빅데이터 분석에 의한 요율산정 방법 비교

: 실손의료보험 적용 사례

2018. 9

이항석

KiRi 보험연구원
Korea Insurance Research Institute

머 리 말

4차 산업혁명이 시작되고 인터넷을 기반으로 모든 것들이 연결(connected)되면서 데이터가 다양한 형태와 여러 장소 및 엄청난 규모로 축적되고 있다. 보험업계도 급속도로 변화하는 환경에 대응하여 보험의 빅데이터 관리와 다양한 기법을 적용하여 보험소비자의 만족도 제고를 위한 신상품의 개발, 적정한 보험료의 산출, IFRS17과 신지급여력제도의 대응방안 마련에 활용할 필요가 있다. 2018년 베를린에서 개최된 ICA (International Congress of Actuaries)를 통해서 발표된 논문의 많은 부분은 빅데이터 분석이다. 즉, 해외보험사의 보험전문가들은 빅데이터 분석을 통하여 새로운 가치창출의 수단으로 다양한 주제를 연구하고 활용하고 있다. 이에 본 보고서에서는 이러한 대응의 한 시도로서 실손의료보험의 요율산정에 빅데이터기법을 적용한다.

실손의료보험은 대다수 국민이 가입하여 국민의 의료 보장성 강화에 기여하고 있지만 여러가지 문제점들로 인하여 여러 차례의 제도개선에도 불구하고 체계적인 요율변수의 선택과 보험료 세분화 및 제도개선에 대한 논란은 계속되고 있다. 본 보고서에서는 빅데이터 분석을 통하여 요율산정에 중요한 요율변수가 무엇인지 살펴보고 기존의 전통적인 방법론과의 차이점을 비교한다.

IFRS17과 신지급여력제도에서 시가평가를 목적으로 다양한 보험계약들에 적합한 세분화된 모델링을 요구한다. 이 보고서는 우리나라 보험업계와 학계에 빅데이터 분석을 사용하여 가치를 창출하는 하나의 예로서 향후 후속 연구를 통하여 보험수요의 예측, 위험률 산출, 해약률 관리, 도덕적 해이, 보험사기 등 다양한 주제로 빅데이터 분석이 이루어지길 기대한다.

마지막으로 이 보고서의 내용은 연구자 개인의 의견이며, 우리원의 공식적인 의견이 아님을 밝혀둔다.

2018년 9월

보 험 연 구 원

원장 **한 기 정**

목차

목차

표 차례

그림 차례

A Ratemaking of Private Health Insurance using Data Mining Techniques

Many insurance companies use data mining techniques to find insights hidden in their data. In this study, a ratemaking of the private health insurance is carried out through various supervised learning. In the case of private health insurance, although it is necessary to calculate a more detailed rate to prevent adverse selection, various ratemaking methods have not yet been applied in practice.

Currently, rating variables of private health insurance are genders, ages and class rates. In spite of the heterogeneous risk characteristics of private health insurance, the use of only restrictive rating variables can lead to sustained loss ratios and a reduction in the private health insurance market by intensifying adverse selection. Therefore, it is necessary to consider introducing the policyholder's performance as a rate variable, which can better explain the risk characteristics of each policyholder.

In order to overcome the shortcomings of one-way classification, ratemaking approach using multivariate method such as generalized linear model (GLM) is used. Furthermore, we apply machine learning techniques such as decision trees, ensemble models, MARS and neural network models to ratemaking in this study. We implement through R programming so that insurance practitioners and researchers can try machine learning algorithms.

요 약

Ⅰ. 서론

■ 빅데이터 시대에 발맞춰 외국보험회사들이 빅데이터기법을 도입하고 있음

○ 기존의 전통적인 요율산정기법에서 교호작용을 고려하지 않아 요율산정에 왜곡을 가져올 수 있으나, Multivariate Method가 단점을 보완함

- 요율산정의 Multivariate Method로 일반화선형모형이 최근에 널리 이용되고 있으며 본 연구에서는 일반화선형모형, 일반화 혼합선형모형, 의사결정나무, MARS 그리고 신경망모형을 통하여 실손의료보험 데이터를 분석함

■ 실손의료보험은 국민 60% 이상이 가입한 상품임에도 불구하고 손해율은 여전히 100%를 상회하고 있음

○ 국민복지에 지속적인 순기능을 위해 현행되고 있는 단일보험료체계가 아닌 세분화된 요율산정의 필요성이 대두됨

■ 최근에는 실손의료보험에 대한 다양한 요율 산출 연구가 진행되고 있음

○ 역선택과 관련된 선행연구로

- 김대환·이봉주(2013)는 국내 실손의료보험시장에 역선택이 존재한다는 점을 확인하였고,
- 이경아·이항석(2016)은 역선택의 원인을 계약자별 편차가 큰 위험특성으로 주장하며 이에 대한 해결책으로 과거 보험금 정보인 경험자료를 활용한 보험료 차등화를 제안함

○ 과거실적을 반영한 요율 산출 방법에 대한 선행연구로 이항석·이수빈·백혜연(2017)은 신뢰도기법을 반영하여 보험료를 산정함

Ⅱ. GLM 빈도 심도 분석

■ 일반화선형모형은 보험자료의 흔한 특성인 비정규성의 특성에 맞는 방법론임

○ GLM은 선형모형의 확장된 모델로 정규성과 등분산을 만족하지 않아도 되어 포괄적인 적용이 가능함

- 선형모형의 가정과는 다르게 반응변수는 정규분포를 따르지 않아도 되고 오차항의 등분산성을 상정하지 않아도 됨

○ 반응변수의 분포를 지수족(Exponential family)으로 가정함

- 지수족에는 포아송분포, 감마분포, 이항분포 등이 있으며 이는 보험데이터에 적합한 분포임

○ 로그 연결함수를 이용하면 승법성이(Multiplicativity) 성립하여 상대도(Relativity)에 적용하기에 용이함

- $\ln(\mu) = \beta_0 + x_1\beta_1 \Rightarrow \mu = e^{\beta_0 + x_1\beta_1}$

■ GLM을 활용한 빈도 분석

○ 보험사고건수는 포아송분포를 가정함

- 사고건수는 이산형변수이므로 이를 반응변수로 갖는 포아송분포가 적합함

■ GLM을 활용한 심도 분석

○ 보험사고금액은 0이상인 데이터를 반응변수로 하는 감마분포를 가정함

○ 위험에 노출된 정도가 다를 때, 즉 익스포져(Exposure)가 다를 때 오프셋(offsets)을 사용하여 보정함

- 보험금은 가입금액에 따라 최대로 받을 수 있는 보험금이 달라지므로 오프셋항을 활용함

Ⅲ. GLMM을 이용한 할인할증제도 적용 방법

■ 일반화 혼합선형모형(GLMM)을 이용해 할인할증제도를 적용함

○ GLMM 방법은 GLM에서 고정효과(Fixed Effect)로 가정하고 계약자들의 실적변수들의 계수는 임의효과(Random Effect)로 가정하여 계약자에게 발생한 보험사고 간의 상관관계를 모형화한 것임

○ 할인할증제도는 최적의 방법론이라기보다는 주로 해외 선진국에서 사회적으로 수용 가능한 형태로 도입되어 옴

- 이미 2017년 4월부터 판매되고 있는 국내 실손의료보험 상품에 대해서는 2년 무사고자에게 보험료를 10% 할인해주는 할인할증제도를 적용하고 있음
- AIA, AXA, BUPA와 같은 해외보험사들은 무청구 이력을 바탕으로 할인할증제도를 운영함

■ 무사고 누적연수를 바탕으로 할인할증제도 실증분석

○ -1/Top scale 형태의 할인할증제도로 실증분석결과 무사고 누적연수가 0년인 계약자보다 4년 이상인 계약자가 61% 할인을 받음

Ⅳ. 의사결정나무와 MARS

■ CART 방법론은 의사결정나무 중에 가장 널리 쓰이는 방법론으로 Breiman et al. (1984)이 개발함

○ 출력변수가 연속형인지 범주형인지에 따라서 회귀나무와 분류나무로 나뉠 수 있는데 본 연구에서는 회귀나무로 분석함

○ 이진분리(Binary split)로 해석성은 좋으나 예측력이 떨어지는 단점이 있음

- 또한 가장 설명력이 있는 변수에 대하여 최초로 분리가 일어나는 특징을 가지므로 요율산정에 있어서 주요변수가 무엇인지 파악 가능함

- complexity parameter(cp)를 통하여 나무크기 조절이 가능한데 cp가 클수록 나무형태가 작아짐

■ CART 분석결과 직전연도 발생건수가 최초 분리변수로 사용되었으며, 나머지 변수들도 중요도 순으로 분리변수로 선택됨

○ 분리변수의 중요도 순은 GLM 분석의 p-value 작은 순과 일치함

- GLM 분석에서 p-value 값은 작을수록 유의한 변수라고 해석될 수 있음

■ MARS는 입력변수가 많은 고차원 회귀문제에 적합한 알고리즘으로 Friedman (1991)이 제안함

○ 기저함수(basis function)로 데이터 자체(X_j)가 아닌 변형된 형태($(X_j - t)_+$)로 입력됨

- 전통적 선형회귀와 같이 오차제곱합을 최소화시키는 β_m 계수들을 추정함
- 기저함수는 전진 선택법을 사용하여 선택되어, 먼저 $B_0(X) = 1$을 모형에 투입하고 오차제곱합을 최소화하는 변수와 매듭점을 찾고 기저함수쌍을 모형에 추가
- 그 후 과대적합(Overfitting)을 막기 위해 후진 소거법으로 설명력이 없는 기저들을 제거함

■ GLM 옵션이 있는 MARS 분석

○ GLM 옵션에서 링크함수를 로그로 지정하면 빈도나 심도가 반응변수가 음수 값을 갖는 것을 방지함

V. 앙상블기법과 신경망모형

■ 신경망모형은 복잡한 구조를 가진 데이터의 예측 문제를 해결하는 비선형 모형화 방법이며 본문에서는 가장 간단한 신경망모형을 통해서 신경망모형의 구조를 살펴봄

○ 입력변수의 선형결합에 비선형 함수를 취하는 사영추적회귀(Projection Pursuit Regression)임

- 입력층에서 은닉층으로 시그모이드 함수를 사용한 선형결합이 이루어지고, 은닉층에서 출력층으로 비선형결합이 이루어짐

■ 은닉층이 다층인 신경망모형을 다차원 신경망모형, 즉 딥러닝이라고 칭함

○ 가중치가 많아 해석하는데 어려움이 있음

■ 신경망모형을 이용한 보험료 차등화

○ 신경망모형을 이용한 빈도와 심도 모형은 MAE(Mean Sbsolute Error) 지표를 기준으로 보았을 때 대체로 우수한 편이나 입력변수에 영향을 많이 받음

- 빈도 모형의 경우 신경망모형이 가장 우수하였으나 그 해석이 어려움
- 빈도는 0 근처의 값이 대부분이므로 입력변수를 데이터와 똑같이 삽입하는 반면 심도는 입력변수가 크기 때문에 정규화하여 모델링함

Ⅵ. 시사점 및 결론

■ 모델 비교

○ MAE(Mean Absolute Error)지표를 바탕으로 5-묶음 교차검증을 각 방법론별로 실시함

- 그 결과, 빈도 모형에서는 신경망모형이 가장 우수하고 GLM과 같은 통계적인 기법보다 머신러닝기법 사용 시 오차가 감소함
- 심도 모형은 로그를 연결함수로 사용한 모형들이 오차가 작게 평가됨

■ 상황별 제한점

○ GLM 모형과 같은 통계 모형은 추정한 계수를 해석하기 편한 장점이 있는 반면

빈도 모형에서는 낮은 예측력을 보임

○ GLMM은 Random Effect를 고려할 수 있는 장점이 있으며 할인할증제도 모델링에 응용 가능함

○ CART는 이진분리로 해석하기 용이하나 예측력이 다른 모델에 비해 좋지 않음. 이를 보완하기 위해 앙상블기법이 쓰임

○ 앙상블기법에는 배깅, 부스팅, 랜덤 포레스트 등이 있는데 부스팅같은 경우 해석력이 부족하고 이상치(Outlier)에 민감함

- 배깅이나 랜덤 포레스트의 경우 복원추출 시 일부 관측치들은 훈련자료에서 빠질 수도 있음

○ 신경망모형 및 딥러닝은 빈도 모형에서 예측력은 좋으나 은닉층에 노드가 많을 때에는 해석하기 쉽지 않은 편임

Ⅰ. 서론

1. 연구배경 및 목적

4차 산업혁명이 시작되고 사람, 물건 등 모든 것들이 인터넷을 통하여 연결되고 있다. 4차 산업혁명의 주요 핵심인 빅데이터 분석, 인공지능(Artificial Intelligence), 로봇, 사물인터넷(Internet of Things), 블록체인(Blockchain) 등도 결국에는 연결(connected)이라는 개념을 기반으로 발전되어 왔다. 이러한 사회를 바탕으로 사람들이 어떻게 행동하고 비즈니스가 어떻게 작동하고 있는지에 대한 정보들이 쌓여 데이터 자본을 형성하게 된다.

빅데이터 시대에 많은 기업과 공공기관이 빅데이터 도입을 검토하고 있다. 빅데이터를 도입한다는 것은 데이터를 수집한 뒤 이를 분석해 데이터 속에 숨은 패턴(Patterns)을 찾아내서 문제 해결에 활용하는 것이다. 이미 해외 보험회사들은 효율적인 데이터 통합을 위해 데이터 센터를 구축하고 이를 바탕으로 최신 ICT 기술을 사업에 활용하는 방법에 대해 연구 중이다. 기존에 활용하지 않았던 종류의 데이터를 활용함으로써 업무효율을 향상시키고, 상품 및 서비스를 차별화하는 전략을 실행하여 기업의 경쟁우위를 확보하려는 것이 해외 보험사들이 빅데이터에 관심을 갖는 목적이다.

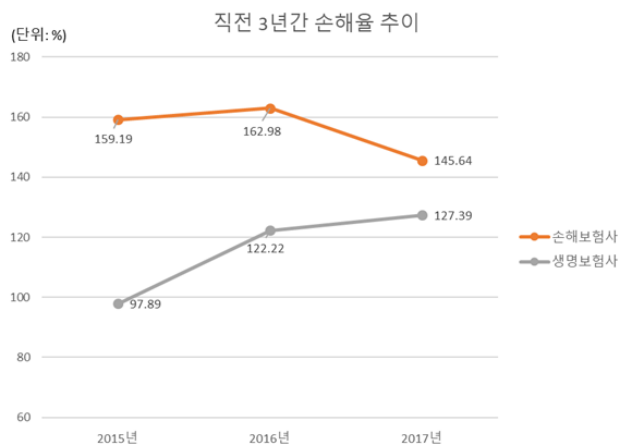
반면에 국내 보험회사는 빅데이터 분석을 위한 전사적 데이터 통합이 이루어진 상태는 아니며, 외부데이터의 활용 또한 미흡한 상태이다. 더하여 여태까지의 실손의료보험의 요율산정은 전통적인 통계기법에 의해서만 시도가 이루어져 왔다. 실손의료보험같은 경우 역선택 방지를 위해 더 세분화된 요율산정이 필요함에도 불구하고 실무적으로 다양한 요율산정 방법들이 아직 많이 적용되지는 않고 있다.

본 연구에서는 통계적 기법 외에도 여러 가지 빅데이터 분석을 통하여 실손의료보

험의 요율산정을 진행한다. 빅데이터 분석을 하는 것은 실손의료보험의 데이터로부터 부가적인 정보를 얻고 이를 바탕으로 보험료 산정을 개선하기 위함이다. 빅데이터 분석에 사용하는 데이터의 개수와 설명변수가 제한적인 측면이 있지만 여러 가지 방법론을 적용하는 것에 초점을 맞춘다.

실손의료보험은 국민건강보험공단이 보장하지 않는 환자의 본인부담 의료비를 포괄적으로 보장해주는 상품으로 국민의 약 66%가 가입한 상품이다. 하지만 실손의료보험의 손해율은 여전히 100%를 상회하고 있다.

〈그림 Ⅰ-1〉 직전 3년간 실손의료보험 손해율 추이



자료: 손해보험협회 공시, 손해율=발생손해액/위험보험료

실손의료보험은 정보비대칭에 따른 역선택이 존재하나, 현재는 요율변수로서 나이, 성별 그리고 상해급수에만 의존하고 있다. 따라서 다양한 보험 모델링에 대한 변화가 필요하다.

본 연구에서는 보험실무자들과 연구자들이 빅데이터 분석을 시도할 수 있도록 R프로그래밍을 통하여 구현할 것이다. R프로그래밍은 접근성이 좋고 누구나 쉽게 따라할 수 있는 장점이 있다. R과 관련한 설치방법 및 코드는 이 보고서의 부록과 참고문헌을 참고하길 바란다.

빅데이터 분석기법은 Univariate Method에서 컴퓨터와 통계의 발전에 따라 Multivariate Method로 발전해왔다. 순보험료법(Pure Premium Analysis)이나 손해율법(Loss Ratio Analysis)은 전통적인 요율산정기법으로 사용되었으나 이는 교호작용을 고려하지 않아 요율산정에 왜곡을 가져올 수 있다. 그러한 단점을 보완하기 위하여 최근에는 일반화선형모형(GLM)과 같은 Multivariate Method를 통한 요율산정 방법이 사용되고 있다. 따라서 이러한 추세에 이어 본 연구에서는 의사결정나무, MARS 그리고 신경망모형과 같은 빅데이터기법들을 요율산정에 적용시키고 분석하고자 한다.

2. 최근 빅데이터기법 연구 동향

가. 보험에서 빅데이터 분석의 필요성

IFRS17과 SolvencyⅡ에서 가치평가를 목적으로 다양한 보험계약들에 적합한 세분화된 모델링을 요구함에 따라 평가모델과 시스템 측면에서 더욱 복잡해 질 것으로 예상되고 있다. Aleandri(2018)에 의하면 SolvencyⅡ 시행에 따라 기존의 계리적 기법이 더욱 더 복잡한 모델로 조금씩 대체되고 있다고 한다. 따라서 이러한 환경변화에 따라 점점 머신러닝(Machine Learning)기법의 필요성이 대두되고 있는 실정이다.

최근에는 보험산업에서도 다양한 분야에 머신러닝기법들이 응용 및 연구되고 있다. 머신러닝기법은 빠르면서도 큰 데이터를 다룰 수 있고, 비용도 많이 들지 않는다는 장점이 있다. 그리고 무엇보다 전통적인 기법보다 변수들 간의 관계를 고려하여 분석한다는 장점이 있다. 이러한 환경으로 미루어 보았을 때 계리사는 경영과정에서 빅데이터를 사용하는데 필요한 기본적인 기술이나 도구에 보다 능숙해질 필요가 있다(Gupta et al. 2018).

보험 관련 데이터 분석은 대부분 출력변수를 예측하는 것을 목표로 하는 지도학습(Supervised Learning)을 한다. 지도학습은 회귀나 분류의 결과 값을 주는 것으로 해당하는 방법론에는 CART, MARS, 신경망모형 등이 있다. 예를 들면 손해보험의 지급준

비금 추정, 그리고 생명보험에서의 변액연금에서 해약환급금 기댓값 추정 등 다양한 보험 분야에서 머신러닝기법을 적용 및 연구하고 있다. 대다수의 연구들에서는 기존의 계리적 기법보다 빅데이터기법들이 더 정확한 값을 예측한다고 기술하고 있다.

나. 머신러닝 알고리즘의 이해

Gupta et al.(2018)는 머신러닝 알고리즘의 블랙박스를 해석할 수 있어야 한다고 강조하고 있다. 블랙박스를 해석하는 것은 우리가 처한 상황에 알맞은 모수에 적절한 방법을 적용하여 분석을 하고 있는지 이해하게 해준다. 그리고 입력변수에 따른 출력변수만 얻는 것이 아니라, 그에 대한 관련 식이 무엇인지 살펴볼 수 있으며, 또한 블랙박스를 해석 가능하게 하여 출력변수값들이 어떻게 예측되었는지 알 수 있다. 본 연구에서는 이에 관한 머신러닝의 기본 개념에 대해 조금 더 자세히 설명하고 특히 의사결정나무와 변수선택에 주안점을 두고 설명을 해보고자 한다.

러닝(Learning)이라고 하는 것은 함수로서 데이터의 패턴을 찾아내는 작업인데 자극에 대한 반응이 어떻게 변하는지(change in response probability to stimulus)를 말한다. 러닝은 크게 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 나눠진다. 쉽게 말하면 지도학습은 종속변수를 생각할 때를 말하고 비지도학습은 종속변수의 개념이 없이 설명변수들끼리의 관계를 고려하여 분류하는 것을 말한다.

사용가능한 데이터가 증가함에 따라 과거로부터 학습하여 미래에 적용하는 방법론에 대한 연구가 지속적으로 진행될 필요가 있다. 머신러닝은 이 목적에 부합한 방법론이라고 볼 수 있다. IT 산업에서는 모델링을 목적으로 머신러닝 알고리즘을 적용하는 것에 주력하고 있으며 보험업계에서도 유사한 목적을 이유로 조금씩 관심을 두기 시작하였다. 이미 보험업계에서 많은 가격산정 방법론(Pricing Methodologies)들은 전통적인 GLM 모형보다 부스팅이나 서포트벡터기계와 같은 정교한 방법론들로 이동하고 있는 추세이다.

다. 생명보험 연구

1) 계약자 행동(해지 위험 모델링)

SolvencyⅡ 시행에 따라 보험회사들의 경제적 가치를 추정하기 위해 쓰던 전통적인 계리기법들은 점점 더 복잡하고 구조적인 모델들로 대체되고 있다. Aleandri(2018)에서는 계약자 행동의 해지위험을 머신러닝을 바탕으로 모델링하고 있다. 현재 실무에서 쓰이는 것보다 더 좋은 추정을 할뿐만 아니라, 로지스틱 회귀와 같은 전통적인 모수 모형들과 비교하였을 때 예측력이 향상된 것을 관찰할 수 있다. 예측모형으로 로지스틱 회귀분석, 배깅, 랜덤 포레스트 그리고 부스팅을 비교한다. AUC 지표로 살펴보았을 때 배깅이 가장 좋은 모형으로 나온다.

2) 변액연금 해약환급금 가치평가

Ha(2018)는 머신러닝기법 중 특히 신경망모형을 이용하여 변액연금에서의 해약환급금을 평가하였다. GMWB(Guaranteed Minimum Withdrawal Benefit)의 상품구조는 해약환급금에 대한 옵션가치가 non-European이기 때문에 매우 복잡하다. 이에 대한 기댓값을 근사하는 것을 해결하기 위해 신경망모형을 적용하고 있다.

라. 손해보험 연구

1) 지급준비금 추정

Wuthrich(2017)은 손해보험사 지급준비금을 계산할 때 신경망모형을 활용하였다. 전통적인 지급준비금 추정은 지급보험금을 합쳐놓은 삼각형에서 이루어진다. 전통적인 지급준비금 추정은 homogeneous하다는 가정을 바탕으로 이용한다. 여기서는 Mack's chain ladder 방법을 바탕으로 개별 클레임들의 정보들이 어떻게 이질적인지 신경망모형을 통해 살펴본다.

