, and K.A. Bradley, "The Effects of Smoking Cessation on the Risk of Chronic Obstructive Pulmonary Disease Exacerbations", *J Gen Intern Med*, Vol.24, pp.457-463, 2009.

[6] Burge, S. and J.A. Wedzicha, "COPD exacerbations: definitions and classifications", *Eur Respir J*, Vol.21, No.41, pp.46s-53s, 2003.

[7] Chawla, N.V., K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-

sampling Technique", *Journal of Artificial Intelligence Research*, Vol.16, pp.321-357, 2002.

[8] Donaldson, G.C., T.A.R. Seemungal, A. Bhowmik, and J.A. Wedzicha, "Relationship between exacerbations frequency and lung function decline in chronic obstructive pulmonary disease", *Thorax*, Vol.57, pp.847-852, 2002.

[9] Gama, J. and G. Castillo, "Adaptive Bayes for User Modeling", EUNITE, 2002.

[10] Hinton, G., L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition", *IEEE Signal Processing Magazine*, Vol.29, No.6, pp.82-97, 2012.

[11] Hurst, J.R., "Susceptibility to Exacerbations in Chronic Obstructive Pulmonary Disease", *The New England Journal of Medicine*, Vol.363, No.12, 2010.

[12] Khuri, A.I. and S. Mukhopadhyay, "Response surface methodology", *TOC*, Vol.2, No.2, pp.128-149, 2010.

[13] Laird, N.M. and J.H. Ware, "Random-Effects Models for Longitudinal Data", *Biometrics*, Vol.38, pp.963-974, 1982.

[14] Luts, J., G. Molenberghs, G. Verbeke, S. Van Huffel, and J.A.K. Suykens, "A mixed effects least suqres support vector machine models for classificatio of longitudinal data", *Computational Statistics and Data Analysis*, Vol.56, pp.611-628, 2012.

[15] Nathalie, J. and S. Shaju, "The class imbalance problem: A systematic study", *Intelligent Data Analysis*, Vol.6, pp.429-449, 2002.

[16] Seemungal, T., R. Happer-Owen, and A. Bhowmik, "Respiratory viruses, Symptoms, and Inflammatory Markers in Acute Exacerbations and Stable Chronic Obstructive Pulmonary Disease", *Am J Respir Crit Care Med*, Vol.164, No.9, pp.429-449, 2001.

[17] Terence, A.R. and A. Jadwiga, "Exacerbation frequency and FEV1 decline of COPD: is it geographic?", *European Respiratory Journal*, Vo.l43, pp.1220-1222, 2014.

[18] Teresa, T., "Progression from Asthma to Chronic Obstructive Pulmonary Disease Is Air Pollution a Risk Factor?", *AM J Respir Crit Care Med*, Vol.194, No.4, pp.429-438, 2016.

[19] Tseng, C.M., Y.T. Chen, S.M. Ou, Y.H. Hsiao, S.Y. Li, S.J. Wang, A.C. Yang, T. Chen, and D. Perg, "The Effect of Cold Temperature on Increased Exacerbation of Chronic Obstructive Pulmonary Disease: A Nationwide Study", *PLOS ONE*, Vol.8, No.3, pp.e57066, 2013.

[20] Tu, Y.H., Y. Zhang and G. Fei, "Utility of the CAT in therapy assessment of COPD exacerbations in China", *BMC Pulmonary Medicine*, pp.14-42, 2014.

[21] Yoon, H.K., Y.B. Park, C.K. Rhee, J.H. Lee, and Y.M. Oh, "Summary of the Chronic Obstructive Pulmonary Disease Clinical Practive Guideline Revised in 2014", *The Korean Academy of Tuberculosis and Respiratory Diseases*, 2017.

[22] Zeger, S.L., K.Y. Lian, and P.S. Albert, "Models for Longitudinal Data: A Generalized Estimating Equation Approach", *Biometircs*, Vol.44, pp.1049-1060, 1988.

[23] http://health.chosun.com/site/data/html_dir/2016/09/27/2016092702474.html.

[24] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/.

[25] http://www.cdc.go.kr/CDC/main.jsp.

[26] http://www.kma.go.kr/index.jsp.

[27] https://www.airkorea.or.kr/index.

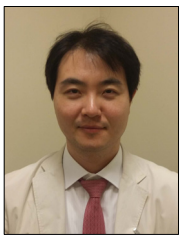
저 자 소 개

정 현 명(Hyeon-Myeong Jeong)

- 2016년 : 인하대학교 통계학과 (학사)
- 2016년~현재 : 인하대학교 통계학과 석사과정 재학 중
- 관심분야 : Big Data Analytics, Data Mining, Statistical Computing

박 헌 진(Heon-Jin Park)

- 1984년 : 서울대학교 계산통계학과 (학사)
- 1987년 : Lowa State University, Statistics (석사)
- 1987년 : Lowa State University, Statistics (박사)
- 1994년~현재 : 인하대학교 통계학과 교수
- 관심분야 : Big Data Analytics, Data Mining, Statistical Computing



이 진 국(Chin-Kook Rhee)

- 2002년 : 가톨릭대학교 의과대학 (학사)
- 2007년 : 가톨릭대학교 의과대학 (석사)
- 2013년 : 가톨릭대학교 의과대학 (박사)
- 현재 : 서울성모병원 호흡기내과 부교수
- 관심분야 : Chronic Obstructive Pulmonary Disease, Asthma

이 종 민(Jong-min Lee)

- 2009년 : 가톨릭대학교 의과대학 (학사)
- 현재 : 서울성모병원 호흡기내과 임상강사
- 관심분야 : Chronic Obstructive Pulmonary Disease, Asthma