Mukherjee & Vijayaraghavan(2018)에서는 전통적인 지급준비금 추정 테크닉과 더불어 머신러닝기법이 필요하다고 주장하고 있다. 세분화된 지급준비금 추정이 필요하기 때문에 방법론으로 군집분석(Clustering)을 사용한다. 예를 들면 자동차 책임보험과 같은 경우 청구가 자동차의 전방 충돌인지 후미 충돌인지에 따라서 심도와 진전 패턴이 달라진다. Mukherjee & Vijayaraghavan(2018)은 다음과 같은 이유에서 머신러닝 기법들이 필요하다고 기술하고 있다. 청구 유형이나 익스포져 특성에 따라 분리가 필요하다. 청구를 세분화하는 것은 체인레더와 청구 한 건당 평균비용을 개선시키고 익스포져의 세분화는 B-F방식이나 F-S방식을 개선시킨다. 세분화된 지급준비금 추정은 위험에 대해 통찰력을 갖게 해주고 이는 가격산정이나 자본모델링을 향상시킨다.

2) 자동차 보험상품 구입여부 모델링

Francis et al.(2018)은 자율학습(Unsupervised Learning) 중 하나인 주성분분석(Principal Component Analysis)과 지도학습 중 MARS와 단순 의사결정나무모형을 통하여 데이터를 분석한다. 분석에 쓰인 데이터는 여행용 자동차 보험상품 구입여부를 반응변수로 모델링하였다. 설명변수로는 우편번호와 고객의 보험상품 정보를 사용하였다. 모든 예측방법에서는 고객의 개인적인 데이터를 포함할수록 더 좋은 결과를 얻을 수 있다. Francis et al.(2018)은 벤치마크로 GLM을 통하여 분석을 하고 빅데이터기법들과 비교하였다. 이 연구에서는 이항분포와 로그를 연결함수로 지정하여 모델링하였다. 다음의 분류나무로 분석을 하였을 때는 의사결정나무모형에 분리가 전혀 일어나지 않았으나 회귀나무로 바꾸었을 때는 분리가 일어났다. 의사결정나무는 랜덤 포레스트기법을 통하여 보완하였다. 하지만 그럼에도 불구하고 GLM이 의사결정나무나 랜덤 포레스트보다 더 좋은 모델로 판단되었고 이러한 결과에 대해서는 다른 데이터셋에 적용해 볼 필요가 있음을 언급하였다. PCA 분석은 통계적인 방법으로 주로 데이터셋에서 서로 상관이 있는 변수들을 원래 데이터보다 차원이 낮은 데이터로 변환할 수 있다는 것이다. PCA 분석에서는 training data의 두 가지 변수를 제거하였다. AUC(Area Under Curve) 지표를 종합하여 보았을 때 모든 분석 중 PCA 분석이 가장 우수한 모델이었다.

3) 자동차보험 빈도 모델링

Marechal(2018)은 가격산정을 위해서 자동차보험 데이터를 바탕으로 빈도 모델링을 하여 전통적인 모형인 GLM과 GAM을 머신러닝방법과 비교하였다. 손해보험 가격산정에 많이 쓰이는 기술은 분류와 회귀로 지도학습 머신러닝 알고리즘이 적합하다. 통계적인 모형과 머신러닝기법을 비교해보면 배깅모델의 이탈도가 가장 작기는 했으나 두드러지는 차이는 없었다. 다만 머신러닝기법은 그 과정이 자동화되어있는 만큼 사업의 로직과 맞는지 주의를 기할 필요가 있고 또한 과대적합의 위험에 대해서도 간과하여서는 안된다는 것을 언급하였다.

4) 건강보험 청구 관련

Mukherjee & Ajmani(2018)은 머신러닝기법을 바탕으로 건강보험 청구 관련 분석을 하였다. micro cluster 방법론을 바탕으로 진단, 입원, 수술자료를 활용하여 건강보험의 패턴을 분석한다. 그리고 CART 기법을 바탕으로 보험사기 식별을 한다.

3. 실손의료보험 관련 선행연구

가. 실손의료보험의 역선택

역선택 유인이 높은 실손의료보험의 상품적 특성은 손해율 상승의 주요 원인 중 하나로 볼 수 있다. 특히 의료보장 분야에서 역선택 유인이 높은 이유는 공급자와 수요자 사이의 정보비대칭성으로 인한 문제가 가장 크고 가입자별 위험특성 편차가 크기 때문이다. 역선택은 위험집단이 이질적일 경우 더욱 악화되는데 이러한 상황에서 단일보험료를 유지할 경우 구조적 보험료 인상과 보험시장 축소를 야기할 수 있다. 역선택을 완화하기 위해서는 정보비대칭성을 완화하고 위험평가를 보완하는 것이 필요하다. 하지만 현재 감독규정은 여전히 나이와 성별 등 제한적 요율변수만을 허용하고 있다. 위

험편차가 큰 위험집단에 대해 계속해서 현재와 같이 제한적 정보만을 반영한 단일보험료를 부과한다면 손해율 상승과 이로 인한 계약자의 이동으로 제도의 지속성은 위협받게 될 것이다. 이러한 보험시장에서의 역선택에 관한 실증분석연구는 지속해서 이어져 왔다. 김대환·이봉주(2013)의 연구에서는 국내 실손의료보험시장에 역선택이 존재한다는 점을 확인하였고, 이경아·이항석(2016)은 실손의료보험 비급여 의료비 부분의 보장특성과 계약자별로 위험특성 편차가 크다는 점이 역선택을 발생시킬 가능성 높다고 주장하며 이에 대한 해결책으로 가입자의 과거 보험금 정보인 경험자료를 활용한 보험료 차등화를 제안하였다.

보험료 산정에 가입자의 경험자료는 비관찰 위험특성에 대한 정보를 반영한다는 점에서 가입자 위험특성을 파악하기에 좋은 정보이다. 또한 불필요한 가입자 행동은 경험자료를 반영하여 보험료를 조정함으로써 그러한 행동을 완화시킬 수 있다. 위험수준별 보험료 차등화는 보험료 산출의 기본원리이며 이미 사회적 합의가 이루어진 명제이다. 이에 이경아·이항석(2016)은 계약자별 과거 보험금 지급정보를 활용한 베이지안 기댓값 원리(Bayesian theorem)를 바탕으로 위험특성에 대한 정보의 제약성을 보완하고자 하였다. 베이지안 기댓값 원리는 과거 보험금 지급정보가 계약자별 비관찰 위험특성을 가장 잘 반영할 수 있으며 계약자별 위험정보를 토대로 참값에 근사한 보험료를 산출할 수 있게 해준다. 이를 토대로 산출된 신뢰도보험료는 가입자 정보가 제한적인 상황에서 공정한 보험료에 가장 근사한 값을 산출할 수 있다는 이론적 타당성을 지니며 손해보험 전반에서 널리 사용되고 있다.

나. 실손의료보험에서 신뢰도기법을 반영한 보험료 산정[1)]

신뢰도 이론은 신뢰할 만한 값을 구하기 위하여 특정 위험집단의 경험데이터를 이용하여 얻어진 추정치 등을 절충하여 최종적인 추정치를 구하는 계리적 기법이다. 이 이론은 특정 위험집단의 경험데이터가 충분하지 못하여 신뢰할 만한 추정치를 구하는 것이 어려울 때 추정의 신뢰성 및 안정성을 높이기 위해 사용된다. 또한 신뢰도 이론

1) 이항석·이수빈·백혜연(2017), pp. 41~73을 발췌 및 요약함

의 개념은 1914년 설립된 CAS(미국 손해보험협회)가 초기 보험료 산정을 위해 현재까지 수집된 데이터를 활용하는 방법으로 처음 제기되기 시작하여 그 후로도 보험회사의 보험료 산출 방법으로 적용되고 있다(김명준·최정아·김영화 2013).

신뢰도 또는 신뢰도 계수 값의 범위는 0~1이며, 신뢰도가 반영된 최종적인 추정치를 구하는 기본 공식은 다음과 같다.

$$\text{New Rate} = Z \times \text{Observed Data} + (1-Z) \times \text{Old Rate} \qquad (\text{I}-1)$$

여기서 Z를 신뢰도(Credibility) 혹은 신뢰도 계수(Credibility Factor)라 부르고, 개인의 과거 경험데이터에 대한 신뢰 정도를 측정한 값이다(Norberg 1992). 또한 식 (Ⅰ-1)에서 'Observed Data'는 최초 관측 자료로 추정대상 집단의 경험데이터를 이용한 추정치이고, 'Old Rate'는 이 집단을 포함한 전체 집단의 자료를 이용하여 산출한 추정치이다. 그리고 앞서 언급한 것과 같이 Z는 'Observed Data'에 할당된 신뢰도이며, (1-Z)는 'Old Rate'에 할당된 신뢰도로 여신뢰도(complement of credibility) 또는 신뢰도의 보완이라고 한다. 최종적인 추정치는 보통 경험통계를 활용하여 추정하는 손해율, 보험료, 보험금, 경험 사망률 등이 될 수 있다.

결국 적정한 추정치를 산출하기 위해서는 어떻게 'Observed Data'와 'Old Rate' 추정치 정보를 가중 평균하여 합리적으로 결합시키는지에 달려있다. 신뢰도 범위는 0과 1사이이며, 신뢰도(Z)가 '1'이라는 것은 추정대상 집단의 경험데이터만으로 추정치가 결정된다고 볼 수 있고, 반대로 신뢰도(Z)가 '0'이라는 것은 새로운 요율인 'New Rate'가 추정대상 집단을 포함한 전체집단에 대한 자료를 이용하여 결정된다고 볼 수 있다.

신뢰도 이론의 기본 원칙은 현재의 데이터를 기초로 한 평균값과 과거의 경험적 추정에 의존한 평균값 사이에 가중치를 반영하여 확률변수의 미래 기댓값을 추정하는 것이다. Jones & Gerber(1975)는 과거 경험 자료인 지급 보험금과 평균보험료들의 가중평균으로 특정 시점의 보험료를 산출하는 'Updating type'에 대한 신뢰도 공식 (Credibility Formula)을 제안하였고, 더 자세한 신뢰도 이론에 대한 설명은 Buhlmann(1967), Buhlmann(1969), Buhlmann & Straub(1970), 그리고 Buhlmann & Gisler(2005)를 참고

하면 된다. 이항석·이수빈·백혜연(2017)에서는 과거 9년치의 사고발생 정보를 사용하였고, Semiparametric credibility 신뢰도 이론을 적용시켜 보험료의 차등화 필요성에 대해서 제안하여 최종적으로 각 위험특성별 보험료를 산출하였다.

4. 연구범위와 방법

Ⅰ장에서는 연구의 배경 및 목적에 대해서 서술하고 있다. Ⅱ장과 Ⅲ장은 각각 GLM과 GLMM을 활용한 요율산정을 보여준다. Ⅳ장에서는 CART와 MARS에 대한 방법론과 예제를 싣는다. Ⅴ장에서는 앙상블기법과 신경망모형에 대해 설명한다. 마지막으로 Ⅵ장에서 각 방법론별 비교 및 상황별 제한점을 제시하며 결론을 맺는다.

Ⅱ. GLM 빈도 심도 분석

1. 일반화선형모형(GLM) 방법론

일반화선형모형(GLM: Generalized Linear Model)은 포괄적인 의미의 선형모형(LM: Linear Model)이며, 반응변수와 설명변수 간의 관계를 정량화하는 모형이다. 정규선형모형과 가장 큰 차이점은 반응변수의 분포가 지수족에서 선택되어 정규성을 만족하지 않아도 된다는 것이다. 따라서 일반화선형모형은 보험금(반응변수 Y)처럼 설명변수 X_i와 선형관계가 성립하지 않는 비정규성(non-normal) 자료에 적합하다.

일반화선형모형은 선형모형의 정규성을 배제하였다는 것 이외에도 선형관계에 대한 가정에서 차이가 있다. 선형모형은 식 (Ⅱ-1)에서처럼 설명변수 x_i와 반응변수 y간 선형관계를 가정한다. 이때 x'_i는 $x_1, x_2, ..., x_p$인 설명변수의 행벡터이고 β는 설명변수의 계수인 β_j의 열벡터이다. 이 식은 x_j 이외 모든 x_k를 상수로 고정했을 때, x_j가 한 단위 증가함에 따라 y가 β_j만큼 증가한다는 것을 의미한다.

$$y_i = x_i^{'}\beta + \epsilon_i \,. \qquad (Ⅱ\text{-}1)$$

반면 일반화선형모형은 식 (Ⅱ-2)에서처럼 설명변수와 반응변수 y_i의 기댓값인 함수 $g(\mu_i)$가 선형관계를 갖는다고 가정한다. 이때 연결함수(Link function) $g(\mu_i)$는 반응변수 y_i의 기댓값 μ_i의 미분가능 단조함수이며 반응변수의 분포를 지수족(Exponential family)으로 가정한다. 확률분포의 확률밀도함수가 식 (Ⅱ-3)의 형태로 표현될 때 이를 지수족이라고 말한다. 이 확률밀도함수는 상수인 정준모수(canonical parameter) θ와 산포모수(dispersion parameter) ϕ, 함수인 $a(\theta)$와 $c(y,\phi)$에 의해서

최종적인 형태가 결정된다.

$$g(\mu_i) = x_i^{'}\beta\,. \tag{Ⅱ-2}$$

$$f(y) = c(y,\phi)\,\exp\left\{\frac{y\theta - a(\theta)}{\phi}\right\}. \tag{Ⅱ-3}$$

지수족 확률분포는 평균이 μ, 분산 $Var(y) = \phi V(\mu)$의 형태이고 $V(\mu)$는 μ의 함수형태이다. 따라서 지수족 확률분포를 따르는 반응변수의 분산은 평균이 증가할 경우 같이 증가하게 되는데 이는 일반적인 보험자료의 특성과 유사하다. 특히 보험금 지급금액 및 사고건수에 자주 쓰이는 지수족 분산함수는 〈표 Ⅱ-1〉과 같다.

〈표 Ⅱ-1〉 정준연결함수와 분산

분포	연결함수	연결함수 이름	분산
Binomial	$\log(\mu/(1-\mu))$	로짓함수	$\mu(1-\mu)$
Gamma	$-1/\mu$	멱수함수	μ^2
Gaussian	μ	항등함수	1
Inverse-Gaussian	$-2/\mu^2$	멱수함수	μ^3
Poisson	$\log\mu$	로그함수	μ
Negative Binomial	$\log\mu$	로그함수	$\mu(1+\alpha\mu)$

자료: Venables, Rispley(2002)

계수 β와 산포모수 ϕ은 반응변수의 분포를 선택하고 해당 분포의 $a(\theta)$에 따라 최대우도추정법(MLE: Maximum Likelihood Estimation)에 따라 관찰된 표본 $y_1, \ldots, y_n$에서 우도가 최대가 되게 하는 값으로 추정한다. 만약 반응변수의 분포를 포아송으로 선택하면 식 (Ⅱ-3)의 확률밀도함수는 $\phi = 1$, $\theta = \ln(\mu)$, $a(\theta) = e^{\theta}$이 되며, 식 (Ⅱ-4)의 로그우도를 최대화함으로써 구해진다.

$$l(\beta,\phi) = \sum_{i=1}^{n} \ln f(y_i;\beta,\phi) = \sum_{i=1}^{n}\left\{\ln c(y_i,\phi) + \frac{y_i\theta_i - a(\theta_i)}{\phi}\right\}. \tag{Ⅱ-4}$$

항등함수를 가진 정규분포와 같은 경우를 제외하고는 뉴튼-랍슨반복이나 피셔스코어링 등 수치적 계산방법을 통해 최대화를 위한 식 (Ⅱ-4)의 1차조건(first order condition)을 구한다.

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p . \quad (Ⅱ\text{-}5)$$

$$\mu = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}. \quad (Ⅱ\text{-}6)$$

식 (Ⅱ-5)에서 양변에 지수를 취하면 곱으로 표현된 식 (Ⅱ-6)을 얻을 수 있다. 이와 같이 일반화선형모형에서 연결함수로 로그를 사용하는 경우 승법성(Multiplicativity)이 성립한다. 특히 보험데이터 분석에서 요율변수 간 상대도(Relativity)를 적용하기에 이러한 특징은 매우 용이하다는 장점이 있다. 예를 들면 성별이 남성(성별(남성)=기준수준, $p = 1, x_1 = 0$)일 때 e^{β_0}에 대하여 성별이 여자일 때 ($x_1 = 1$)가 $e^{\beta_0 + \beta_1}$이면 e^{β_1}만큼이 기준수준 남성에 대한 여성의 상대도가 된다.

위험집단에서 각 데이터들의 위험에 노출된 정도가 다를 때 보정하는 것으로 오프셋(offsets)을 사용한다. 이를 사용함으로써 다른 변수들이 위험노출정도에 좌우되지 않는다. 다음 절의 심도 분석에서는 가입금액에 따라 최대 사고금액이 달라지므로 이를 오프셋항으로 적용한다. 변수 n은 익스포져이고, 식 (Ⅱ-7)에서 $\ln(n)$을 오프셋이라고 칭한다.

$$\ln\left(\frac{\mu}{n}\right) = x'\beta \Rightarrow \ln(\mu) = \ln(n) + x'\beta. \quad (Ⅱ\text{-}7)$$

$$\mu = n e^{x'\beta}. \quad (Ⅱ\text{-}8)$$

2. 보험료 차등화-GLM을 이용한 분석

GLM을 이용하여 실손의료보험 실제 데이터에 대한 빈도와 심도를 분석하고자 한다. 빈도와 심도는 데이터의 모양이 다르기 때문에 각각 포아송분포와 감마분포로 가정한다. 빈도의 종속변수는 사고건수로 예를 들면 0건, 1건과 같은 이산형 변수인 반면에 심도의 종속변수는 사고금액으로 0원 초과의 연속형 변수이다. 빈도와 심도 분석 결과를 각각 살펴보면 다음과 같다.

가. 빈도 분석 결과

GLM을 이용하여 실제 보험사의 실손의료보험 데이터를 이용한 빈도 분석 결과가 〈표 Ⅱ-2〉이다. 이 표와 같이 사고건수에 대해서는 포아송분포를 사용하였으며 로그 연결함수를 사용하였다. 종속변수가 사고건수인 이산형 변수이기 때문에 보험금지급 청구건수의 분포로 주로 많이 쓰이는 포아송분포를 사용한 것이다. 아래의 표는 식 (Ⅱ-6)의 β값들을 추정한 결과를 포함하고 있다.

〈표 Ⅱ-2〉 사고건수 GLM 포아송분포 가정

반응변수	사고건수				
반응분포	Poisson				
연결함수	Log				
△이탈도[1]	1084				
AIC[2]	5675.8				
모수	$\hat{\beta}$	s. e.	z-value	Pr(〉|z|)	significance[3]
절편	-2.5589	0.2156	-11.869	〈 2e-16	***
성별: 여	0.4087	0.0671	6.085	1.16E-09	***
연령	0.0122	0.0044	2.762	0.0057	**
상해급수 2급	-0.0714	0.0765	-0.934	0.3503	-
상해급수 3급	-0.2424	0.1207	-2.007	0.0447	*
15년도 발생건수	0.5533	0.0151	36.578	〈 2e-16	***

주: 1) 모수추가로 인한 이탈도 감소량(Null deviance-Residual deviance)
2) AIC(Akaike's Information Criterion)=-2ln(L)+2p
3) ***: 0.001, **: 0.01, *: 0.05

표를 해석하는 방법을 설명하면 다음과 같다. 만약 가입자가 남성, 40세, 상해급수 1급, 직전연도 발생건수가 0이면 평균사고건수는 $e^{-2.5589+0.0122*40}$으로 나타낼 수 있다. 반면 같은 조건에 여성일 때에는 $e^{-2.5589+0.4087\times1+0.0122*40}$으로 표현한다. 이 두 가지 경우를 비교해보면 여성일 때의 사고건수가 남자에 비해 $e^{0.4087}=1.5049$배 더 높을 것으로 예상해 볼 수 있다. 이러한 값을 남성이 기준수준(Base level)일 때의 여성의 상대도라고 말하며, 여성의 사고건수가 남성의 사고건수보다 1.5049배 더 많을 가능성이 있다는 것을 모형을 통해 표현할 수 있다. 아래의 표는 각 설명변수별로 기준수준 대비 상대도를 계산한 결과이다. 이러한 결과를 이용하면 특정 조건을 만족하는(설명변수에 0 또는 1 대입한 결과) 경우의 빈도 발생 정도를 기준수준에 비교하여 예상해 볼 수 있다.

〈표 Ⅱ-3〉 사고건수 GLM 상대도

설명변수		상대도
성별	남성	1.0000
	여성	1.0549
연령	+1	1.0123
상해급수	1	1.0000
	2	0.9311
	3	0.7848
15년도 발생건수	+1	1.7390

나. 심도 분석 결과

본 분석에서는 심도인 사고금액을 분석하기 위해 지급금액이 0원 초과인 데이터를 바탕으로 연속형 변수인 보험금 지급금액에 주로 많이 적용되는 감마분포를 가정하여 분석하였다. 〈표 Ⅱ-4〉는 GLM을 이용한 사고금액에 대한 추정결과이고, 앞서 빈도에서와 같이 로그 연결함수를 사용하였다. 〈표 Ⅱ-3〉과 〈표 Ⅱ-5〉를 비교해보면, 빈도는 남성에 비해 여성에게 사고건수가 많이 발생할 확률이 1.0549배 더 높게 나타났으나, 심도는 남성에 비해 여성의 사고금액이 더 클 확률이 1.0997배로 성별로 구분했을 때 빈도보다 심도에서의 차이가 약간 더 크게 나타났다.

〈표 Ⅱ-4〉 사고금액 GLM 감마분포 가정

반응변수	사고금액				
반응분포	Gamma				
연결함수	Log				
오프셋	가입금액				
△이탈도[1]	42.56				
AIC[2]	20681				
모수	$\hat{\beta}$	s. e.	t-value	Pr(>|z|)	significance[3]
절편	-0.9816	0.3226	-3.0433	0.0024	**
성별: 여	0.0951	0.0977	0.9726	0.3311	-
연령	0.0187	0.0066	2.8336	0.0047	**
상해급수 2급	-0.0988	0.1130	-0.8741	0.3823	-
상해급수 3급	-0.0276	0.1699	-0.1624	0.87108	-
15년도 발생건수	0.1105	0.0327	3.3838	0.0008	***

주: 1) 모수추가로 인한 이탈도 감소량(Null deviance-Residual deviance)
2) AIC(Akaike's Information Criterion)=-2ln(L)+2p
3) ***: 0.001, **: 0.01,*: 0.05

〈표 Ⅱ-5〉 사고금액 GLM 상대도

설명변수		상대도
성별	남성	1.0000
	여성	1.0997
연령	+1	1.0189
상해급수	1	1.0000
	2	0.9059
	3	0.9728
15년도 발생건수	+1	1.1168

이와 같이 GLM 분석을 통해 상대도를 산출하게 되면 해당 계약자의 위험특성에 따른 빈도 또는 심도에 대한 발생가능성을 예측할 수 있게 되는 것이다. 위험특성별로 보험료를 차등화하고자 한다면 이러한 추정치를 이용하여 표준적으로 적용하는 기준보험료에 위험특성별 상대도를 곱하는 방식으로 보험료율을 각기 다른 위험특성별로 차등하여 적용할 수 있는 것이다.

Ⅲ. GLMM을 이용한 할인할증제도 적용 방법

1. 할인할증제도

국민건강보험의 보장성 강화 정책에 따라 실손의료보험에 대한 역할과 관심 또한 증가하고 있다. 실손의료보험에 대한 신뢰성을 제고하기 위해서는 무엇보다 보험료 산출 구조의 신뢰성을 높이는 것이 가장 중요하다. 현재 감독규정상 성, 연령, 위험급수 등과 같이 극히 제한된 요율변수들을 이용하여 계약자들의 보험료를 단일보험료로 부과하다 보니 현 요율 체계에서는 계약자 간 형평성이 훼손되는 부분이 일정 부분 발생하고 있다. 이러한 요율 체계의 신뢰성과 형평성 제고를 위하여 최근 실손의료보험의 다양한 요율 산출 방법에 대한 연구가 진행되고 있다. 이번 절에서는 특히 일반화 혼합선형모형(GLMM: Generalized Linear Mixed Model)을 이용한 할인할증제도 적용 방법에 대해서 소개하고, 실제 보험사 자료를 이용하여 실증분석을 진행하고자 한다.

계약자에 대한 사전적으로 관찰할 수 있는 정보 이외에도 계약자의 실적 변수와 같이 계약자의 비관찰 위험을 실손의료보험의 보험료 산출 시 반영하는 방법에 대해 본 절에서는 고려하고자 한다. 사전적으로 관찰 가능한 변수들의 계수들은 앞서 소개한 GLM에서 고정 효과(Fixed Effect)로 가정하고, 계약자의 실적 변수들의 계수는 임의효과(Random Effect)로 가정하여 계약자에게 발생한 보험사고 간의 상관관계를 모형화한 GLMM 방법을 활용하고자 한다. GLMM 분석 방법에 대해서는 다음 절에서 조금 더 상세히 소개할 예정이며, 우선 할인할증제도에 대해서 소개하고자 한다.

2017년 4월부터 판매되고 있는 실손의료보험 상품에 대해서는 이미 2년 무사고자에 대해 보험료를 10% 이상 할인해주는 할인할증제도를 적용하고 있다. 할인할증제도라는 것은 어느 방법이 완벽하게 최적의 방법론이라 정의하는 것이 어렵고, 주로 해외

선진국에서는 사회적으로 수용 가능한 형태로 도입되어 왔다. 물론 자동차보험과 달리 실손의료보험에 할인할증제도를 적용하는 것에 대해 도덕적으로 많은 논쟁이 있는 것도 사실이다. 자동차보험의 경우 할인할증제도로 인해 운전자의 운전 습관을 변화시키거나 사고율을 낮추는 등 사회적, 개인적으로 모두 긍정적인 효과를 불러일으키기도 한다. 그러나 실손의료보험에 자동차보험과 동일하게 무사고 이력만을 고려하여 할인할증제도를 적용하는 것은 윤리적으로 적절하지 않을 수도 있다. 그 이유는 유병자의 경우 지속적으로 진료를 받고, 의료행위를 하는 것이 불가피하기 때문에 이러한 점을 고려하지 못한 채 무조건적으로 의료행위에 따라 보험료를 계속적으로 할증시키는 것이 적절하지 못할 수 있다. 따라서 실손의료보험에 할인할증제도를 적용하기 위해서는 다른 보험상품에 비해 제도 적용에 대한 사회적 수용성과 보험사의 재무건전성을 모두 만족하는 계리적인 분석과 검토가 도입 전 반드시 필요하다.

간단히 해외 보험사의 할인할증제도를 소개해 보면 다음과 같다. AIA의 경우에는 계약자의 건강한 생활과 습관을 유지하는 것을 목적으로 하여 보험금 미청구 이력이 최소 3년 연속일 경우 최소 5%에서 최대 15%까지 할인을 해주는 상품들이 있다. 또한, 건강상태를 일정 수준 이상 유지하기 위해 치과진료를 받거나, 백신주사를 맞거나, 또는 온라인으로 금연 선언을 하는 등 다양한 건강을 위한 행위에 따라 포인트를 적립하도록 하여 누적된 포인트에 따라 갱신 보험료를 할인해주는 혜택이 포함된 프로그램도 운영하고 있다. AXA의 경우에는 보험금 미청구 기간이 연속 2년 이상일 경우 미청구 기간에 따라 차별적인 할인율을 적용하고 있다. 이러한 할인할증제도를 적용하기 위해서는 사회적으로도 수용 가능하고, 동시에 보험사의 재정적인 면에도 큰 영향을 끼치지 않는 최대, 최소 할인율 폭 설정에 대한 고민이 필요하다. 마지막으로 BUPA의 경우에는 우리나라의 자동차보험과 유사하다고 볼 수 있다. BUPA는 계약자를 여러 등급으로 세분화하여 등급 간 전이과정에 따라 최종적인 등급에 도달할 경우 해당 등급별 할인율을 적용하는 방식을 사용하고 있다. 이러한 방식을 벤치마킹하기 위해서는 최초 가입할 경우의 시작 등급과 최소 그리고 최대 등급 간의 할인율 차이 폭 등에 대해 신중하게 선택할 필요가 있다.

실손의료보험은 다른 보험상품과 달리 계약자의 건강 상태에 따라 보험금 지급 사

유가 발생하기 때문에 계약자의 의지로 빈도나 심도를 조정하는 것이 거의 불가능하다. 따라서 반드시 실손의료보험에 할인할증제도를 적용하기 위해서는 다른 보험상품들에 비해 이런 윤리적인 문제까지도 해결할 수 있도록 더욱 더 신중하게 방법론을 연구하고 선택해야만 한다는 점을 명심해야 한다.

2. GLMM 방법론

일반화 혼합선형모형(GLMM: Generalized Linear Mixed Model)은 앞서 소개한 GLM의 확장된 모형들 중 하나이다. GLM에서는 설명변수의 계수들을 모두 고정 효과(Fixed Effect)로 간주하고 있으나, GLMM에서는 예를 들어, 비관찰 정보 중 하나인 계약자의 사고 실적 변수의 계수에 대하여 임의 효과(Random Effect)를 가정하고 있다. 다시 정리하면, 동일 계약자의 n년 무사고 누적연수를 분석하기 위해서는 한 계약자(한 보험증권)를 n년 동안 매해 관찰해야만 한다. n년 동안 매해 발생하는 계약자의 사고 실적들은 서로 상관관계가 있기 때문에 n년 무사고 누적연수(설명변수)를 군집(보험증권들 또는 계약자들)마다 고정된 효과로 가정하면 안 되며, 같은 확률변수에 의한 결과로 보는 임의 효과(Random Effect)로 가정해야 한다. 따라서 앞서 소개한 GLM의 식 (Ⅱ-5)에서 $X_1 = x_1, ..., X_p = x_p$들은 계약자 특성을 나타내는 계약자의 사전적 관찰 정보인 설명변수로 생각할 수 있고, 이들의 계수는 군집(보험증권들 또는 계약자들)마다 동일한 값으로 고정 효과를 가정하면 되나, GLMM에서는 여기에 추가로 군집에 따라 설명변수들의 계수들을 임의적(확률적)이라 간주하는 것이다. 다음 절에서는 실증분석 시 무사고 누적연수 0년, 1년 등을 각각 설명변수 z로 가정하고, 각 설명변수들에 대한 모수 벡터를 γ라 가정하여 다음 식을 사용할 것이다.

$$\ln(\mu) = x'\beta + z'\gamma,\ \gamma \sim N(0, G). \qquad (Ⅲ-1)$$

위 식에서 설명변수 z의 계수인 γ는 공분산 행렬 G를 갖는다고 가정한다. 본 연구

에서는 이러한 GLMM 식을 이용하여 계약자에 대한 사전적 관찰 정보들과 비관찰 정보들을 동시에 활용하여 할인할증제도하에 각 위험특성별로 차등화된 보험료들을 산출하기 위해 상대도를 산출하였다.

GLM에서 반응변수 Y와 설명변수 $X_1 = x_1, ..., X_p = x_p$는 다음과 같은 관계가 성립한다. $x_1,..., x_p$가 주어졌을 때 반응변수 Y의 평균을 μ라 하고, μ가 지수족 분포(Exponential family)를 따른다고 가정하면 확률밀도함수 $f_Y(y)$는 정준모수(canonical parameter) θ, 산포모수(dispersion parameter) ϕ, 그리고 $\alpha(\theta)$, $c(y, \phi)$ 함수들에 의해 최종적인 형태가 결정된다.

$$f_Y(y) = c(y, \phi)\exp\left\{\frac{y\theta - \alpha(\theta)}{\phi}\right\}. \quad \text{(Ⅲ-2)}$$

$$E[Y|x_1, ..., x_p] = \mu,\ Var[Y|x_1, ..., x_p] = \phi \bullet V(\mu). \quad \text{(Ⅲ-3)}$$

반응변수 Y의 기댓값인 μ에 대한 연결함수($g(\mu)$)가 미분가능한 단조함수라 하고, 다음과 같은 선형관계를 갖는다고 가정한 것이 일반화 선형모형(GLM)이다.

$$g(\mu) = x_i^{'}\beta\,. \quad \text{(Ⅲ-4)}$$

식 (Ⅲ-3)에서 지수족 확률분포의 평균은 μ이고, 분산은 μ에 대한 함수($V(\mu)$)이기 때문에 반응변수의 평균이 증가하면 분산도 함께 증가한다. 이러한 평균과 분산 관계는 보험자료의 특성을 반영하기에 적합하며, 다음의 표에 있는 다양한 지수족 분포와 그에 대한 연결함수(Link function)들을 분석 시 사용하면 된다.

〈표 Ⅲ-1〉 지수족 분포에 대한 연결함수

분포	link 이름	link 함수	평균
Normal	Identity	μ	$\mu = X\beta$
Exponential	Inverse	μ^{-1}	$\mu = (X\beta)^{-1}$
Gamma			
Inverse Gaussian	Inverse squared	μ^{-2}	$\mu = (X\beta)^{-1/2}$
Poisson	Log	$\ln(\mu)$	$\mu = \exp(X\beta)$
Bernoulli	Logit	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(X\beta)}{1+\exp(X\beta)} = \frac{1}{1+\exp(-X\beta)}$
Binomial			
Categorical			
Multinomial			

본 연구에서는 실증분석을 위해 사고건수(반응변수 Y)가 포아송(Poisson)분포를 따른다고 가정하고, GLM에서 Log 연결함수를 사용하였다. 이 Log 연결함수는 승법성(Multiplicativity)때문에 요율변수들($X_1 = x_1, \ldots, X_p = x_p$) 간 상대도를 적용하는 것이 굉장히 용이하다(이항석·이가은·이경아 2017).

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p . \quad (Ⅲ\text{-}5)$$

$$\mu = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} . \quad (Ⅲ\text{-}6)$$

예를 들어, 설명변수가 1개이고, x_1이 성별을 나타내는 범주형 자료라 한다면, 주로 자료의 양이 많은 것을 기준(Base)으로 설정하면 된다. 기준이 남성일 경우는 x_1=0이고, 여성일 경우는 x_1=1로 보아 더미(Dummy)변수로 변환시켜야 한다. 이때 최종적으로 실증분석에서 산출할 상대도라는 것은 기준(Base) 대비 추정하고자 하는 변수 값을 의미하며 다음의 식과 같다.

$$\frac{\mu_{여성}}{\mu_{남성}} = \frac{e^{\beta_0 + \beta_1 \times 1}}{e^{\beta_0 + \beta_1 \times 0}} = e^{\beta_1}. \quad (Ⅲ\text{-}7)$$

식 (Ⅲ-7)에서 기준으로 설정한 남자에 대한 요율 상대도 값은 1이고, 여성에 대한 상대도 값 e^{β_1}는 여성의 사고 발생 가능성이 기준인 남성에 비해 e^{β_1}배 더 높다고 해석하면 된다(기승도·김대환 2009). 더 자세한 GLM과 GLMM에 대한 설명은 Jong & Heller(2008), Goldburd, Khare & Tevet(2016), Faraway(2016)의 연구들을 참고하길 바란다.

3. GLMM을 이용한 할인할증제도의 실증분석

GLM 분석과 달리 GLMM 분석 방법을 활용하면 사전 요율변수들은 고정 효과를 가정하고, 비관측 위험(Θ)은 임의 효과를 가정한다. Y를 실제 사고건수(빈도)라 하면, Y가 요율 산정 시 실적변수와는 달리 포함되지 않은 위험요인들의 잔류 효과(Residual Effect)를 나타내는 비관측 위험 Θ가 주어졌을 때 포아송분포를 따른다고 가정하였다. 여기서 Λ는 무작위로 뽑은 계약자에 대한 사전적 평균 사고 건수를 뜻하고 ($\Lambda = e^{x'\beta}$), $\Theta \sim \log normal(0, \sigma^2)$를 따른다고 본다. 그러면 GLMM 분석을 통해 다음의 표로부터 비관측 위험 Θ에 대한 분산을 구할 수 있게 된다.

$$Y|\Theta \sim Poi(\Lambda \bullet \Theta). \qquad \text{(Ⅲ-8)}$$

〈표 Ⅲ-2〉 질병외래에 대한 GLMM Fixed Effect 결과

구분		Estimate	S. E.	z value	Pr(〉|z|)	
β_0		-3.200671	0.1537	-20.829	〈 2e-16	***
성별: 여		0.6766	0.0525	12.879	〈 2e-16	***
상해 급수	2급	-0.0637	0.0615	-1.036	0.3003	-
	3급	-0.2463	0.0851	-2.893	0.0038	**
연령		0.0197	0.0033	5.985	2.17e-09	***

〈표 Ⅲ-3〉 질병외래에 대한 GLMM Random Effect 결과

계약자	1	2	3	4	...	N	S.D.
γ_0	...	...	...	...	...	...	1.241

식 (Ⅲ-9)는 상대도를 구하는 식이다. 이 식에 $\theta \sim \log normal(0, 1.241^2)$을 이산화하여 $F_\Theta(\theta)$와 그에 해당하는 θ를 구하여 대입하면 된다. 그리고 λ_k는 k번째 위험등급 내 평균 사고건수를 산출하여 대입하면 되고, ω_k는 k번째 위험등급의 비중을 구해서 대입하면 된다. 이러한 방법으로 구한 최종 상대도는 다음의 〈표 Ⅲ-4〉와 같다.

$$r_l = \frac{\sum_k \omega_k \int_0^\infty \theta \pi_l(\lambda_k \theta) dF_\Theta(\theta)}{\sum_k \omega_k \int_0^\infty \pi_l(\lambda_k \theta) dF_\Theta(\theta)}. \qquad \text{(Ⅲ-9)}$$

〈표 Ⅲ-4〉 질병외래 계약자 포트폴리오 및 상대도 분석

무사고 누적연수	계약자 포트폴리오	상대도 (r_l)	상대도 비율
0년	0.2022	2.4744	1.0000
1년	0.1173	1.3303	0.5376
2년	0.0828	1.0120	0.4090
3년	0.0632	0.8394	0.3392
4년 이상	0.5345	0.3868	0.1563

본 분석에서는 실손의료보험 데이터 중에서 질병외래 계약자에 대한 상대도를 분석하였다. 먼저 할인할증제도를 위험집단에 적용 가능할지에 대한 판단을 하기 위해서는 계약자 포트폴리오 비중을 비교해보면 된다. 〈표 Ⅲ-4〉와 같이 계약자 포트폴리오가 어느 한 위험등급에 과도하게 비중이 많이 쏠려있다고 볼 수 없는 경우에는 이러한 계약자 그룹에 대해 무사고 누적연수(위험집단 구분 변수)를 기준으로 할인할증을 적용하는 것이 무리는 아니라고 볼 수 있다. 〈표 Ⅲ-4〉와 같이 할인할증제도를 적용해 볼 수 있는 대상 집단들에 대하여 차별적인 할인할증 폭을 적용해보면 누적연수가 0년인

계약자보다 4년 이상인 계약자가 약 85%(0.85=1-0.15) 할인을 받게 된다고 볼 수 있다. 이와 같이 계약자들을 과거 관측 및 비관측 정보들을 이용하여 위험특성별로 구분한 후 상대도를 구한다면 각 위험집단별로 보험료율을 차등화하여 적용해 볼 수 있는 것이다.

Ⅳ. 의사결정나무와 MARS

1. 실손의료보험 자료를 활용한 CART 분석

가. CART(Classification and Regression Tree)

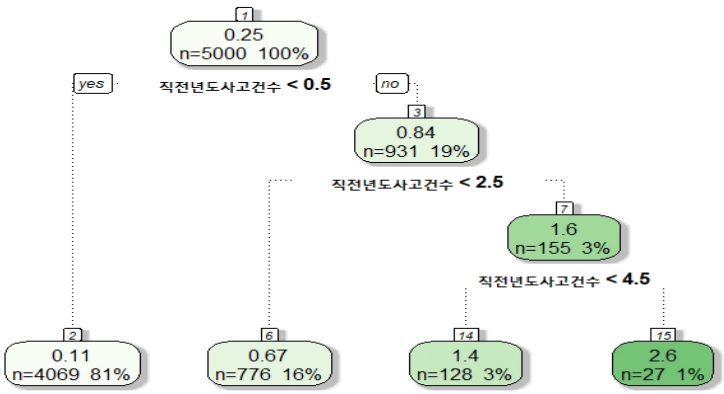
CART는 데이터를 가장 잘 분류해주는 도구라고 할 수 있는데 데이터를 잘 분류해주는 분리변수를 선택하고 분리지점을 정해준다. 그리고 가지치기를 통해서 분류의 정도를 결정할 수 있다. 데이터마이닝 방법론들 중 가장 널리 쓰이는 방법론으로 반응변수가 범주형 또는 연속형일 때 가능한 의사결정나무의 한 알고리즘이다. 이는 의사결정나무가 다지분류가 아닌 두 갈래로만 나누어져 예측력은 낮지만 해석하기가 용이한 장점이 있다. 이번 절에서는 실손의료보험 자료를 활용한 CART 분석의 예제를 먼저 보여주고 방법론적인 내용은 다음 절에서 소개한다. 예제로 사용한 데이터의 독립변수는 성별, 연령, 상해급수, 직전연도 발생건수이고 종속변수는 사고건수이다. 다음 절에서는 종속변수를 심도인 사고금액으로 하여 분석한다.

1) 빈도 분석

아래의 그림은 빈도 의사결정나무로 각 노드마다의 예측값과 전체의 몇 퍼센트가 해당 노드에 있는지에 대한 정보를 담고 있다. 가장 상위 노드에서 직전연도 발생건수 0.5건을 기준으로 가지가 분류된다. 0.25인 첫 번째 노드에서 직전연도 발생건수가 0.5 이하인 집단의 예측값은 0.11로 전체의 81%를 차지한다. 발생건수가 0.5 이상인 노드의 예측값은 0.84건으로 전체의 19%이다. 더하여 직전연도 0.5건 이상으로 분류된 노드는 직전연도로 다시 구분된다.

의사결정나무는 변수들 중 가장 설명력이 있는 변수에 최초로 분리가 일어난다. 이러한 점으로 미루어 보아 아래 의사결정나무는 직전연도 발생건수가 중요한 변수라는 것을 알 수 있다.

〈그림 Ⅳ-1〉 빈도 의사결정나무 cp=0.01



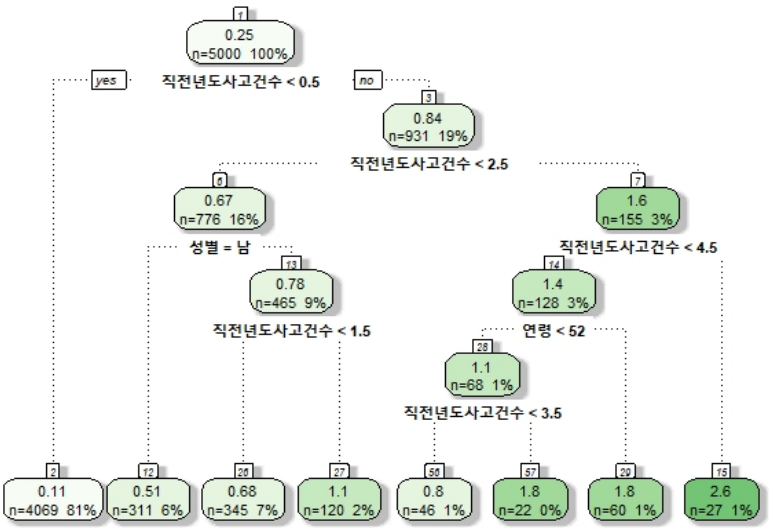
아래 표는 뿌리마디부터 끝마디까지 나무가 성장하면서 달라지는 지표들을 요약한 것이다. 첫 번째 행에 있는 cp(complexity parameter)는 나무의 크기를 통제할 때 쓰인다. 만약 cp가 작으면, 끝마디의 숫자가 많아지는 것에 대한 벌점이 작으므로 나무의 크기가 커진다. 〈그림 Ⅳ-1〉과 〈그림 Ⅳ-2〉를 비교하면 cp=0.005(〈그림 Ⅳ-2〉)일 때의 나무의 사이즈가 cp=0.01(〈그림 Ⅳ-1〉)일 때보다 큰 것을 볼 수 있다. cp에 관한 더 자세한 내용은 다음 장에서 서술한다. 두 번째 행에 있는 나뭇가지가 분리됨에 따라 rel.error가 감소한다. 여기서 rel.error는 R-squared와 관련된 지표이다.

〈표 Ⅳ-1〉 빈도 의사결정나무 요약

구분	CP	nsplit	rel. error	xerror	xstd
1	0.1693	0	1.0000	1.0001	0.0601
2	0.0521	1	0.8308	0.8311	0.0491
3	0.0126	2	0.7786	0.7878	0.0451
4	0.0100	3	0.7661	0.7832	0.0445

〈그림 Ⅳ-2〉는 cp가 0.005로 〈그림 Ⅳ-1〉의 의사결정나무에 비해 세분화되어 있는 것을 볼 수 있다. 설명변수 분류의 순서를 보면 상위에서 하위항목으로 갈수록 발생건수, 성별, 연령 순이다. 이는 GLM 분석 결과의 p-value 값이 작은 순과 같다. GLM에서는 p-value 값이 작을수록 유의미한 변수라고 해석한다.

〈그림 Ⅳ-2〉 빈도 의사결정나무 cp=0.005



처음의 세 번 분리는 cp가 0.01(〈표 Ⅳ-1〉)일 때와 같다.

〈표 Ⅳ-2〉 빈도 의사결정나무 요약 cp=0.005

구분	CP	nsplit	rel. error	xerror	xstd
1	0.1693	0	1.0000	1.0002	0.0601
2	0.0521	1	0.8308	0.8316	0.0492
3	0.0126	2	0.7786	0.7941	0.0471
4	0.0067	3	0.7661	0.7811	0.0457
5	0.0060	4	0.7594	0.7897	0.0463
6	0.0057	5	0.7534	0.7890	0.0463
7	0.0053	6	0.7478	0.7873	0.0462
8	0.0050	7	0.7424	0.7864	0.0462

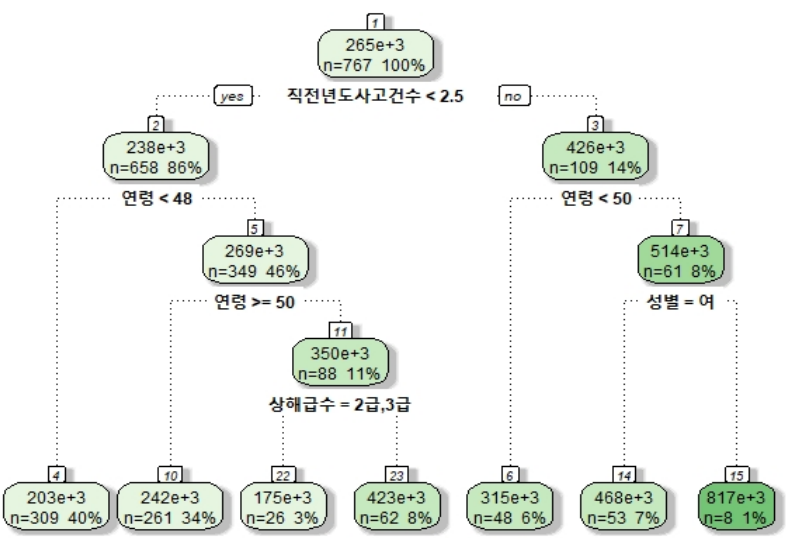
2) 심도 분석

심도 의사결정나무모형은 빈도모형보다 다양한 설명변수가 채택되었지만 여전히 직전연도 발생건수가 최초분리 변수로서 설명력이 높은 변수라는 것을 보여준다. 위 모델의 데이터는 빈도데이터에서 지급액이 0 이상인 것들을 조건으로 한 후 사용한다. 좌측으로 분리되면 분리기준에 수긍하는 것이고 우측으로 분리되면 수긍하지 않는다는 것을 의미한다.

〈표 Ⅳ-3〉 심도 의사결정나무

구분	CP	nsplit	rel. error	xerror	xstd
1	0.0416	0	1	1.0009	0.1604
2	0.0133	1	0.9584	0.9621	0.1559
3	0.0109	2	0.9452	1.0067	0.1570
4	0.0106	5	0.9125	1.0090	0.1573
5	0.0100	6	0.9019	1.0096	0.1566

〈그림 Ⅳ-3〉 심도 의사결정나무



2. CART(Classification and Regression Trees)[2)]

의사결정나무는 쉽게 말하면 의사결정규칙(Decision rule)으로 이루어진 나무 모양을 그리는 것이라고 할 수 있다. 의사결정나무는 과거에 수집된 데이터들을 분석하고, 이 데이터들 사이에 존재하는 패턴들의 특성을 속성의 조합으로 나타내는 분류 모형이다. 이는 새로운 데이터에 대해 분류(Classification)하거나 해당 범주의 값을 예측하는 목적으로 쓰인다. 모형화(Predictive Modeling) 자체가 분류 및 예측모형으로도 사용될 수 있다. 또한 탐색(Exploratory data analysis)으로 모형화에 앞서, 이상치(Outlier)의 검색, 변수의 선택, 교호작용 파악 등에 사용된다. 종속변수의 유형이 범주형, 연속형인지에 따라 각각 분류나무(Classification Tree)와 회귀나무(Regression Tree)로 분류한다.

CART의 장점으로는 모형을 해석하고 이해하기 쉽고 입력변수를 선정하는 데에도 매우 유용하다는 점이다. 또한 극단치(Outlier)에 덜 민감(Robust)하다. 그리고 비모수적인 방법이어서 분포에 대한 가정이 필요 없고 비선형적인 방법이다.

의사결정나무(Decision Trees, 이하 DT)는 주어진 입력값을 이용하여 출력값을 예측하는 모형이며 그 종류로는 분류나무(Classification Trees)와 회귀나무(Regression Trees) 모형이 있다. 의사결정나무는 예측한 결과를 나무형태의 그래프로 나타낼 수 있다는 사실에 기인하여 이름이 붙여졌다. 이 장에서는 분류 및 회귀 의사결정나무의 구조, 형성과정, 분리기준으로 사용되는 여러 가지 불순도의 측도, 여러 가지 알고리즘에 대하여 자세히 서술한다.

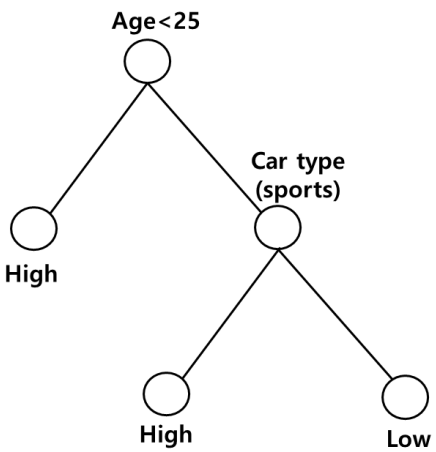
가. 의사결정나무의 구조

의사결정나무(DT)의 구조는 크게 노드(Node), 가지(Branch) 그리고 깊이(Depth)로 구성되어있다. 가지(Branch)라는 것은 하나의 마디부터 끝마디까지 하나로 연결된 마디들을 가리키고 깊이(Depth)는 가지를 이루고 있는 마디의 개수를 말한다. 각 Node

2) 의사결정나무 방법론은 박창이·김용대·김진석·송종우·최호식(2011)을 참고하여 요약 및 정리함

마다 불리는 이름이 있다. 뿌리마디(Root Node)는 DT가 시작되는 마디로 전체 데이터를 구성한다. 자식마디(Child Node)는 하나의 마디로부터 분리되어 나간 마디이다. 부모마디(Parent Node)는 자식마디의 상위마디를 일컫는다. 끝마디(Terminal Node)는 말 그대로 끝 마디로 더 이상의 분할이 이루어지지 않는 마디를 말한다. 중간마디(Internal Node)는 나무구조의 중간에 있는 마디이다.

〈그림 Ⅳ-4〉 의사결정나무 예시



위 그림은 간단한 의사결정나무 예시를 보여주고 있다. 맨 위쪽에 나이를 나누는 노드는 뿌리마디(Root Node)이다. 그 다음에 한 칸 내려와서 Car type이 Sports인지를 나눠주는 노드를 중간마디(Internal Node)라고 한다. 그리고 끝단에서 Risk가 High인지 Low인지 나타내주는 노드는 끝마디(Terminal (Leaf) Node)라고 할 수 있다.

회귀나무(Regression Trees)에서 사용되는 분리기준은 분산의 감소량이다. 예측오차를 최소화하는 것과 동일한 개념으로 분산의 감소량을 최대화하는 것을 최적분리의 기준으로 삼아 자식마디를 형성하면 된다.

나. 의사결정나무의 형성

의사결정나무의 형성과정은 크게 성장(Growing), 가지치기(Pruning), 타당성 평가 그리고 해석과 예측으로 이루어진다. 성장 단계라는 것은 각 마디에서 최적의 분리규칙을 적절하게 찾아서 나무를 성장시키는 과정이며, 적절한 정지규칙을 만족하면 중단하는 것으로 한다. 가지치기 단계라는 것은 오차를 크게 할 위험이 높을 경우나 부적절한 추론규칙을 가지고 있는 가지 또는 불필요한 가지를 제거하는 것을 의미한다. 타당성 평가 단계라는 것은 이익도표(Gain chart), 위험도표(Risk chart) 또는 시험자표를 활용하여 의사결정나무를 평가하는 것을 의미한다. 마지막으로 해석과 예측 단계라는 것은 구축된 나무모형을 해석하고 예측모형을 설정한 후 예측에 적용하는 것을 의미한다.

의사결정나무는 출력변수가 연속형인 회귀나무(Regression Tree)와 범주형인 분류나무(Classification Tree)로 나눌 수 있다. 회귀나무와 분류나무의 형성과정을 아래에서 살펴본다.

1) 회귀나무(Regression Trees)

회귀나무는 회귀나무를 어디까지 성장시킬지가 관심사일 것이다. p개의 입력변수와 하나의 종속변수로 이루어진 N개의 관측 데이터가 있다고 가정한다. $(x_i, y_i), i = 1, ..., N$으로 정의하고, $x_i = (x_{i1}, ..., x_{ip})$인 행벡터로 정의한다. 알고리즘은 분리변수(Split variable)와 분리점(Split point)을 결정하고 또한 나무모형이 어떻게 생길지도 결정해야 한다. 먼저 전체 영역을 M개의 영역 $R_1, ..., R_M$으로 나누고 상수값 c_m을 각 영역의 예측값으로 하는 나무모형은 다음과 같이 표현된다.

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m). \qquad (Ⅳ\text{-}1)$$

여기서 회귀나무 기준으로 오차제곱합 $Q_m(T) = \sum_{i=1}^{n} (y_i - f(x_i))^2$을 그 측도로서 사용한다. 그러면 이에 대해 최솟값을 갖는 $\widehat{c_m}$은 영역 R_m에서 y_i의 평균일 것이다.

$$\widehat{c_m} = avg(y_i | x_i \in R_m). \quad (Ⅳ-2)$$

주어진 분리변수 x_j가 연속형인 경우 분리점을 s라 하면 두 영역 $R_1(j,s) = \{x : x_j \leq s\}$와 $R_2(j,s) = \{x : x_j > s\}$을 정의할 수 있다. 범주형 분리변수일 경우에는 전체 범주를 부분집합 2개로 나눈다. 예를 들면 전체 범주가 {남, 여}일 때 $R_1(j) = \{남\}$과 $R_2(j) = \{여\}$로 나눌 수 있다. 그러면 분리기준을 정하는 것은 분리변수 j와 분리점 s를 찾는 최적화 문제로 볼 수 있다.

$$\min_{j,s}\left(\min_{c_1}\sum_{x_i \in R_1(j,s)}(y_i - c_1)^2 + \min_{c_2}\sum_{x_2 \in R_2(j,s)}(y_i - c_2)^2\right). \quad (Ⅳ-3)$$

분리변수가 주어지고 나면 어렵지 않게 분리점 s를 찾을 수 있으며 적절한 최적화를 통하여 최적 분리기준(j,s)를 찾을 수 있다. 우선 최적 분리를 찾고 난 후에는 두 영역에 대하여 반복하여 동일한 과정 거치면 된다.

나무모형이 너무 크면 자료를 과대적합할 가능성이 있고, 반대로 나무모형이 너무 작으면 자료를 과소적합할 가능성이 있다. 즉, 의사결정나무에서는 나무의 크기가 모형의 복잡도(Complexity)를 의미하며, 최적의 나무 크기는 사용하는 자료들로부터 추정한다. 일반적으로 사용되는 방법은 마디에 속하는 자료가 일정 수 이하일 때 분할을 정지하고 비용-복잡도 가지치기(Cost-complexity pruning)를 이용하여 성장시킨 나무를 가지치기하게 된다.

성장시킨 나무모형 T_0를 가지치기하여 얻을 수 있는 나무모형을 $T \subset T_0$로 나타내자. $|T|$는 T에서의 끝마디 개수, N_m은 T의 영역 R_m에 속하는 자료 수, $\widehat{c_m}$은 영역 R_m에 속하는 자료에 대한 y값들의 평균, 그리고 불순도는 $Q_m(T) = \frac{1}{N_m}\sum_{x_i \in R_m}(y_i - \widehat{c_m})^2$로 나타낸다. 이 때 최적화할 비용함수는 다음과 같이 정의된다.

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|. \quad (Ⅳ-4)$$

가지치기는 α에 대하여 $C_\alpha(T)$를 최소화하는 $T_\alpha \subset T_0$를 찾는 문제가 된다. 여기서 $\alpha \geq 0$는 Complexity parameter이며 데이터 분석가가 나무모형의 크기와 자료에 대한 적합도를 조절하기 위한 조율모수로서 선택할 수 있다. α값이 크면 T_α의 크기는 작아지고, 반대의 경우도 마찬가지이다. 그리고 $\alpha = 0$이면 가지치기는 일어나지 않고 T_0를 최종모형으로 준다.

추정값 $\hat{\alpha}$은 자료로부터 흔히 5 또는 10-묶음 교차확인오차로 얻을 수 있다. 가지치기된 최종 모형은 $T_{\hat{\alpha}}$으로 나타낼 수 있고 시험자료가 $x \in R_m$이면 $\hat{y} = \widehat{c_m}$으로 예측한다.

2) 분류나무(Classification Tree)

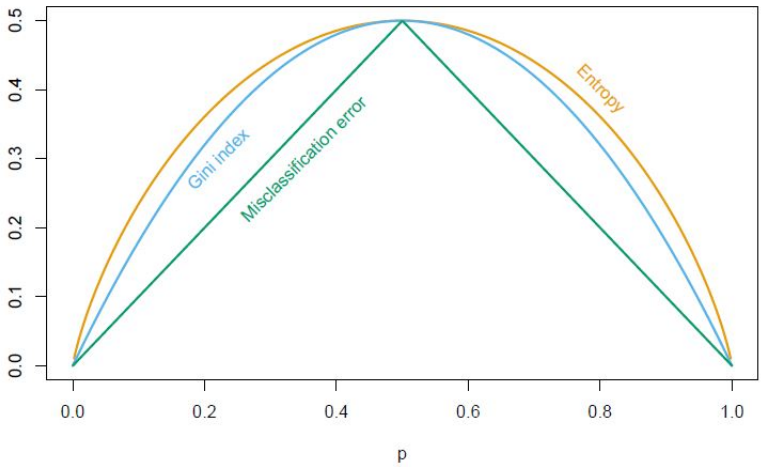
출력변수가 범주형인 분류나무는 불순도의 측도로 주로 사용되는 카이제곱 통계량, 지니지수(GINI index), 엔트로피지수(Entropy index) 등을 이용하여 회귀나무와 동일한 방식으로 성장시키게 된다. 분류나무의 가지치기는 흔히 오분류율을 불순도의 측도로 사용하여 회귀나무와 동일한 방식으로 실시하여 최종 분류나무모형 $T_{\hat{\alpha}}$을 얻게 된다. $\widehat{p_{mk}}$를 최종 모형의 영역 R_m에 속하는 자료 중 출력변수의 범주가 k인 자료의 비율이라 하자. $x \in R_m$이면 그 예측값은 $\hat{y} = \arg max_k \widehat{p_{mk}}$로 주어진다. 즉, 분류나무는 각 마디에서 다수결원칙(Majority vote)으로 정하는 것이다.

다. 분류나무의 여러 가지 불순도 측도

불순도의 측도는 의사결정나무의 성장 단계에서 최적의 분리변수(Splitting Variable)와 기준값(Threshold)을 정하는데 사용된다. 회귀나무에서는 불순도의 측도를 $Q_m(T)$로 정의하였다면 분류나무에서 사용되는 측도는 카이제곱(χ^2) 통계량, 지니지수, 엔트로피지수, 분류오차 등이다.

분류나무의 경우 데이터의 분리/분할은 각 자식마디에 속하는 자료의 순수도(Purity) 또는 불순도(Impurity)가 가장 크게 증가 또는 감소하도록 진행된다.

〈그림 Ⅳ-5〉 불순도 측도 비교



자료: Hastie, Tibshirani, Friedman(2008), p. 309

불순도 예시 표를 바탕으로 각 측도들이 어떻게 계산되는지 살펴본다.

〈표 Ⅳ-4〉 불순도 예시

구분	남성	여성	전체
왼쪽 마디	47(42)	23(28)	70
오른쪽 마디	73(78)	57(52)	130
부모마디	120	80	200

1) 카이제곱 통계량

위의 표는 실제 도수(O)와 기대 도수(E)(괄호 안 숫자)를 보여주고 있다. 예를 들어 왼쪽의 남성의 기대 도수라는 것은 $70 \times 120/200 = 42$이며, 다른 셀들의 기대 도수 역시 동일한 방법으로 구할 수 있다. 카이제곱 통계량이라는 것은 각 셀에 대하여 ((기대 도수-실제 도수)의 제곱/기대 도수)의 합으로 정의된다. 그리고 카이제곱 통계량이 최대가 되는 분리를 사용한다. 이 표에서 카이제곱 통계량은 다음과 같이 계산된다.

$$\frac{(42-47)^2}{42}+\frac{(28-23)^2}{28}+\frac{(78-73)^2}{78}+\frac{(52-57)^2}{52}=2.2893. \quad (Ⅳ\text{-}5)$$

2) 지니지수(GINI index)

지니지수는 CART에서 쓰이는 지표이며, 지니지수는 다음과 같이 정의된다.

$$GINI(t)=1-\sum_j [p(j|t)]^2. \quad (Ⅳ\text{-}6)$$

$$GINI_{split}=\sum_{i=1}^{k}\frac{n_i}{n}GINI(i). \quad (Ⅳ\text{-}7)$$

지니지수가 최소가 되는 분리를 선택한다. 앞의 표에 대하여 지니지수를 구하면 다음과 같이 주어진다.

$$\begin{aligned}&2(\mathbb{P}(\text{왼쪽에서 남성})\mathbb{P}(\text{왼쪽에서 여성})\mathbb{P}(\text{왼쪽})\\&+\mathbb{P}(\text{오른쪽에서 남성})\mathbb{P}(\text{오른쪽에서 여성})\mathbb{P}(\text{오른쪽}))\\&=2\left(\frac{47}{70}\times\frac{23}{70}\times\frac{70}{200}+\frac{73}{130}\times\frac{57}{130}\times\frac{130}{200}\right)=0.4745. \quad (Ⅳ\text{-}8)\end{aligned}$$

3) 엔트로피지수(Entropy measure)

엔트로피지수는 C4.5에서 불순도 측도로 사용되는 것으로 다음과 같이 정의된다.

$$Entropy(t)=-\sum_j p(j|t)\log_2 p(j|t). \quad (Ⅳ\text{-}9)$$

$$\begin{aligned}&\text{엔트로피지수} = \text{엔트로피(왼쪽)}\mathbb{P}(\text{왼쪽}) \quad (Ⅳ\text{-}10)\\&\qquad\qquad + \text{엔트로피(오른쪽)}\mathbb{P}(\text{오른쪽})\end{aligned}$$

여기서

$$\text{엔트로피(Left)} = -\mathbb{P}(\text{왼쪽에서 남성})\log_2 \mathbb{P}(\text{왼쪽에서 남성}) - \mathbb{P}(\text{왼쪽에서 여성})\log_2 \mathbb{P}(\text{왼쪽에서 여성}) \quad (Ⅳ\text{-}11)$$

로 정의되며, 오른쪽 마디에 대한 엔트로피도 이와 동일한 방법으로 정의될 수 있다. 앞의 표에서 엔트로피지수를 구하면 다음과 같이 얻을 수 있다.

$$-\left(\frac{47}{70}log_2\left(\frac{47}{70}\right)+\frac{23}{70}log_2\left(\frac{23}{70}\right)\right)\frac{70}{200} - \left(\frac{73}{130}log_2\left(\frac{73}{130}\right)+\frac{57}{130}log_2\left(\frac{57}{130}\right)\right)\frac{130}{200} = 0.9626. \quad (Ⅳ\text{-}12)$$

4) 분류오차(Misclassification error)

$$Error(t) = 1 - \max P(i|t) \quad (Ⅳ\text{-}13)$$

라. 여러 가지 의사결정나무 알고리즘

위쪽에서는 주로 CART 방법론에 초점을 맞춰 살펴보았다. CART는 가장 많이 쓰이는 의사결정나무 알고리즘으로 Breiman, Friedman, Stone & Olshen(1984)가 개발하였다. 출력변수가 범주형인 경우 불순도를 앞 절에서 살펴본 지니지수를 통해 계산하고 출력변수가 연속형인 경우는 분산을 이용한다. CART는 이진분리(Binary split)하여 해석성이 좋다는 장점이 있다. 개별 입력변수 또는 입력변수들의 선형결합들 중에서 최적의 분리를 찾으면 된다.

CART 다음으로 유명한 의사결정나무 방법론에는 ID3이 있고 그 후에 더 발전된 C4.5와 C5.0이 있다. ID3은 처음 호주의 연구원인 Quimlan(1993)에 의하여 개발되었다. 위에서 언급한 CART와는 달리 각 마디에서 다지분리(Multiple split)가 가능하며 범주형 입력변수에 대해서는 범주의 수만큼 분리가 일어난다. 초기 버전은 범주형 예측변수에만 국한되어있었으나 최근에 발전된 C5.0은 CART와 매우 유사해졌다. 불순

도의 측도로는 앞 장에서 살펴본 엔트로피지수를 사용한다.

CHAID(Chi squared Automatic Interaction Detection)는 카이제곱 검정에 근거한 모수적인 방법이며 불순도의 측도로는 카이제곱 통계량을 사용한다. Hartigan(1975)이 제안한 방법으로 Morgan & Sonquist(1963)의 AID를 발전시킨 것으로 볼 수 있다. CHAID는 가지치기를 하지 않는 대신 나무모형의 성장을 적당한 크기에서 중지하면 된다. 그리고 한계점은 입력변수가 반드시 범주형 변수여야 한다는 점이 있다.

마. CART의 특징

CART는 이진분리의 if-then 형식의 이해하기 쉬운 규칙을 생성하며 분류작업이 쉽다. 또한 연속형 변수와 범주형 변수의 형태 모두 입력변수로 취급할 수 있으며 비모수적 방법이라는 장점이 있다. CART는 가장 설명력이 있는 변수에 대하여 최초로 분리가 일어난다는 특징 때문에 요율산정에 있어서 중요변수를 알아낼 수 있다.

단점으로는 출력변수가 연속형인 회귀모형에서는 예측력이 감소한다는 것이다. 일반적으로 복잡한 나무모형은 예측력이 저하되고 해석이 어렵다. 상황에 따라서는 많은 양의 계산 작업이 필요할 수도 있으며, 베이즈 분류경계가 사각형(Rectangle)이 아닌 경우에는 결과가 좋지 않을 수도 있다. 특히 자료가 조금만 달라져도 전혀 다른 결과를 얻을 정도로 분산이 크고 불안정한 방법이다. 앙상블 알고리즘을 적용하여 의사결정나무의 분산을 줄일 수도 있다.

3. MARS(Multivariate Adaptive Regression Splines)

가. MARS 방법론

MARS 방법론은 주로 Francis(2003), Hastie, Tibshirani & Friedman(2008)을 주로 참고하여 정리하였다. Friedman(1991)이 제안한 MARS는 입력변수가 많은 고차원의

회귀문제에 적합한 방법이다. 설명변수들이 종속변수에 대해 직선이 아닌 꺾인 선형 형태(Splines)로 설명된다. 단계별 선형회귀의 일반화 또는 의사결정나무의 개선으로 볼 수 있다. MARS는 $(x-t)_+$와 $(t-x)_+$ 형태의 매듭점(Knot point) t에서의 조각별 선형회귀의 기저함수(Basis function)를 사용한다. 여기서 $(x)_+ = xI(x>0)$를 의미하며, 괄호 안의 값 x의 양수부분만을 나타낸다.

$$(x-t)_+ = \begin{cases} x-t, & x>t \\ 0, & otherwise \end{cases} \qquad (Ⅳ\text{-}14)$$

$$(t-x)_+ = \begin{cases} t-x, & x<t \\ 0, & otherwise \end{cases}.$$

기저함수의 클래스는 각 입력변수 X_j에 대하여 관측값 x_{ij}들을 매듭점으로 한 선형 스플라인들로 표현된다. MARS 모형은 식 (Ⅳ-15)로 표현되며 $B_m(X)$은 B상의 기저함수 혹은 B에 속하는 둘 이상의 기저함수들의 곱이다.

$$B = \{(X_j - t)_+, (t - X_j)_+ : t \in x_{1j}, \ldots, x_{nj} \text{ for } j = 1, \ldots, p\}. \qquad (Ⅳ\text{-}15)$$

모델링은 원래의 입력값이 아닌 식 (Ⅳ-16)과 같이 변형된 형태로 전진단계선형회귀(Forward stepwise linear regression)로 한다. 어떤 $B_m(X)$을 선택할지 주어지면, 전통적 선형회귀와 같이 오차제곱합을 최소화시키는 β_m계수들을 추정한다.

$$\mu(X) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X). \qquad (Ⅳ\text{-}16)$$

B상의 한 기저함수만을 사용하는 경우에는 주효과만을 사용하는 가법모형이고, B상의 두 기저함수들의 곱까지 허용하는 경우에는 2인자 교호작용이 있는 모형에 해당된다.

MARS에서 기저함수는 전진 선택법을 사용하여 선택된다. 우선 β_0에 해당하는 $B_0(X)=1$을 모형에 투입하고, 각 단계에서 오차제곱합 $\sum_{i=1}^{n}(y_i - \mu(x_i))^2$을 최소화하는 변수와 매듭점을 찾고, 해당 기저함수쌍을 모형에 추가한다. 예를 들어 M개의 기

저함수가 선택되었다고 하자. 그러면 이번에는 자료의 과대적합을 막기 위해 후진 소거법으로 설명력이 없는 기저들을 제거한다. 이 때 사용되는 기저함수 선택기준은 식 (Ⅳ-17)로 $\hat{\mu}_m(x)$는 m개의 항들에 기반한 $\mu(x)$의 적합값이고 $C(m)$은 모수의 개수로 정의되는 복잡도함수이다. 최종적으로 $m^* = \arg\min GCV(m)$개의 항을 갖는 모형을 선택한다.

$$GCV(m) = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\mu_m}(x_i)\right)^2}{(1 - C(m)/n)^2}. \qquad (Ⅳ\text{-}17)$$

의사결정나무는 각 영역에서 동일한 상수값을 추정하는 반면 MARS에서는 선형 스플라인을 사용하므로 조각별로 선형인 연속함수로 추정하게 된다. 따라서 일반적으로 MARS는 의사결정나무에 비해서 예측력이 높다. MARS에서 조각별 선형 기저함수를 사용하는 이유는 비선형적인 모형을 조각별 선형함수로 근사하기 위한 것이다. 자료의 특성에 따라서는 교호작용을 허용할 수도 있는데 대체로 과대적합의 문제로 2인자 이상의 고차의 교호작용은 고려하지 않는다.

MARS와 CART는 서로 상이한 방법론처럼 보임에도 불구하고 굉장히 유사한 점들이 있다(Hastie, Tibshirani & Friedman 2008). MARS에서 기저함수를 $I(x-t>0)$과 $I(x-t\le 0)$으로 대체한다고 가정하면 MARS의 전진선택법 알고리즘과 CART가 나무를 성장시키는 알고리즘을 동일하게 볼 수 있다. 또한 MARS 모형에서 기저함수의 짝이 곱 형태가 포함되어 있으면 CART 모형에서 나뭇가지가 분리되는 것과 동일하다. 이러한 것은 CART에서 노드에서 한 번 이상의 분리를 할 수 없도록 제한하고 이진분리로 나타내게 한다.

나. 실손의료보험 자료를 활용한 MARS 분석

1) 빈도 분석

〈표 Ⅳ-5〉 질병외래 빈도 MARS 분석

구분	기저함수	계수
BF_0	(Intercept)	0.5518
BF_1	성별 여	0.0888
BF_2	$(1-15$년도 발생건수$)_+$	-0.4806
BF_3	$(15$년도 발생건수$-1)_+$	0.2732
BF_4	$(15$년도 발생건수$-2)_+$	0.1810

성별, 연령, 상해급수, 직전연도 발생건수를 입력변수로 갖고 있음에도 불구하고 빈도모형에서는 성별과 발생건수 변수가 더 중요한 변수로 선택되어 사용된다. 〈표 Ⅳ-5〉는 다음 식과 같이 쓸 수 있다.

$$Y = 0.5518 + 0.0888BF_1 - 0.4806BF_2 + 0.2732BF_3 + 0.181BF_4. \qquad (Ⅳ\text{-}18)$$

여성일 때의 계수가 양수인 것을 미루어 보아 기대빈도가 남성보다 여성일 때 높은 것으로 해석된다. 직전연도 발생건수에 대하여 1건과 2건이 기준이 되는데 0건일 때의 기대빈도는 0.4806만큼 감소하는 반면 2건일 때에는 0.2732만큼 증가한다. 2건 초과일 때에는 과거실적 1건당 0.4542(=0.2732+0.1810)씩 기대빈도가 높아진다.

2) 심도 분석

〈표 Ⅳ-6〉 질병외래 심도 MARS 분석 thresh=0.001

구분	기저함수	계수
BF_0	절편	172,646
BF_1	성별 여	28,346
BF_2	$(\text{연령}-47)_+$	118,545
BF_3	$(\text{연령}-48)_+$	-198,690
BF_4	$(\text{연령}-49)_+$	257,027
BF_5	$(50-\text{연령})_+$	740
BF_6	$(\text{연령}-50)_+$	-292,756
BF_7	$(\text{연령}-52)_+$	171,010
BF_8	$(\text{연령}-54)_+$	-60,951
BF_9	$(2-\text{15년도 발생건수})_+$	1,100
BF_{10}	$(\text{15년도 발생건수}-2)_+$	210,175
BF_{11}	$(\text{15년도 발생건수}-3)_+$	-320,938
BF_{12}	$(\text{15년도 발생건수}-4)_+$	190,426

심도 분석은 지급금액이 0건 이상인 데이터를 사용하여 보험금액을 모형화한다. terminal condition은 빈도모형과 같게 했을 때를 먼저 살펴보고 조절하는 방향으로 한다. terminal condition 값은 모형에 항을 하나 추가하였을 때 개선되는 R^2값이다. 예를 들어 terminal condition이 0.01이라는 것은 모형에 항을 하나 추가하였을 때 R^2값이 0.01보다 커지지 않으면 모형의 성장을 멈춘다는 것이다. 빈도 분석일 때보다 계수들의 항이 증가한 것을 볼 수 있다. 성별이 여성인 경우 남성보다 28,346원 기대 지급금액이 증가한다. 성별에 대한 해석은 쉬우나 연령이나 직전연도 발생건수 같은 경우 다수의 항이 있어 해석하기 불편할 수도 있다.

〈표 Ⅳ-6〉의 모형과 같이 해석이 불편한 모형이 산출된 경우에는 함수 안의 옵션을 바꿔 보다 간편한 모형을 모델링할 수 있다. terminal condition을 0.01로 개선하여 다시 모델링하면 위의 모델에서 중요도가 높은 변수들로 구성된 모형을 얻는다. 연령은

50세를 기준으로 50세 미만일 때에는 1세 감소할 때 기대 지급금액이 7,365원 감소한다. 50세 초과인 연령대에서는 연령이 1세 증가할 때마다 451원 감소한다. 50세를 기준으로 뒤집어진 'V'모양을 볼 수 있다.

〈표 Ⅳ-7〉 질병외래 심도 MARS 분석 thresh=0.01

구분	기저함수	계수
BF_0	절편	293,289
BF_1	$(50-\text{연령})_+$	-7,365
BF_2	$(\text{연령}-50)_+$	-451
BF_3	$(2-15\text{년도 발생건수})_+$	-14,529
BF_4	$(15\text{년도 발생건수}-2)_+$	63,344

다. GLM option이 있는 MARS

MARS 분석은 단계별 선형회귀의 일반화로 반응변수가 0 미만인 값이 나올 가능성이 있다. 하지만 보험데이터의 빈도나 심도를 모형화한 경우 반응변수가 0 미만인 값이 나오는 것은 현실과 부합하지 않으므로 이를 방지하기 위하여 GLM option을 사용한다. R 프로그램의 earth 패키지에서 MARS 분석을 할 때에 GLM option을 삽입하고 연결함수를 로그로 취하면 0 이상의 값을 갖는다.

1) 빈도 분석

〈표 Ⅳ-8〉의 결과를 바탕으로 한 식은 다음과 같다. 앞 절과 다른 점이 있다면 반응변수에 로그 연결함수가 적용되어 있고 양변에 지수를 취하면 곱으로 표현된다는 것이다.

$$\ln(y) = -0.7607 + 0.4035BF_1 - 1.6295BF_2 + 0.4135BF_3 - 0.1465BF_4. \quad (Ⅳ\text{-}19)$$

$$y = e^{-0.7607 + 0.4035BF_1 - 1.6295BF_2 + 0.4135BF_3 - 0.1465BF_4}. \quad (Ⅳ\text{-}20)$$

앞 절에서 GLM option이 없을 때와 같은 변수들이 채택되는데 곱으로 표현할 수 있는 장점이 있다. 성별과 발생건수만 고려된 아래의 모형에 따르면 여성의 기대빈도는 1.49배 높다. 이는 앞 장의 네 가지 변수가 사용된 GLM 모형의 상대도와 비교하였을 때 변수가 적게 채택되어 계수 값이 집중된 것이라고 볼 수도 있다. 직전연도 발생건수는 1건 일어났을 때를 기준으로 0건일 때는 0.196배, 2건일 때에는 1.5121배, 3건 이상일 때에는 1건이 증가할 때마다 1.306배라고 해석할 수 있다.

〈표 Ⅳ-8〉 질병외래 빈도 GLM option이 있는 MARS 분석

구분	기저함수	계수
BF_0	Intercept	-0.7607
BF_1	성별 여	0.4035
BF_2	$(1-15년도\ 발생건수)_+$	-1.6295
BF_3	$(15년도\ 발생건수-1)_+$	0.4135
BF_4	$(15년도\ 발생건수-2)_+$	-0.1465

〈표 Ⅳ-9〉 질병외래 빈도 GLM option이 있는 MARS 상대도

설명변수		상대도
성별	남성	1.0000
	여성	1.4971
15년도 발생건수	0건	0.1960
	1건	1.0000
	2건	1.5121
	3건 이상	1.3060

2) 심도 분석

위의 심도 분석과 다른 점이라면 GLM option을 사용한 것이다. 계수의 부호방향이 위의 결과값과 같게 나왔다. 곱으로 표현되어서 비교하기 쉬운 점이 있다.

〈표 Ⅳ-10〉 질병외래 심도 GLM option이 있는 MARS thresh=0.01

구분	기저함수	계수
BF_0	절편	12.5917
BF_1	$(50-\text{연령})_+$	-0.0293
BF_2	$(\text{연령}-50)_+$	-0.0055
BF_3	$(2-\text{15년도 발생건수})_+$	-0.0537
BF_4	$(\text{15년도 발생건수}-2)_+$	0.1920

Ⅴ. 앙상블기법과 신경망모형

1. 앙상블기법[3)]

앙상블(Ensemble)기법은 CART라는 도구가 괜찮다는 철학하에 만들어진 것이다. 하지만 CART의 성능이 우수하지 못할 수 있기 때문에 이를 개선하기 위해 만들어졌다. 주어진 자료를 이용하여 여러 개의 예측모형을 먼저 만들고, 그 예측모형들을 결합하여 최종적으로 하나의 예측모형을 만드는 방법이다. 최초로 제안된 앙상블 알고리즘은 1996년에 만들어진 Breiman의 배깅(Bagging)이다. 그 이후 앙상블 방법은 예측력을 획기적으로 향상시킬 수도 있음이 경험적으로 입증되었다. 이번 장에서는 가장 널리 사용되는 앙상블기법들인 배깅(Bagging), 부스팅(Boosting), 랜덤 포레스트(Random forest)를 설명한다.

가. 배깅(Bagging)

배깅은 불안정한 예측모형에서 불안정성을 제거하고 예측력을 향상시키기 위하여 개발되었다(Breiman 1996). 자료의 작은 변화에 예측모형이 크게 변하는 경우를 학습방법이 불안정하다고 말한다. 예를 들어 의사결정나무에서는 첫 번째 노드의 분리변수를 찾을 때 비슷한 예측력을 갖는 변수가 다수 존재하여 자료의 작은 변화에도 첫 번째 분리변수가 바뀌는 경우가 생긴다. 첫 번째 노드의 분리변수가 바뀌면 자식노드에 포함되는 자료가 완전히 바뀌고 이에 최종 의사결정나무가 완전히 달라진다. 이처럼 학습방법의 불안정성은 예측력의 저하를 가져오고, 예측모형의 해석을 어렵게 만든

3) 앙상블기법은 박창이·김용대·김진석·송종우·최호식(2011)을 참고하여 요약 및 정리하였음

다. 데이터마이닝의 목적을 달성하기 위해 불안정한 학습방법을 안정적으로 만드는 작업은 필수적이며, 그러한 이유로 배깅이 개발되었다.

배깅(Bagging)은 Boostrap Aggregating의 준말로 자료에 대하여 여러 개의 붓스트랩(Boostrap) 자료를 생성하여 각각에 대한 예측모형을 생성한 후 조합하여 최종적으로 하나의 예측모형을 만드는 방법이다. 여기서 붓스트랩 자료라는 것은 주어진 자료를 이용하여 동일한 크기의 표본을 무작위로 복원추출한 자료를 의미한다.

$\mathcal{L} = (x_i, y_i)$은 훈련자료를 나타낸다고 할 때, 배깅 알고리즘을 정리하면 다음과 같다.

1. B개의 붓스트랩 자료 $\mathcal{L}^{*(b)}, b = 1, \ldots, B$를 만든다.
2. 각 붓스트랩 자료 $\mathcal{L}^{*(b)}$에 대해서 예측모형 $f^{(b)}(x)$를 구축한다.
3. B개의 예측모형을 결합하여 최종 모형 $\hat{f}$을 만든다.
   최종모형을 만드는 방법은
   (a) 회귀모형인 경우 $\hat{f}(x) = \sum_{b=1}^{B} f^{(b)}(x)/B$와 같이 평균을 취한다.
   (b) 분류모형인 경우 $\hat{f}(x) = \arg max_k \left(\sum_{b=1}^{B} I(f^{(b)}(x) = k) \right)$와 같이 투표(Voting)를 한다.

최적의 의사결정나무를 구축할 때 가장 어려운 부분은 가지치기이다. 가지치기를 위한 여러 가지 방법이 제안되었지만, 배깅은 각각의 의사결정나무를 구축할 때 가지치기를 하지 않아도 된다는 장점 때문에 많이 활용되는 방법이다. 배깅은 가지치기를 하지 않은 최대로 성장한 의사결정나무를 사용하기 때문에 계산량을 대폭 줄일 수 있는 장점이 있다. 배깅에서 가지치기가 왜 필요하지 않은지에 대한 설명은 다음 절에서 한다.

배깅이 왜 예측력을 크게 향상시킬 수 있는 방법론인지에 대하여 많은 이론적인 연구가 진행되었다. 그러한 연구들 중 Breiman의 이론을 소개하고자 한다. 여기서 소개

되는 이론적 설명은 예측모형을 만들 때 주로 대두되는 몇 가지 고려사항들에 대하여 답을 제공하고 있다. 특히 배깅에서 가지치기를 할 필요가 없는 이유에 대해 설명을 하고자 한다.

주어진 훈련자료 $\mathcal{L}$ 을 이용하여 구축된 예측모형 $\hat{f}(x)$는 $\mathcal{L}$ 에 의존한다. 이를 강조하기 위하여 $\hat{f}(x) = f(x, \mathcal{L})$이라고 쓰고, 주어진 예측모형$f(x, \mathcal{L})$에 대하여 평균예측모형 $f_A(x) = E_{\mathcal{L}} f(x, \mathcal{L})$로 정의한다. 여기서 기댓값은 훈련자료가 얻어진 모집단의 분포를 이용하여 구한다는 점에 유의하기 바란다. 다음 정리는 평균예측모형의 기대손실(또는 위험)이 단일 예측모형의 기대손실보다 작다는 것을 보여준다.

〈정리 1〉 (X, Y)를 $\mathcal{L}$과 독립인 미래의 관측값이라 하자. 제곱손실함수 $L(y,a) = (y-a)^2$에 대하여 $f(x, \mathcal{L})$와 $f_A(x)$의 기대손실 R 과 R_A를 다음과 같이 정의한다.

$$R = E_{(X,Y)} E_{\mathcal{L}} L(Y, f(X, \mathcal{L})),\ R_A = E_{(X,Y)} L(Y, f_A(X)).$$

그러면 항상 $R \geq R_A$가 성립한다.

P r*oof*. 제곱함수는 볼록함수이므로 Jensen 부등식에 의해서

$$E_{(X,Y)} E_{\mathcal{L}} f^2(X, \mathcal{L}) \geq E_{(X,Y)} f_A(X)^2$$

이 성립한다. 따라서

$$\begin{aligned} R &= E_{(X,Y)}[Y^2] - 2E_{(X,Y)}[Y E_{\mathcal{L}} f(X, \mathcal{L})] + E_{(X,Y)} E_{\mathcal{L}}[f^2(X, \mathcal{L})] \\ &\geq E_{(X,Y)}[Y^2] - 2E_{(X,Y)}[Y f_A(X)] + E_{(X,Y)}[f_A(X)^2] \\ &= E_{(X,Y)}[(Y - f_A(X))^2] = R_A. \end{aligned}$$

위의 증명에서 중요한 사실 하나를 확인할 수 있는데, $R - R_A$는

$$E_{X,Y}\{E_{\mathcal{L}} f^2(X, \mathcal{L}) - E_{\mathcal{L}} f(X, \mathcal{L}))^2\} = E_{X,Y}(Var_{\mathcal{L}} f(X, \mathcal{L}))$$

이다. 즉, $f(x, \mathcal{L})$의 분산(또는 불안정성)이 크면 평균예측모형이 원래의 예측모형을 크게 향상시키며, 반대로 분산이 작으면(또는 안정적이면) 평균예측모형의 예측력의 향상 정도가 줄어든다.

훈련자료로 얻은 모집단의 분포를 모르기 때문에 학습자료를 모집단으로 생각하고 이것의 평균예측모형을 구한 것이 배깅의 예측모형이다. 배깅은 주어진 예측모형의 평균예측모형을 구하고 분산을 줄여줌으로써 예측력을 높인다. 즉, 배깅은 예측모형의 편의(Bias)에는 영향을 미치지 않고 분산에만 영향을 미친다. 따라서, 배깅에 적합한 예측모형은 편의가 없고 분산이 큰 과대적합된 모형이다. 의사결정나무에 배깅을 적용할 때 나무를 최대한 성장시키고 가지치기를 하지 않는 것도 배깅의 효과를 극대화하기 위함이다.

나. 부스팅(Boosting)

부스팅은 예측력이 약한 예측모형(Weak learner)들을 결합하여 예측력이 최적에 가까운 강한 예측모형을 만드는 것을 말한다. 즉, 경계에 있는 데이터에 가중치를 더욱 부여(Boost)하여 만들어진 모형이다. 여기서 약한 예측모형이란 랜덤하게 예측하는 것보다 약간 좋은 예측력을 지닌 모형을 말한다. 반면 강한 예측모형이란 예측력이 최적에 가까운 예측모형을 말한다. 실제 자료분석을 위해 제안된 최초의 부스팅 알고리즘은 이진 분류문제에서 Freund & Schapire(1997)에 의해서 개발된 AdaBoost(Adaptive Boost) 알고리즘이다.

AdaBoost 알고리즘에서 주의 깊게 살펴보아야 할 부분은 단계 2의 (c)와 (d)이다. 부스팅에 사용되는 예측모형 f_m은 랜덤한 추측보다 조금 더 좋은 예측력을 갖는다고 가정하면, f_m의 (가중) 오분류율은 0.5보다 작게 되므로 (c)의 $c_m > 0$이 된다. 그러면 (d)에서 각 관측치에 할당되는 가중치가 f_m에 의해서 오분류된 관측치에서는 증가하고 정분류된 관측치에서는 기존의 값과 같게 된다. 가중치를 정규화하여 합이 1이 되도록 하면, AdaBoost 알고리즘은 매 반복마다 오분류된 관측치의 가중치는 증가시키고 정분류된 관측치는 감소시키면서 예측모형을 만들어 간다.

1. 가중치 $w_i = 1/n, i = 1, ..., n$를 초기화한다.
2. $m = 1,...,M$ 에 대하여 다음 과정을 반복한다.
   (a) 가중치 w_i 를 이용하여 분류기 $f_m(x) \in \{-1,1\}$를 적합한다.
   (b) err_m를 다음과 같이 계산한다.

$$err_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq f_m(x_i))}{\sum_{i=1}^{n} w_i}$$

   (c) $c_m = \log((1 - err_m)/err_m)$로 설정한다.
   (d) 가중치 w_i를 $w_i = w_i \exp(c_m I(y_i \neq f_m(x_i)))$로 업데이트 한다.
3. 단계 2에서 얻은 M개의 분류기를 결합하여 최종 분류기 $sign\left(\sum_{m=1}^{M} c_m f_m((x))\right)$를 얻는다.

AdaBoost 알고리즘의 본래의 목적은 훈련오차를 빨리 그리고 쉽게 줄이는 것이다. 약한 학습기 f_m의 오분류율이 항상 0.5-$\gamma(\gamma > 0)$이면 훈련오차는 지수적으로 빠르게 0으로 수렴함이 증명되었다(Freund & Schapire 1997). 이러한 성질은 AdaBoost 알고리즘이 자료의 압축문제에 적합한 방법으로 여겨지는 계기가 되었다.

이제 부스팅 알고리즘의 여러 해석들 중에 주요한 두 가지 해석에 대해서 살펴보도록 한다. 이러한 해석을 통해서 다양한 종류의 부스팅 알고리즘이 어떻게 개발되었는지를 설명한다.

1) 가파른 강하 알고리즘으로서의 AdaBoost

Schapire & Singer(1999)에 의해서 AdaBoost 알고리즘은 최적화에서 잘 알려진 가파른 강하(Steepest descent) 알고리즘으로 해석될 수 있다는 것이 밝혀졌다. $y_i \in \{-1,1\}$로 라벨이 주어진 이진 분류문제를 생각해볼 때, $\mathcal{F}$는 약한 학습기의 집합이라고 하자. 또한 AdaBoost의 최종 예측모형은 약한 학습기들의 선형결합으로 이루어져 있으므로,

$L(\mathcal{F})$는 $\mathcal{F}$상의 학습기들의 모든 선형결합이라고 하자. Schapire & Singer(1999)에 따르면 AdaBoost는 지수손실함수 $L(y,f)=\exp(-yf)$에 대한 경험위험 최소추정량

$$\hat{f}=\arg min_{f\in L(F)}\sum_{i=1}^{n}\exp(-y_i f(x_i))/n \qquad \text{(V-1)}$$

을 구하는 가파른 강하 알고리즘이다. 참고로 지수 손실함수는 AdaBoost 알고리즘을 통해서 소개된 손실함수로서 Fisher 일치성을 만족한다.

2) 기울기 강하 알고리즘으로서의 부스팅

Friedman(2001)은 부스팅 알고리즘을 최적화 알고리즘의 하나인 기울기 강하(Gradient descent) 알고리즘으로 해석하였으며, 이를 통하여 지수 손실함수 이외의 다양한 손실함수에서 부스팅 알고리즘을 개발하였다. 이러한 알고리즘을 그래디언트 부스팅(Gradient boosting)이라고 부른다.

먼저 기울기 강하 알고리즘에 대하여 소개하면, p차원 공간에서 정의된 미분 가능하고 볼록인 함수 $f(x)$의 최솟값을 찾는 문제를 고려해보자. 다음의 기울기 강하 알고리즘은 주어진 해로부터 기울기 값이 작은 쪽으로 현재의 해를 이동시키면서 축차적으로 최솟값을 찾는 방법이다.

1. 해를 $x^c=x_0$로 초기화 한다.
2. 다음 단계를 해 x^c가 수렴할 때까지 반복한다.
   (a) x^c에서 기울기를 다음과 같이 계산한다.
   $$\nabla=(\partial f(x)/\partial x_{1,\ldots,}\partial f(x)/\partial x_p)^T\big|_{x=x^c}$$
   (b) x^c의 이동거리 ρ을 계산한다.
   (c) x^c를 ∇방향으로 ρ만큼 이동하여 새로운 해를 구한다.
   $$x^c=x^c+\rho\nabla$$

위의 기울기 강하 알고리즘을 부스팅에 적용하면 다음과 같다. 주어진 손실함수 L과 주어진 함수집합 $\mathcal{F}$에 대해서 경험위험함수 $R(f)=\sum_{i=1}^{n}L(y,f(x_i))/n$을 최소화하려고 한다. 주어진 함수 f에서의 경험위험함수의 기울기는 $\nabla(f)=\dot{L}(y,f(x))$로 정의된다. 여기서 $\dot{L}(y,a)=\partial L(y,a)/\partial a$이다.

1. 해를 $f^c=f_0$로 초기화한다.
2. 다음 단계를 해 f^c가 수렴할 때까지 반복한다.
   (a) f^c에서 기울기 $\nabla(f)$를 계산한다.
   (b) $\nabla(f)$와 가장 가까운 기저 학습기(Base learner) g를 다음과 같이 찾는다.

$$g=\arg min_{h\in F}\sum_{i=1}^{n}(h(x_i)-\nabla(f)(x_i))^2$$

   (c) f^c의 이동거리 ρ를 계산한다.

$$\rho=\arg min_{z\in R}R(f^c+zg)$$

   (d) f^c를 g방향으로 ρ만큼 이동하여 새로운 해를 구한다.

$$f^c=f^c+\rho g$$

배깅과의 차이점으로 배깅은 분류기들이 상호 영향을 주지 않지만 부스팅은 이전 분류기의 학습 결과를 토대로 다음 분류기의 데이터의 샘플가중치를 조정한다는 것이 있다. 최초의 부스팅 알고리즘은 AdaBoost 알고리즘으로 가중선형결합 후 최종 분류기를 설정하는 알고리즘이다. AdaBoost 알고리즘은 초기에는 모두 동일한 확률로 복원추출을 하지만 매 반복마다 오분류된 관측치의 가중치는 증가시키고 정분류된 가중치는 감소시키면서 예측모형을 만들어 간다.

Friedman(2001)은 의사결정나무를 기본학습기로 하는 그래디언트 부스팅 알고리즘을 L_2 손실함수와 로지스틱 손실함수에 대해서 개발하였으며 이 두 알고리즘을 각각 L_2 부스팅과 로짓 부스팅으로 명명하였다. 예를 들어, L_2 부스팅의 손실함수는

$L(y,a)=(y-a)^2/2$이며 기울기 함수는 $\nabla(f)=-(y-f(x))$로서 음의 잔차가 된다. 따라서 L_2 부스팅은 현재의 해 f^c의 잔차인 $r_i=y_i-f^c(x_i)$를 가장 잘 설명하는 기저 학습기 g를 찾고 f^c를 g방향으로 ρ만큼 이동시킨다.

학습기의 복잡도가 더해지는 기본학습기의 수에 비례한다고 생각할 수 있으므로 그래디언트 부스팅은 너무 많이 반복하면 과대적합 문제가 발생할 수 있다. 과대적합을 피할 수 있는 방법으로는 반복수를 조절하는 것인데, 수렴할 때까지 반복하는 것이 아니라 일정수의 반복만 수행하는 것이다. 이때 반복수는 벌점모수가 되며 이를 적절히 조합함으로써 최적의 예측력을 갖는 모형을 찾을 수 있다.

과대적합을 피하기 위한 더 효율적인 방법으로는 축소추정 방법을 이용하는 것이다. 현재의 해 f^c를 기울기인 g방향으로 이동시킬 때 최적의 이동량인 ρ를 사용하지 않고 아주 작은 γ만큼만 이동함으로써 현재의 해를 $f^c=f^c+\gamma g$와 같이 갱신하는 것이다. 이때 γ는 벌점모수가 되며 보통 반복수를 아주 크게 놓고 γ를 조절하여 최종 학습기의 복잡도를 조절함으로써 과대적합을 피하게 된다. 축소추정을 이용한 부스팅 방법은 다음과 같다.

1. 해를 $f^c=f_0$로 초기화하고 아주 작은 $\gamma>0$를 선택한다.
2. 다음을 수렴할 때까지 반복한다.
   (a) f^c에서 기울기 $\nabla(f)$를 계산한다.
   (b) $\nabla(f)$에서 가장 가까운 기저 학습기 g를 다음과 같이 찾는다.

$$g=\mathrm{arg}min_{h\in F}\sum_{i=1}^{n}(h(x_i)-\nabla(f)(x_i))^2.$$

   (c) f^c를 g방향으로 γ만큼 이동하여 새로운 해를 구한다.

$$f^c=f^c+\gamma g$$

그래디언트 부스팅 알고리즘에서 기본학습기의 선택도 최종 학습기의 복잡도에 영향을 미친다. Friedman(2001)은 기본학습기의 선택이 함수에 대한 ANOVA 분해에서

교호작용 차수의 선택과 동일함을 간단한 의사결정나무로 설명하였다.

가장 간단한 의사결정나무인 그루터기를 기저 학습기로 사용하는 경우를 생각해 보면 주어진 그루터기의 구조는 분리에 사용된 변수와 분리 기준 값, 그리고 두 최종노드에서의 예측값으로 설명할 수 있다. 주어진 그루터기 $g(x)$는 다음 수식과 같이 나타낼 수 있다.

$$g(x) = a_l I(x_k \leq s) + a_r I(x_k > s). \tag{V-2}$$

여기서 x_k는 분리에 사용된 변수, s는 분리 기준값, a_l과 a_r는 두 최종노드에서의 예측값이다. 반복수가 M이고 축소추정 모수가 γ인 그래디언트 부스팅의 최종 학습기는 다음과 같이 나타낼 수 있다. 여기서 $x^{(m)}$은 m번째 기본학습기의 분기에 사용된 변수이다.

$$\begin{aligned} f(x) &= \sum_{m=1}^{M} \gamma g_m(x) \\ &= \sum_{m=1}^{M} \gamma\left(a_l^{(m)} I(x^{(m)} \leq s^{(m)}) + a_r I(x^{(m)} > s^{(m)})\right). \end{aligned} \tag{V-3}$$

따라서 최종학습기 f에서 변수 x_k가 미치는 영향은 다음과 같이 정리할 수 있다.

$$f_k(x) = \sum_{m=1}^{M} \gamma g_m(x) I(x^{(m)} = x_k) \tag{V-4}$$

$$f_k(x) = \sum_{k=1}^{M} f_k(x) \tag{V-5}$$

식 (V-4)이 식 (V-5)가 되는데, 이러한 모형은 일반화 가법모형(Hastie & Tibshirani 1990)이다. 일반화 가법모형은 고차원 함수의 추정에 많이 사용되는 모형으로서 교호작용은 존재하지 않으며 각 변수들을 적당한 비선형함수들의 가법모형으로 표현된다. x_k가 예측에 미치는 영향은 성분(Component) f_k를 그려봄으로써 시각적으로 확인할 수 있다.

다. 랜덤 포레스트(Random Forest)

랜덤 포레스트는 의사결정나무의 분산이 크다는 특징을 감안하여 배깅과 부스팅보다 더 많은 무작위성(Random)을 주어 약한 학습기들을 생성한 후 이를 선형결합하여 최종 학습기를 만드는 기법이다. 랜덤 포레스트에 대한 이론적 설명이나 최종 결과에 대한 해석은 어렵다는 단점이 있지만 예측력은 매우 높은 방법으로 알려져 있다. 특히 입력변수의 개수가 많을 때에는 배깅이나 부스팅과 비슷하거나 더 좋은 예측력을 보이는 경우가 많고, 조율모수가 없어서 실제 자료분석에 쉽게 사용될 수 있다.

랜덤 포레스트는 무작위성을 최대로 주기 위하여 붓스트랩과 더불어 입력변수들에 대한 무작위 추출을 결합하기 때문에 연관성이 약한 학습기를 여러 개 만들어 내는 기법이라 할 수 있다.

1. 훈련자료 $\mathcal{L} = \{y_i, x_i\}_{i=1}^{n}, x_i \in \mathbb{R}$에 대하여 n개를 자료를 이용한 붓스트랩 표본 $\mathcal{L}^* = \{y_i^*, x_i^*\}_{i=1}^{n}$을 생성한다.

2. $\mathcal{L}^*$에서 입력변수들 중 $k(k < p)$개만 무작위로 뽑아 의사결정나무를 생성한다. 이때 의사결정나무는 정해놓은 s단계까지 진행한다.

3, 이렇게 생성된 의사결정나무들을 선형결합하여 최종학습기를 만든다.

여기에서 붓스트랩 표본을 몇 개나 생성할 것인지, k/d 값을 어떻게 할 것인지, 선형결합의 형식을 어떻게 할 것인지에 대한 여러 가지 선택이 있다. 보통 붓스트랩 표본의 개수는 지나치게 적어서는 곤란하며, 선형결합의 형식은 각각의 의사결정나무들의 결과들에 대하여 회귀분석에서는 평균, 분류문제에서는 다수결원칙을 적용하는 방식이 많이 사용된다.

랜덤 포레스트의 이론적 배경을 분류문제에서 살펴보자. 일반적으로 분류문제에 있어서는 0-1 손실함수를 이용하지만, 랜덤 포레스트에 대한 이론적 설명에는 마진함수(Margin function)와 그에 기반한 랜덤 포레스트의 강도(Strength)를 이용한다.

먼저 랜덤 포레스트에 사용된 의사결정나무이 모집단을 $\mathcal{F}(\theta)$로, 이 모집단의 의사결정나무는 $\{f(\bullet, \Theta_k), k=1,...,K\}$로 표기하기로 한다. 만약 (x,y)가 (X, Y)의 분포로부터 생성된 분포이고 각각의 분류함수 $f(\bullet, \Theta_k)$가 투표방식으로 분류를 한다면,

$$\frac{1}{K}\sum_{k=1}^{K} I(f(x,\Theta_k)=y) - \mathrm{arg}max\left(\frac{1}{K}\sum_{k=1}^{K} I(f(x,\Theta_k=l)\right) < 0 \qquad \text{(V-6)}$$

이면 잘못된 분류를 하게 될 것이다. 위의 사실을 이용하여 랜덤 포레스트 $\mathcal{F}(\Theta)$의 마진함수, 예측오차, 강도를 다음과 같이 정의할 수 있다.

$$\begin{aligned} M_{F(\Theta)}(x,y) &= P_\Theta(f(x,\Theta)=y) - \max_{l\neq y} P_\Theta(f(x,\Theta)=l), \qquad \text{(V-7)} \\ R_{F(\Theta)} &= P_{X,Y}(M_{F(\Theta)}(X,Y)<0), \\ S_{F(\Theta)} &= E_{X,Y}[M_{F(\Theta)}(X,Y)]. \end{aligned}$$

$R_{F(\Theta)}$은 결국 0-1 손실함수를 이용한 예측오차와 동일하며, $S_{F(\Theta)}$은 분류함수와 과연 얼마나 큰 마진으로 분류하는가를 나타낸다. 예를 들면, 실제 y가 1일 때 로지스틱 회귀분석에 의한 0과 1에 대한 확률 추정값이 각각 0.45, 0.55라고 하자. 더 높은 확률을 주는 쪽으로 분류한다면 0-1 손실은 0이며 마진은 0.1이 된다. 반대로 0과 1에 대한 확률 추정값이 각각 0.1, 0.9이면 0-1 손실은 0이며 마진은 0.8이 된다. 이 두 개의 분류함수는 0-1 손실은 같으나 두 번째 분류함수가 더 좋은 분류성능을 갖는다고 생각할 수 있으며, 이를 측정한 것이 강도이다. 즉, 분류에 있어서 어느 정도 명확하게 분류하는가를 고려한 것이 $S_{F(\Theta)}$라고 할 수 있다. $S_{F(\Theta)}$ 값은 [-1, 1]에 속하며 그 값이 클수록 분류를 잘한다고 할 수 있으나 분류 자체가 분류경계 근처에서만 이루어지는 경우 $S_{F(\Theta)}$의 값은 작을 수 있다. 따라서 분류를 잘한다고 해서 반드시 $S_{F(\Theta)}$가 큰 것은 아니다.

Breiman(2001)은 랜덤 포레스트의 성능에 대하여 다음과 같이 이론을 설명하였다.

$$r(\theta,x,y) = I(f(x,\theta)=y) - I(f(x,\theta)=\arg max_{l\neq y} P_{\Theta}(f(x,\Theta)=l)), \quad \text{(V-8)}$$
$$\sigma^2(\theta) = Var_{X,Y}(r(\theta,X,Y)),$$
$$\rho(\theta,\theta') = Corr_{X,Y}(r(\theta,X,Y), r(\theta',X,Y))$$

라 하자. 아래의 정리는 앞에서 정의한 개념들에 기반하여 $R_{F(\Theta)}$의 상한을 구한 것이다.

〈정리 3〉 만약 $S_{F(\Theta)} > 0$이면

$$R_{F(\Theta)} \leq \rho\left(\frac{1-S_{F(\Theta)}^2}{S_{F(\Theta)}^2}\right). \quad \text{(V-9)}$$

여기서, ρ는 독립적으로 Θ의 분포를 따르는 Θ_1, Θ_2에 대하여

$$\rho = \frac{E_{\Theta_1,\Theta_2}\{\rho(\Theta_1,\Theta_2)\sigma(\Theta_1)\sigma(\Theta_2)\}}{E_{\Theta_1,\Theta_2}\{\sigma(\Theta_1),\sigma(\Theta_2)\}}$$

로 정의된다.

$Proof.$ $g(x,y) = M_{F(\Theta)}(x,y)$로 놓으면

$$R_{F(\Theta)} = P_{X,Y}(g(X,Y) - S_{F(\Theta)} < -S_{F(\Theta)}) \quad \text{(V-10)}$$
$$\leq \frac{Var_{X,Y}(g(X,Y))}{S_{F(\Theta)}^2}$$

로 표현된다.

여기서 $Var(g(X,Y))$의 상환은 다음과 같이 구할 수 있다.

$$\begin{aligned} Var_{X,Y}(g(X,Y)) &= E_{X,Y}\{E_{\Theta}(r(\Theta,X,Y))\}^2 - \{E_{X,Y}E_{\Theta}(r(\Theta,X,Y))\}^2 \\ &= E_{X,Y}E_{\Theta_1,\Theta_2}(r(\Theta_1,X,Y)r(\Theta_2,X,Y)) \\ &\quad - \{E_{\Theta}E_{X,Y}(r(\Theta,X,Y))\}^2 \\ &= E_{\Theta_1,\Theta_2}[E_{X,Y}(r(\Theta_1,X,Y)r(\Theta_2,X,Y))] \\ &\quad - E_{\Theta_1,\Theta_2}[E_{X,Y}(r(\Theta_1,X,Y))E_{X,Y}(r(\Theta_2,X,Y))] \\ &= E_{\Theta_1,\Theta_2}[Cov_{X,Y}(r(\Theta_1,X,Y)),r(\Theta_2,X,Y)] \\ &= \rho \end{aligned}$$

여기서, Θ_1, Θ_2는 Θ와 분포가 같으면서 서로 독립이다. 또한

$$\begin{aligned} E_{\Theta}\sigma^2(\Theta) &= E_{\Theta}E_{X,Y}\{r^2(\Theta,X,Y)\} - E_{\Theta}\{E_{X,Y}r(\Theta,X,Y)\}^2 \\ &\le E_{\Theta}E_{X,Y}\{r^2(\Theta,X,Y)\} - S^2_{F(\Theta)} \\ &\le 1 - S^2_{F(\Theta)} \end{aligned}$$

이므로

$$Var_{X,Y}(g(X,Y)) \le \rho(1 - S^2_{F(\Theta)})$$

이 성립하며 식 (V-10)으로부터 식 (V-9)가 성립함을 보일 수 있다.

식 (V-9)로부터 $R_{F(\Theta)}$는 $S_{F(\Theta)}$ 값이 크고 ρ가 작을 때 작은 값을 갖는다는 것을 알 수 있다. 특히 ρ가 작다는 것은 랜덤 포레스트에서 생성된 의사결정나무들 간에 독립성이 강하다는 것으로 생각할 수 있다. 따라서 이 경우에 랜덤 포레스트에서 붓스트랩 표본을 뽑고 임의로 입력변수들을 선택함으로써 예측오차가 줄어드는 것이 설명된다.

1) 부분 의존성 도표

일반적인 회귀분석의 경우에는 예측에 사용되는 입력변수가 2개 이내이면 적절한 그래프를 통해 입력변수의 출력변수에 대한 영향이나 중요도를 쉽게 확인할 수 있다.

그러나 변수의 개수가 많고 의사결정나무, 부스팅, 랜덤 포레스트와 같이 예측방법이 단순하지 않을 경우에는 각 변수들의 영향력을 시각화하기가 쉽지 않다. 부분 의존성 도표(Partial dependence plot)는 Friedman(2001)에 의해 제안된 것으로서 학습기에 사용되는 입력변수들 중 일부를 골라내 그것의 영향력을 시각화하는 방법이다.

$x \in \mathbb{R}$인 입력변수를 이용하는 학습기 $\hat{F}(x)$를 생각해보자. $l(l < p)$에 대하여 Π_l 를 크기가 l인 {1,...,p}의 부분집합이라 하고, Π_{-l}={1,...,p}$- z_l$이라 하자. $z_l = \{x_k : k \in \pi_l\}$ 이고 $z_{i,l} = \{x_{ik}, k \in \pi_l\}$이라 하면, 이 때 z_l의 영향력을 나타내는 값으로 다음을 고려한다.

$$\frac{1}{n}\sum_{i=1}^{n} \hat{F}(z_l, z_{i,-l}). \tag{V-11}$$

위의 값은 z_{-l}의 확률밀도함수 $p_{-l}(z_{-l})$에 대하여

$$\hat{F}_l(z_l) = \int \hat{F}(z_l, z_{-l}) p_{-l}(z_{-l}) dz_{-l} \tag{V-12}$$

로 표현된다. 이 값을 이용해 시각화하는 것이 부분 의존성 도표로서 $l = 1$인 경우에는 2차원 그래프, $l = 2$이면 3차원 그래프로 표현할 수 있다. $\hat{F}(\bullet)$이 가법모형이면 $\hat{F}(x) = \sum_{i=1}^{p} \hat{F}_i(x_i)$이므로

$$\frac{1}{n}\sum_{i=1}^{n} \hat{F}(z_l, z_{i,-l}) = \hat{F}(z_l) \tag{V-13}$$

인 특성을 가지고 있다.

다음 모형은 Friedman(2001)의 모의실험을 조금 수정한 것이다.

$$F^*(x_i) = \sum_{l=1}^{20} a_l g(z_i, l) + \varepsilon_i \tag{V-14}$$

여기에서 ϵ_i는 서로 독립이며 $N(0, 2^2)$을 따르고, $a_l \sim U[-1, 1]$이고, $z_{i,l}$은

$N(0,4I_{10})$을 따르는 10개의 입력변수들 중에서 평균 2인 지수난수 r에 대하여 $n_l = [1.5 + r]$ 개를 임의로 뽑은 것이다. 그리고 $g_l(z_l)$의 구체적인 함수형태는 다음과 같다.

$$g_l(z_l) = \exp\left(-\frac{1}{2}(z_l - \mu_l)^T V_l (z_l - \mu_l)\right) \quad \text{(V-15)}$$

이며, 여기서 μ_l은 입력변수들과 같은 분포를 따르고, $V_l = U_l D_l U_l^T$로 U_l은 임의의 직교행렬이며 $D_l = diag(d_{1l},...,d_{nl},l)$, $\sqrt{djl} \sim U[0.1,0.9]$이다. 이렇게 생성된 자료는 가법모형보다 복잡하며 2차 이상의 교호작용도 포함하게 된다.

R의 랜덤 포레스트에서는 변수에 대한 중요도 지수를 제공한다. 특정 변수에 대한 중요도 지수는 그 변수를 포함하지 않는 경우에 어느 정도 예측오차가 줄어드는지를 보여주는 것이다.

2. 신경망모형[4)]

생물학적 신경망의 구조에 착안하여 학습 알고리즘으로 개발된 인공신경망모형(Neural network)은 복잡한 구조로 이뤄진 데이터의 예측 문제를 해결하기 위해 주로 사용되는 비선형모형이다. 컴퓨터 성능이 개선되면서 다층신경망과 역전파(Back propagation) 알고리즘이 합쳐지면서 신경망모형에 대한 응용된 분야들이 크게 확장되어 왔다. 하지만 방법론이 수학적으로 정교할지라도 실무자에게는 결과가 어떻게 나오는지에 대한 소통가능성과 그 결과를 바탕으로 보험회사의 정책수립에 반영할 수 있는지가 중요하다. 신경망이란 것은 매우 높은 예측력을 보이나 그에 반해 해석에 대한 어려움이 존재하여 해석이 중요한 분야 중에 하나인 신용평가에서는 잘 사용되지 않지만 음성인식 등에는 응용되어 사용되고 있다. 본 절에서는 인공신경망모형의 작동원리에 초점을 맞추도록 한다.

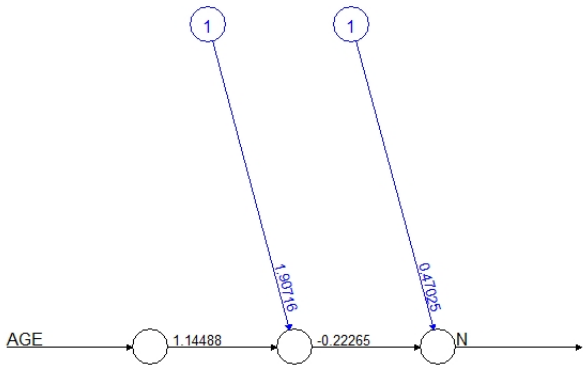
4) 신경망모형 방법론은 Hastie, Tibshirani & Friedman(2008)과 박창이·김용대·김진석·송종우·최호식(2011)을 참고하여 요약 및 정리하였음

신경망모형은 뉴런들을 서로 연결하여 입력한 값에 대하여 가장 최적의 결과값을 예측하는 것이 기본적인 작동원리이다. 생물학적 신경망은 시냅스가 모여서 전자가 오면 처음에는 졸고 있다가 일정 수준 이상으로 오게 되면 활성화되어 입력된 전자들을 깨우는 역할만 하고, 새로운 것을 만들어서 내보낸다. 이와 유사하게 인공신경망모형은 입력변수를 로짓변환하여 0과 1사이에 값을 가지면 은닉노드를 활성화시키고, 은닉노드에서 출력함수를 통하여 출력변수를 생성한다. 신경망은 통계적인 관점에서 보면 입력변수들이 선형적으로 결합되어 있는 것에 비선형 함수를 취하는 사영추적회귀(Projection pursuit regression)로 볼 수 있다.

가. 단층 신경망모형 구조

〈그림 Ⅴ-1〉은 연령을 입력변수로 하고 발생건수를 출력변수로 이용한 가장 간단한 신경망모형 구조를 보여준다. 신경망의 구성은 입력층(Input layer), 은닉층(Hidden layer), 출력층(Output layer)으로 되어 있다. 아래 그림의 입력층은 각 입력변수에 대응되는 노드로 구성되며 노드의 수는 입력변수의 개수와 같다. 아래 그림에서 1이라고 적혀있는 노드는 상수항을 가리킨다. 입력층으로부터 전달된 변수값들의 선형적 결합을 통해 은닉층에서는 비선형함수로 처리하고, 출력층이나 딥러닝의 경우는 다른 은닉층에 전달하는 역할을 한다. 출력층은 출력변수에 대응되는 노드로서 발생건수를 출력변수로 사용한다. 분류모형인 경우에는 클래스의 수 만큼의 출력노드가 생성된다. 신경망모형은 주로 한 방향으로 진행되는 feedforward 형태이다.

〈그림 V-1〉 신경망모형 간단한 구조 예시



위의 그림을 식을 통해 나타내면 다음과 같다. 은닉노드의 값은 입력노드의 선형결합으로 이루어진다.

$$z_1 = \sigma(1.90716 - 1.14488x). \tag{V-16}$$

$\sigma(\bullet)$는 시그모이드 함수(Sigmoid function)로 다시 쓰면 다음의 식과 같고 0에서 1 사이의 값을 갖는다.

$$z_1 = \frac{1}{1+e^{-(1.90716+1.14488x)}}. \tag{V-17}$$

출력값은 은닉노드들의 선형결합 $t_1 = 0.32804 - 3.2884z_1$의 함수로 모형화한다. $g(\bullet)$은 항등함수로 쓰이는 경우가 많다.

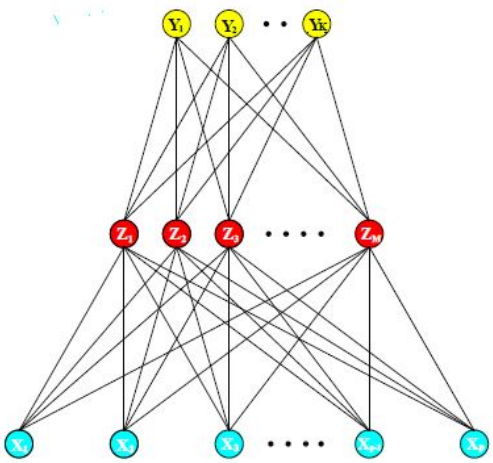
$$f_1(x) = g_1(t_1) = 0.47025 - 0.22265z_1. \tag{V-18}$$

나. 다층신경망모형(딥러닝) 구조

위에서는 이해를 돕기 위해 가장 간단한 신경망모형의 식을 살펴보았다면 여기서는

일반화된 식을 통해 클래스의 수가 K인 분류 문제를 바탕으로 모형을 살펴본다. 출력노드 $k(=1,...,K)$에서는 자료가 k번째 클래스에 속할 때는 출력변수가 1이고 나머지는 0으로 코딩하는 방식으로 클래스 k에 속할 확률을 모형화한다. 회귀문제는 $K=1$인 경우에 해당된다. 즉, 출력노드가 하나라는 뜻이다.

〈그림 V-2〉 신경망모형 구조



출처: The elements of statistical learning, p.394

은닉노드 값 z_m은 입력노드 값들의 선형결합이고 출력값은 z_m들의 선형결합 t_k들의 함수로 다음과 같이 모형화한다.

$$z_m = \sigma(\alpha_{0m} + \alpha_m^T x), m = 1,...,M. \quad \text{(V-19)}$$

$$t_k = \beta_{0k} + \beta_k^T z, k = 1,...,K. \quad \text{(V-20)}$$

$$f_k(x) = g_k(t), k = 1,...,K. \quad \text{(V-21)}$$

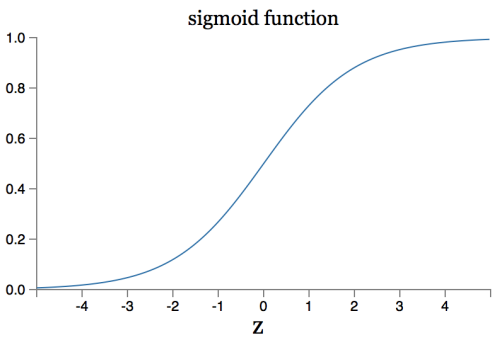
여기서 $z = (z_1,...,z_M)^T$이고 $t = (t_1,...,t_K)^T$이다.

여기서 $\sigma(\bullet)$는 활성함수(Activation function)이고, 시그모이드(Sigmoid) 함수를 주로 사용한다. 원래는 노드의 활성화 유무를 표시하기 위해 step function을 사용하

려하였으나 미분이 불가능하여 그와 비슷한 시그모이드 함수를 사용한다. 시그모이드 함수는 단극성과 양극성으로 나뉘며, 단극성 시그모이드 함수는 다음과 같이 정의되는 증가함수로 0과 1 사이의 값을 갖는다.

$$\sigma(v) = \frac{1}{1+e^{-v}}. \tag{V-22}$$

〈그림 V-3〉 시그모이드 함수



양극성 시그모이드 함수는 다음과 같이 정의되며 출력값은 -1과 1 사이의 값을 갖으며, $\sigma(0) = 0$이다.

$$\sigma(v) = \frac{1-e^{-v}}{1+e^{-v}}. \tag{V-23}$$

그리고 RBF(Radial Basis Function)($\sigma(v) = \exp(-v^2/2)$)를 활성함수로 사용하는 경우에는 RBF 신경망이라고 한다.

$g_k(t)$는 출력함수(Output function)이고, 이 출력함수는 출력값 t를 최종적인 비선형으로 변환해주는 역할을 하는 함수이다. 회귀에서는 $g_k(t) = t_k$인 항등함수(Identity function)가 사용되고 K-클래스 분류에서는 softmax 함수가 다음과 같이 사용된다. softmax 함수는 항상 양의 값을 갖고, 합이 1이 되며, 다범주 로지스틱회귀에서도 많이 활용되고 있다.

$$g_k(t) = e^{t_k} / \sum_{l=1}^{K} e^{t_l}. \quad \text{(V-24)}$$

다. 다층신경망의 적합

신경망모형은 미지의 가중치(Weights) 모수들로 구성되어 있다.

$\{\alpha_{0m}, \alpha_m, m = 1, ..., M\}$ $M(p+1)$개와 $\{\beta_{0k}, \beta_k, k = 1, ..., K\}$ $K(M+1)$개의 가중치를 θ로 나타낸다. 회귀문제에서는 식 (V-10)과 같이 비용함수로서 오차제곱합을 사용하고, 분류문제에서는 비용함수로서 오차제곱합이나 deviance를 사용하며 $G(x) = \arg max_k f_k(x)$를 이용하여 분류한다.

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} (y_{ik} - f_k(x_i))^2 \quad \text{(V-25)}$$

$$R(\theta) = -\sum_{k=1}^{K} \sum_{i=1}^{n} y_{ik} \log f_k(x_i). \quad \text{(V-26)}$$

$R(\theta)$의 비선형성 특성으로 인해 전역 최솟값(Global Minimizer)을 구하는 것은 불가능하기 때문에 과대적합을 불러일으킬 수 있다. 따라서 전역 최솟값을 구하는 것 대신에 좋은 국소 최솟값을 구하기 위해 직접적인 벌점화나 알고리즘의 조기 종료(Early stopping) 등의 간접적인 벌점화를 결합하는 방법을 사용한다.

신경망에서는 $R(\theta)$를 최소화하는 θ를 찾기 힘든 경우에 사용하는 대표적인 반복 알고리즘으로 기울기 강하(Gradient descent) 알고리즘을 적용하는 역전파를 사용한다. 모형의 구성으로 인해 기울기를 chain rule을 통한 미분으로 쉽게 얻을 수 있다. 다음은 오차제곱합을 비용함수로 사용하는 경우에 대한 역전파 알고리즘이다.

$z_{mi} = \sigma(\alpha_{0m} + \alpha_m^T x_i)$와 $z_i = (z_{1i}, ..., z_{Mi})^T$라 하자.

그러면 $R(\theta)=\sum_{i=1}^{n}R_i=\sum_{i=1}^{n}\sum_{k=1}^{K}(y_{ik}-f_k(x_i))^2$이고 편도함수는 다음과 같이 주어진다.

$$\frac{\partial R_i}{\partial \beta_{km}}=-2(y_{ik}-f_k(x_i))g'_k(\beta_k^T z_i)z_{mi}, \tag{V-27}$$
$$\frac{\partial R_i}{\partial \alpha_{ml}}=-2\sum_{k=1}^{K}(y_{ik}-f_k(x_i))g'_k(\beta_k^T z_i)\beta_{km}\sigma'(\alpha_m^T x_i)x_{il}.$$

r번째 반복의 값이 주어지면 다음과 같이 $(r+1)$번째 업데이트 값을 조정한다.

$$\beta_{km}^{(r+1)}=\beta_{km}^{(r)}-\gamma_r\sum_{i=1}^{n}\frac{\partial R_i}{\partial \beta_{km}^{(r)}}, \tag{V-28}$$
$$\alpha_{ml}^{(r+1)}=\alpha_{km}^{(r)}-\gamma_r\sum_{i=1}^{n}\frac{\partial R_i}{\partial \alpha_{ml}}.$$

여기서 γ_r은 학습률(Learning rate)이라고 부른다.

식 (V-13)에서 $\frac{\partial R_i}{\partial \beta_{km}}=\delta_{ki}z_{mi}$와 $\frac{\partial R_i}{\partial \alpha_{km}}=s_{mi}x_{il}$라 하면, δ_{ki}와 s_{mi}는 각각 입력층과 은닉층에서의 현재 모형의 오차로 볼 수 있다. 이러한 오차들은 다음의 역전파 등식을 만족시킨다.

$$s_{mi}=\sigma'(\alpha_m^T x_i)\sum_{k=1}^{K}\beta_{km}\delta_{ki}. \tag{V-29}$$

식 (V-14)를 이용하면 식 (V-13)의 업데이트를 다음과 같이 구현할 수 있다. 전방 패스(Forward pass)에서는 주어진 가중값에 대하여 모형으로부터 예측값 $\hat{f}_k(x_i)$를 계산한다. 후방 패스(Backward pass)는 오차 δ_{ki}를 계산하고 식 (V-13)을 이용하여 역전파시켜서 오차 s_{mi}를 계산한다. 두 오차는 업데이트를 위해 기울기를 계산하기 위해 쓰인다. 이러한 알고리즘은 역전파 알고리즘 또는 델타 규칙(Delta rule)이라 한다. deviance를 목적함수로 사용할 때 역전파 알고리즘에서도 이와 같은 방법으로 유도하면 된다.

라. 신경망모형 구축 시 고려사항

1) 초기값

역전파 알고리즘의 초기값 결과에 영향을 많이 주므로 초기값의 선택은 매우 중요한 문제이다. 보통은 0 근처에서 초기값이 무작위하게 선택되기 때문에 초기 모형은 선형모형에 가깝지만 가중치 값이 증가하면서 비선형모형이 된다. 초기값이 0과 정확하게 일치할 경우에는 반복에 따라 결과값이 전혀 변하지 않는다. 그러나 초기값이 너무 큰 값에서부터 시작하게 되면 좋지 않은 해를 얻을 수도 있다는 문제점이 있음에 유의해야겠다.

2) 과대적합 문제

신경망에서는 많은 가중치를 추정해야 하기 때문에 그만큼 과대적합 문제가 빈번히 이슈가 된다. 과대적합을 피하기 위해서는 알고리즘의 조기종료와 가중치 감소기법 등이 주로 이용된다.

첫째, 조기종료라는 것은 모형을 적합하는 과정에서 검증오차가 증가하게 되면 반복을 중지하는 것이다. 앞서 설명한 것과 같이 초기값은 선형모형에 가까운 형태이기 때문에 이러한 조기종료는 최종모형을 선형모형으로 축소시키는 효과가 있다. 둘째, 가중치 감소기법은 선형모형의 능형회귀와 유사한 벌점화 기법이다. 이 기법은 벌점화된 목적함수 식 (V-30)을 최소화한다. 여기서 식 (V-31)을 사용하기도 한다.

$$R(\theta)+\lambda J(\theta) \tag{V-30}$$

$$J(\theta)=\sum_{k,m}\frac{1+\beta_{km}^2}{\beta_{km}^2}+\sum_{m,l}\frac{1+\alpha_{m,l}^2}{\alpha_{m,l}^2} \tag{V-31}$$

가중치 제거 방법은 가중치 감소보다 작은 계수값들을 더욱 줄여주는 효과가 있다.

3) 입력변수

첫째, 신경망모형은 모형 자체가 복잡하기 때문에 입력자료의 선택에 매우 민감하게 반응한다. 신경망모형에 적합한 자료는 다음과 같다. 입력변수가 범주형일 경우는 모든 범주에서 일정 빈도 이상의 값을 갖는 자료이고, 연속형일 경우는 변수들 간에 값들의 범위가 큰 차이가 없는 자료이다. 그리고 입력변수의 수가 너무 적거나 많지 않고, 범주형 출력값의 각 범주의 빈도가 비슷한 자료이다.

둘째, 신경망모형에서 고려할 사항은 연속형 입력변수의 변환 또는 범주화이다. 연속형 변수는 분포가 평균을 중심으로 비대칭일 경우에는 결과가 좋지 않을 수 있다. 예를 들어, 사고금액 분포는 일반적으로는 대부분의 계약자 사고금액들이 평균 미만이고, 일부 특정한 계약자의 사고금액은 매우 큰 패턴을 보이기도 한다. 따라서 이러한 분포를 보이는 변수의 경우에는 분포가 평균을 중심으로 대칭이 되도록 로그 변환 등을 고려해 볼 수 있다. 그리고 또 다른 방법으로는 연속형 변수를 범주화하는 방법도 있다.

셋째, 새로운 변수의 생성이다. 때로는 최초의 입력변수들을 그대로 사용하는 대신 조합하여 새로운 변수를 생성하여 입력변수로 사용할 경우 아주 좋은 결과를 얻을 수 있다. 예를 들면, 고객의 수입, 학력 등을 입력변수로 그대로 사용하지 않고 이러한 변수들을 이용하여 구매지수를 만든 후에 이 구매지수를 입력변수로 사용하여 특정한 상품의 구매여부를 예측해 볼 수 있다.

마지막으로 범주형 입력변수의 가변수화이다. 회귀분석에서와 같이 신경망에서도 범주형 변수는 가변수로 만들어 사용한다. 회귀분석과의 차이점은 신경망모형에서는 가변수로 설정하는 방법에 따라서 결과가 민감하게 반응한다는 점에 유의해야 한다. 예를 들어, 남자와 여자를 각각 0과 1로 가변수화 하는 것은 각각 -1과 1로 가변수화 하는 것과 그 결과가 예상과 많이 달라질 수 있다는 것이다. 따라서 일반적으로 모든 범주형 변수를 가변수화 할 때는 같은 범위를 갖도록 하는 것이 바람직하다.

4) 은닉층과 은닉노드의 수

신경망을 적용할 때 직면하는 주요한 문제들 중 하나는 모형 선택, 다시 말해 은닉층의 수와 은닉노드의 수를 결정하는 것이다. 은닉층과 은닉노드의 수가 너무 많아지면 추정할 모수인 가중치들이 너무 많아지기 때문에 과대적합 문제가 발생할 수 있다. 그러나 반대로 너무 적으면 과소적합의 문제가 발생할 수도 있다. 은닉층이 하나인 신경망은 범용 근사자(Universal Approximation)이다. 그러므로 일반적으로 신경망모형을 적용할 때는 은닉층을 하나로 하고, 은닉 노드수를 적절히 선택하게 되면 큰 문제가 발생하지 않을 것이다. 둘째 은닉노드의 수는 교차확인오차를 사용하여 결정하는 방법보다는 적절히 큰 값으로 놓은 후에 가중치 감소(Weight decay)라는 모수에 대한 벌점화를 적용하는 것이 좋다.

5) 다중 최솟값

신경망에서는 일반적으로 비용함수 $R(\theta)$는 비볼록함수이고 여러 개의 국소 최솟값들(Local minima)을 가진다. 따라서 무작위로 선택된 여러 개의 초기 값들에 대하여 신경망을 적합한 후 얻은 해들을 비교하여 가장 오차가 작은 것을 선택하여 최종 예측치를 얻거나 예측값의 평균이나 최빈값을 구한 후에 최종으로 예측치를 선택하는 방법을 고려해 볼 수 있다. 그리고 또 다른 방법으로는 훈련자료에 대하여 신경망을 기저학습법으로 사용하는 배깅(Bagging)을 적용하는 방법이 있다.

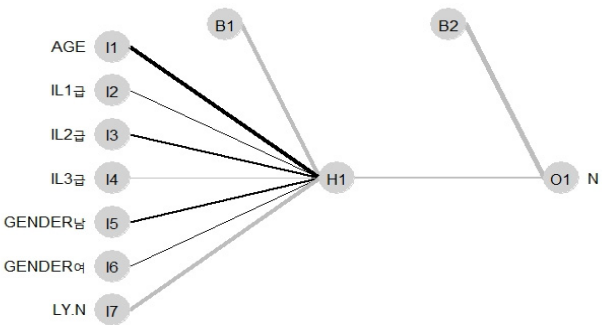
3. 실손의료보험 자료를 활용한 신경망모형 분석

가. 신경망모형

1) 빈도

연령, 상해급수, 성별 그리고 직전연도 발생건수를 독립변수, 즉 입력변수로 하고 16년도 발생건수를 출력변수(Output)로 모델링한다. 신경망모형은 앞서 한 다른 모델링과 다르게 성별과 상해급수와 같은 입력변수는 가변수화 하여 넣는다.

〈그림 V-4〉 빈도 신경망모형 은닉노드=1개

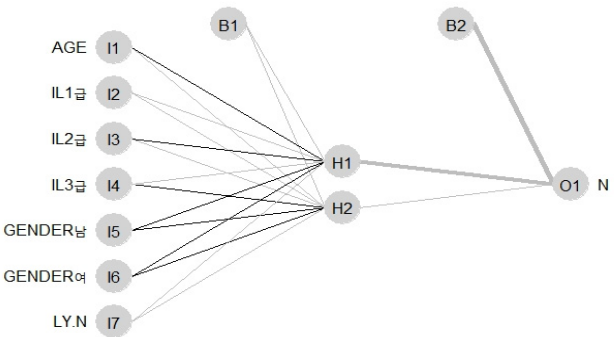


입력변수들의 가중치들은 선형결합하여 은닉노드와 연결된다. 식 (V-32)에서 은닉노드는 시그모이드 함수로 인해 0에서 1 사이의 값을 갖는다. 다시 z_1 값은 식 (V-33) 값에 대입하여 출력변수를 산출한다. 연령 변수의 계수값이 크고 또한 연령 입력변수의 값이 크므로 z_1은 1에 가까운 값을 갖는다. 이는 은닉노드가 활성화되었다고 볼 수 있다.

〈표 V-1〉 빈도 신경망모형 가중치 은닉노드=1개

시작점	종료점	가중치
B1	H1	-0.6344
I1	H1	0.6485
I2	H1	0.0932
I3	H1	0.2183
I4	H1	-0.0048
I5	H1	0.3300
I6	H1	0.1634
I7	H1	-0.6946
B2	O1	-0.7759
H1	O1	-0.3356

〈그림 V-5〉 빈도 신경망모형 은닉노드=2개



$$z_1 = \sigma(-0.6344 + 0.6485x_1 + \cdots - 0.6946x_7). \quad \text{(V-32)}$$

$$t_1 = -0.7759 - 0.3356z_1. \quad \text{(V-33)}$$

신경망모형은 모델링할 때 가중치들의 초기값을 설정해 줄 수도 있고 혹은 랜덤하게 할 수도 있다. 랜덤하게 나온 모형과 동일한 모형을 나오게 하려면 set.seed( )함수를 모델링하기 전에 써주면 된다.

〈표 V-2〉 빈도 신경망모형 가중치 은닉노드=2개

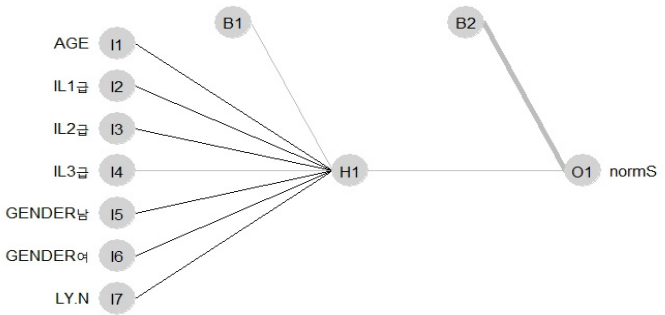
시작점	종료점	가중치	시작점	종료점	가중치
B1	H1	-0.1202	B1	H2	-0.6064
I1	H1	0.1648	I1	H2	-0.3030
I2	H1	-0.4940	I2	H2	-0.2408
I3	H1	0.5561	I3	H2	-0.1911
I4	H1	-0.5261	I4	H2	0.6420
I5	H1	0.6778	I5	H2	0.1245
I6	H1	0.1751	I6	H2	0.2736
I7	H1	-0.2249	I7	H2	-0.4935
-	-	-	B2	O1	-832.9959
-	-	-	H1	O1	-832.7232
-	-	-	H2	O1	-0.0340

〈그림 V-5〉는 은닉노드가 두 개인 신경망모형이다. 은닉노드가 〈그림V-4〉에 비해 하나 추가되었음에도 불구하고 복잡한 모형을 띈다.

2) 심도

신경망모형은 입력변수에 민감하게 반응하기 때문에 연속형일 경우는 변수들 간에 차이가 많이 없는 것을 사용해야 한다. 본 연구에서 사용하고 있는 심도의 데이터는 빈도의 데이터에 비해 범위가 넓다. 따라서 자료를 그대로 넣으면 모형이 수렴하지 않으므로 정규화과정을 거치는 것이 바람직하다.

〈그림 Ⅴ-6〉 심도 신경망모형 은닉노드=1개

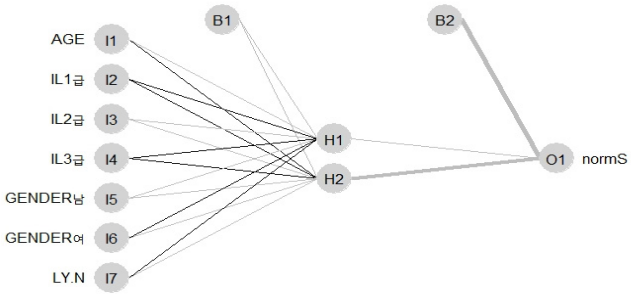


〈표 Ⅴ-3〉 심도 신경망모형 가중치 은닉노드=1개

시작점	종료점	가중치
B1	H1	-0.3008
I1	H1	-0.6138
I2	H1	-0.4094
I3	H1	-0.4986
I4	H1	-0.2651
I5	H1	0.3421
I6	H1	0.0010
I7	H1	-0.3332
B2	O1	6231.6896
H1	O1	0.4335

심도 데이터는 자료의 최솟값을 빼주고 자료의 최댓값과 최솟값의 차이로 나누어 줘서 정규화하는 방법으로 모델링한다. 입력변수는 빈도와 마찬가지로 가변수화하여 넣어 준다.

〈그림 Ⅴ-7〉 심도 신경망모형 은닉노드=2개



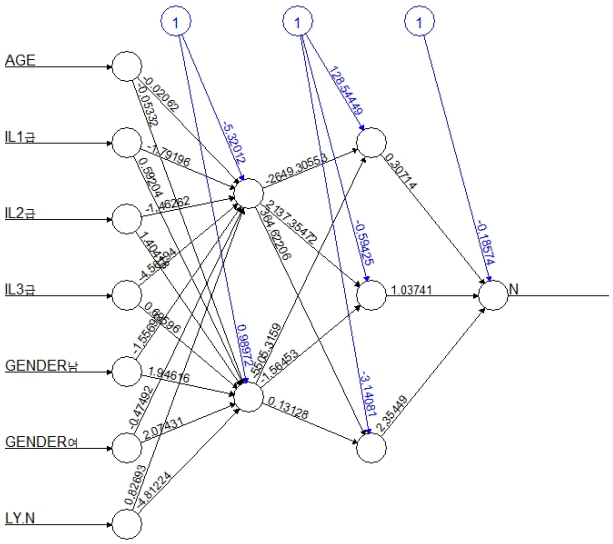
〈표 Ⅴ-4〉 심도 신경망모형 가중치 은닉노드=2개

시작점	종료점	가중치	시작점	종료점	가중치
B1	H1	0.6605	B1	H2	-0.5081
I1	H1	-0.5827	I1	H2	-0.4308
I2	H1	0.5234	I2	H2	0.1625
I3	H1	-0.2391	I3	H2	0.4895
I4	H1	-0.3888	I4	H2	0.0038
I5	H1	-0.1377	I5	H2	0.3337
I6	H1	-0.5985	I6	H2	0.5270
I7	H1	-0.6966	I7	H2	0.4009
-	-	-	B2	O1	5808.0587
-	-	-	H1	O1	-0.5730
-	-	-	H2	O1	0.1317

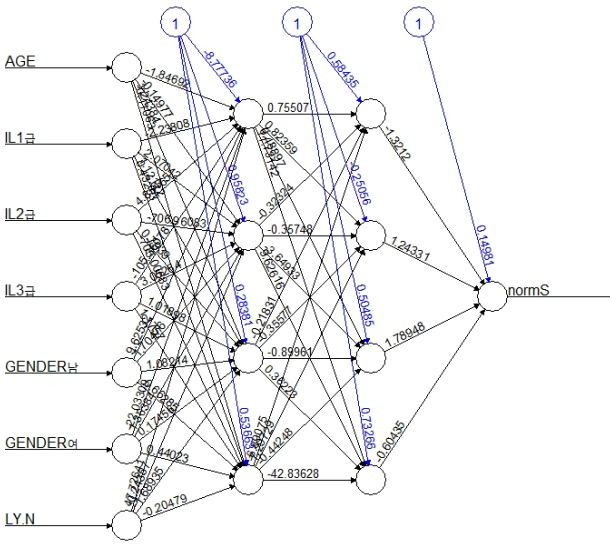
나. 다차원 신경망모형(딥러닝)

신경망모형의 은닉층이 2층 이상일 때를 딥러닝(Deep learning)이라고 한다. 다른 장에서 봤던 모형들보다 훨씬 많은 모수들이 추정된다. 이것은 신경망모형이 해석력이 떨어지고 흔히 모형 안이 블랙박스라고 부르는 이유이다. 중간과정을 해석하기는 쉽지 않고 다만 입력변수와 그에 따른 출력변수가 계산된다.

〈그림 Ⅴ-8〉 딩러닝모형 빈도 예시



〈그림 Ⅴ-9〉 딩러닝모형 심도 예시



딥러닝모형은 “nnet” 패키지에서는 구현하기 어렵고 “neuralnet” 패키지에서 모델링할 수 있다. R 패키지에서 제공하는 것은 fully connected된 모형인데 tensorflow나 keras와 같은 오픈소스에서는 그렇지 않은 모형에 대해서 계산이 가능하다. fully connected가 아닌 모형의 가중치들에는 제한을 주는데 이는 매우 복잡하여 역전파 알고리즘을 못 쓰게 된다. 즉, 그래디어트 계산이 어려워지는데 이것을 해결해 주는 것이 tensorflow나 keras와 같은 오픈소스이다. 이는 본 연구의 범위를 넘어가는 주제로 관심이 있다면 아래 웹사이트를 방문하는 것을 추천한다.

(https://tensorflow.rstudio.com/)

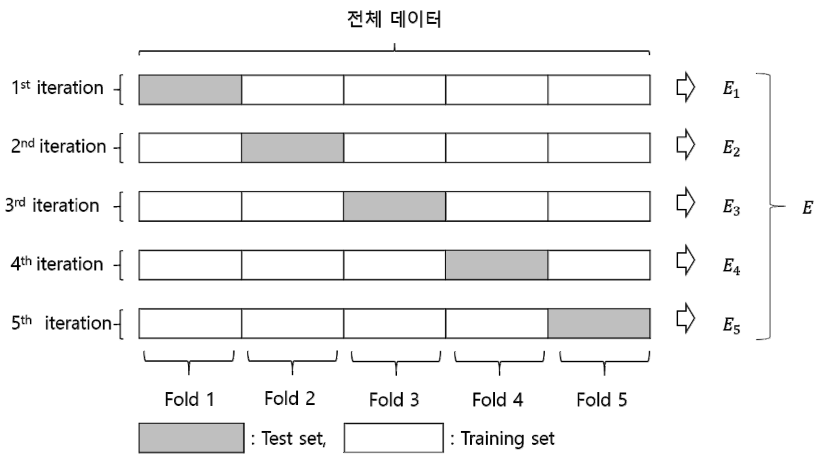
Ⅵ. 시사점 및 결론

1. 모형 비교

가. 평균 절댓값 오차(Mean Absolute Error) 비교

k-묶음 교차검증(k-fold cross validation)을 통해서 앞서 살펴본 방법론들을 비교한다. 여기서 k-묶음이란 것은 전체데이터를 k개로 분할한 것을 의미하는데, 이는 무작위로 서로 겹치는 것을 분할한다. 본 연구에서는 5-묶음 교차검증을 실시하였다. 전체 데이터를 5등분하면 그중 한 묶음은 음영처리된 것은 테스트 데이터를 의미하고 나머지는 훈련 데이터로 모형을 만드는데 쓰이는 데이터를 의미한다. 훈련 데이터를 통하여 모형이 만들어지면 테스트 데이터를 바탕으로 그 모형을 검증한다.

〈그림 Ⅵ-1〉 5-묶음 교차검증



검증할 때에 MSE(Mean Squared Error)를 측정지표로 사용하는 경우도 많지만 여기서는 MAE(Mean Absolute Error)를 사용하여 교차검증의 측정 지표로 사용한다. MSE를 지표로 사용하면 실제값과 예측값의 차이가 1 이상일 때의 오차를 1 이하일 때보다 더 크게 측정하게 되는데 해당 데이터의 빈도 데이터는 1 근처의 값이 많으므로 적합하지 않다.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\text{실제값} - \text{예측값}| \quad \text{(Ⅵ-1)}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}|\text{실제값} - \text{예측값}|^2 \quad \text{(Ⅵ-2)}$$

아래 〈표 Ⅵ-1〉은 빈도 모델들의 교차검증 결과이다. 우선 의사결정나무 모형의 교차검증결과가 다른 모형들에 비해 예측력이 낮은 것을 볼 수 있다. 이를 보완하기 위한 앙상블기법을 이용하면 의사결정나무보다 나은 검증결과를 얻을 수 있다. MARS의 모형도 로그 연결함수를 사용하기 위해 GLM 옵션을 넣었을 때 오차가 낮아지기는 하나 미미한 수준이다. 또한 실무에서 많이 사용하고 있는 GLM보다는 다른 머신러닝모델들의 오차가 작은 것을 볼 수 있다.

〈표 Ⅵ-1〉 빈도 모델 교차검증 결과

구분	MAE
GLM	0.3546
MARS	0.3217
MARS with GLM option	0.3210
CART cp=0.01	1.2229
CART cp=0.005	1.2392
신경망 hidden=1	0.3177
신경망 hidden=2	0.2615
배깅	0.3218
랜덤 포레스트	0.3461

심도 모델은 빈도 모델만큼 의사결정나무와 그 외의 모델들의 MAE가 차이나지는 않는다. 〈표 Ⅵ-2〉를 살펴보면 심도 모델에서는 로그연결함수를 사용하였을 때 그 오차가 줄어드는 것을 볼 수 있다. GLM option이 있는 MARS 모형이 가장 오차가 적었으며 그 다음으로 감마분포를 가정한 GLM의 MAE가 적었다.

〈표 Ⅵ-2〉 심도 모델 교차검증 결과

구분	MAE
GLM	200870.65
MARS	203539.92
MARS with GLM option	197988.91
CART cp=0.01	208798.11
CART cp=0.005	209081.35
신경망 hidden=1	201141.43
신경망 hidden=2	223396.34
배깅	203458.21
랜덤 포레스트	201725.93

나. 평균 비교 예시

다음의 표는 데이터의 평균값과 다양한 모형을 바탕으로 예측값을 비교한 것이다. 종속변수는 기대빈도, 설명변수는 연령, 상해급수, 성별, 직전연도 발생건수를 사용하였고 이에 대한 그룹을 세분화하여 보여주고 있다. GLM(교호작용 반영)은 각 설명변수끼리의 교호작용을 반영한 모형을 뜻한다. 통계적 기법인 GLM을 바탕으로 한 예측값들이 데이터의 평균에서 가장 벗어나 있는 반면 머신러닝기법들은 평균에 가장 가깝게 추정되었다.

〈표 Ⅵ-3〉 40세 상해급수 1급 예시

성별	직전연도 발생건수	평균: N	GLM	GLM (교호작용 반영)	CART	배깅	랜덤 포레스트	MARS	MARS (link = log)	신경망 모형 (h=20)
남	0건	0.0903	0.1243	0.1102	0.0869	0.0869	0.1205	0.0485	0.0820	0.0873
남	1건	0.4783	0.2145	0.2005	0.4897	0.4888	0.4293	0.5832	0.4820	0.4162
여	0건	0.1283	0.1973	0.1934	0.1307	0.1356	0.1717	0.1351	0.1200	0.1190
여	1건	0.7079	0.3405	0.3431	0.6617	0.6904	0.6451	0.6698	0.7053	0.6769

〈표 Ⅵ-4〉 50세 상해급수 1급 예시

성별	직전연도 발생건수	평균: N	GLM	GLM (교호작용 반영)	CART	배깅	랜덤 포레스트	MARS	MARS (link = log)	신경망 모형 (h=20)
남	0건	0.0846	0.1434	0.1443	0.0869	0.0842	0.1282	0.0918	0.0973	0.0876
남	1건	0.5000	0.2474	0.2547	0.4897	0.4926	0.4910	0.6265	0.5717	0.5082
여	0건	0.1607	0.2276	0.2289	0.1307	0.1630	0.2029	0.1784	0.1424	0.1315
여	1건	0.8293	0.3928	0.3938	0.8484	0.8551	0.7792	0.7131	0.8365	0.8003

2. 모형별 시사점 및 제한점

본 절에서는 앞서 살펴보았던 모형별 시사점 및 제한점을 살펴본다. 본 연구에서 살펴본 통계적 기법에는 일반화선형모형(GLM)과 일반화선형혼합모형(GLMM)이 있다. GLM은 보험회사 요율산정에서 최근까지 많이 쓰이고 있는 통계적 기법이다. GLM은 해석하기 용이하고 선형회귀보다 포괄적인 분석이 가능하다는 장점이 있다. 최근의 보험데이터 분석 관련 선행연구들에서는 GLM과 머신러닝기법을 비교하는데 후자의 성능이 더 우수하다는 결과가 많은 편이다. GLMM은 GLM의 고정효과에 임의효과를 추가한 모형이다. 특히 동일 계약자가 반복되는 데이터에서 사용하는데 실적변수가 동일 계약자 안에서는 의존적인 것을 모델링할 수 있다.

본 연구에서는 머신러닝의 기법으로 의사결정나무, 앙상블기법, MARS, 신경망모형

을 실손의료보험 자료에 적용해보았다. 이 알고리즘들은 지도학습기법으로 회귀나 분류의 결과 값을 주어 손해보험 가격산정에 쓰인다(Marechal 2018). CART는 예측모형으로는 우수한 모형이 아닐 수 있지만 나무모형의 그림과 이진분리로 해석력이 좋고 데이터의 시각화에 용이하다. 또한 요율산정을 할 때에 CART 모형은 분석에 가장 중요한 변수를 최초분리변수로 선택한다. 실손의료보험 자료를 통한 분석에서 현행요율변수가 아닌 실적변수가 최초로 분리되는 것을 관찰하였다. 또한 각 노드의 분리기준을 통하여 연속형 변수를 어떻게 범주화시킬 것인지에 대한 도움을 주기도 한다(Werner & Moldin 2016).

의사결정나무의 예측력 향상을 돕기 위하여 앙상블기법들이 발전되었다. 배깅과 랜덤 포레스트가 비슷한 알고리즘을 가지고 있다. 이 둘의 공통점은 자료로부터 본래 자료 크기를 복원 추출하여 여러 개의 나무모형을 만드는 알고리즘이라는 것인데 제한점은 붓스트랩핑할 때에 일부자료가 누락될 수 있다는 것이다. 또한 둘의 차이점은 배깅의 경우 모든 설명변수를 다 쓰는 반면 랜덤 포레스트는 이를 랜덤하게 추출하여 모델링한다. 그렇기 때문에 대형자료에서 중요한 변수를 찾는데 용이하다는 장점이 있다.

배깅은 분류기(Classifiers)들이 독립적으로 생성되는 반면 부스팅은 이전의 분류기에 의존하여 생성된다. 쉽게 말하자면 분류의 경계에 있는 잘못 분류된 자료들에 가중치를 더 주는 것이다. 이는 이상치에 민감하고 해석력이 부족하며 과대적합 경향이 있다.

MARS 분석은 단계선형함수(Piecewise linear function)로 비선형의 데이터에 적용한다. 많은 면에서 회귀분석과 비슷하지만 비선형성, 교호작용, missing data와 같이 선형회귀로는 어려움을 겪는 것들에 대해서 해결해 준다. 그리고 MARS는 매듭점을 찾아주는데 이는 요율산정에서 연속형 변수를 범주화시킬 때 사용된다(Werner & Moldin 2016). 또한 GLM option이 있는 MARS 분석을 통해 보험 자료에 적합한 분석을 할 수 있고 반응변수에 대해 설명력 있는 변수 순으로 선택한다. R에서 MARS와 earth 함수를 검증하였을 때 두 함수 모두 비슷한 결과가 산출된다(Francis et al. 2018).

신경망모형을 선택하는 가장 큰 이유 중 하나는 예측력이 우수하기 때문이다. 하지만 만약 종속변수와 독립변수 간에 선형관계가 있다면, 전통적인 방법인 회귀분석이나 요인분석(Factor analysis)이 신경망모형보다 더 나은 결과를 가져온다(Francis

2001). 또한 독립변수와 종속변수 간의 함수적인 관계를 알고 있을 때에는 회귀가 신경망모형을 능가한다(Warner & Misra 1996). 신경망모형의 가장 큰 장점은 비선형의 데이터에서 모델링을 하는 것이다. 신경망모형은 입력변수에 민감하게 반응하므로 분석 시에는 자료를 변형할 필요가 있다.

앞에서는 CART와 MARS뿐만 아니라 블랙박스라고만 여겨졌던 신경망모형의 수학적 구조에 대해서도 알아보았다. 다양한 머신러닝기법들의 알고리즘과 특징을 살펴본 이유는 방법론을 이해하고 올바르게 사용하기 위함이다. 그래야만 데이터에 적용하고 통찰력 있는 분석을 할 수 있다. 추후에는 실손의료보험 자료뿐만 아니라 다양한 보험자료에 머신러닝기법들을 적용해보는 학계와 실무진들의 연구가 필요할 것이다.

참고문헌

강현철·한상태·최종후·이성건·김은석·엄익현(2014), 『빅데이터 분석을 위한 데이터 마이닝 방법론』, 자유아카데미

기승도·김대환(2009), 『일반화선형모형[GLM]을 이용한 자동차보험 요율상대도 산출 방법 연구』, 보험연구원

김대환·이봉주(2013), 「실손의료보험의 역선택 분석」, 『보험학회지』, 제96집

박창이·김용대·김진석·송종우·최호식(2011), 『R을 이용한 데이터마이닝』, 교우사

이경아·이항석(2016), 「실손의료보험의 역선택과 보험료 차등화」, 『리스크관리연구』, 27(3)

이항석·이가은·이경아(2017), 「실손의료보험에서 요율산정과 일반화선형모형(GLM)의 활용」, 『리스크관리연구』, 28(4)

이항석·이민하·백혜연(2018), 「실손의료보험 할인할증제도의 실증분석」, 『보험학회지』, 116

이항석·이수빈·백혜연(2017), 「실손의료보험에서 신뢰도기법을 반영한 보험료 산정」, 『보험학회지』, 111

임준·황인창·이성은(2015), 「보험산업의 빅데이터 활용현황 및 향후과제」, 『KIRI 리포트』, 341

전희주·최용석·최종후·기승도·김은석(2009), 『보험자료를 활용한 일반화 선형모형』, 사이플러스

조재린·정성희(2018), 『계리적 관점에서 본 실손의료보험 개선 방안』, 보험연구원

Aleandri, M.(2018), "Modelling Policyholder Behaviour through Machine Learning Techniques.", International Congress of Actuaries 2018

Breiman, L., Friedman, J., Stone, C., Olshen, R.(1984), *Classification and Regression Trees*, Wadsworth International Group

Charpentier, A.(2015), *Computational Actuarial Science with R*, CRC Press

Faraway, J.(2016), *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, CRC Press

Francis, L.(2003), "Martian Chronicles: Is MARS better than Neural Networks?", CAS

_____(2001), "Neural Networks Demystified", CAS

Francis, L., Endo, F., Kannon, M., Oda, H., Wolfstein, A.(2018), "ASTIN big data working party phase Ⅱ: Predictive modeling", International Congress of Actuaries 2018

Friedman, J.(2001), "Greedy Function Approximation : A Gradient Boosting Machine", *Annals of Statistics*, 29, 1189-1232

_____ (1991), "Multivariate Adaptive Regression Splies", *Annals of Statistics*, 19(1)

Fritsch, S., Guenther F., Suling M., Mueller, S.(2016), "Package 'neuralnet'"

Goldburd, M., Khare, A., Tevet, D.(2016), "Generalized linear models for insurance rating", *Casualty Actuarial Society*

Gupta, T., Mukherjee, S., Rajasekaram, P.(2018), "Decoding machine learning algorithm", International Congress of Actuaries 2018

Gutherie, K.(2018), "Teaching and Testing Predictive analytics: Innovations in Pedagogy and assessment", International Congress of Actuaries 2018

Ha, H.(2018), "An evaluation of withdrawal benefits in variable annuities via machine learning", International Congress of Actuaries 2018

Hadidi, N.(2003), "Classification Ratemaking Using Decision Trees", CAS

Hartigan, J.(1975), *Clustering Algorithms*, Wiley Series in Probability and Mathematical Statistics

Hastie, T., Tibshirani, R.(1990), *Generalized additive models*, Chapman and Hall, London

Hastie, T., Tibshirani, R., Friedman, J.(2008), *The elements of Statistical Learning*, Springer

Jong, P., Heller, G.(2008), *Generalized Linear Models for Insurance Data*, International Series on Actuarial Science, Cambridge

Loser, F.(2018), "Machine learning vs Actuarial Methods in claim prediction", International Congress of Actuaries 2018

Marechal, X.(2018), "Machine learning applications for non-life pricing", International Congress of Actuaries 2018

Mukherjee, S., Ajmani, A.(2018), "application of machine learning in health insurance – to reduce claims leakage and improve underwriting", International Congress of Actuaries 2018

Mukherjee, S., Vijayaraghavan, A.(2018), "application of classical reserving tech alongside ML algorithms – optimal use of big data", International Congress of Actuaries 2018

Schapire, R. E., Singer, Y.(1999), "Improved boosting algorithms using confidence - rated predictions, Machine Learning", 37

Simon, P.(2013), *Too big to ignore: the business case for big data*, Wiley

Therneau, T., Atkinson, B., Ripley, B.(2018), "Package 'rpart'", R Document

Warner, B., Misra. M.(1996), "Understanding Neural Networks as Statistical Tools", *American Statistician*, 50(4)

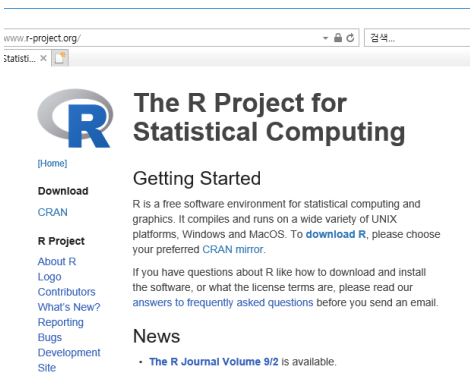
Werner, G., Modlin, C.(2016), "Basic Ratemaking", CAS

Wuthrich, M.(2017), "Neural networks applied to chain-ladder reserving", International Congress of Actuaries 2018

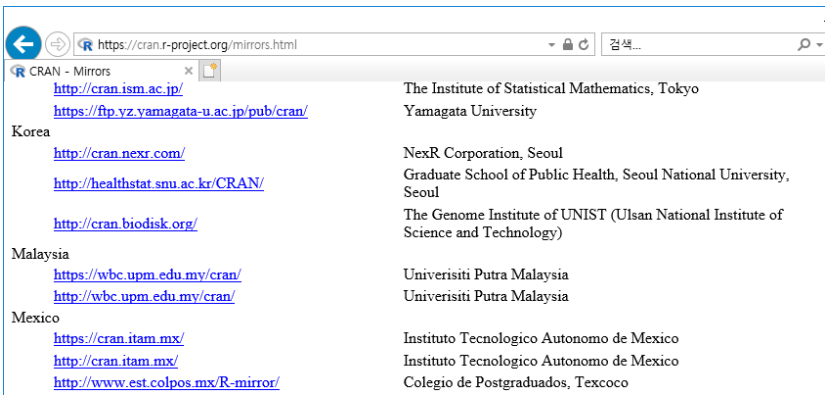
부록. R 프로그래밍 설치 및 코드

1. R 프로그래밍 설치방법

〈부록 그림 1〉 R 설치 자료화면 1



〈부록 그림 2〉 R 설치 자료화면 2



① http://www.r-project.org에 접속

② 좌측 첫 번째 메뉴 CRAN을 클릭

③ 나라별 주소 중 Korea에서 3개 중 무엇이든 다운로드 및 설치

2. R 프로그래밍 코드

#해당 패키지가 없을 시에는 install.packages("패키지명")를 통하여 설치를 하고 library(패키지명) 함수를 시행한다. 아래에는 보고서에서 사용한 패키지들과 함수들을 나열하였다. 함수 안에는 식과 나머지 옵션을 넣는 형태이다. 만약 변수를 범주화 시키고 싶으면 factor(변수명)을 사용하고 연속형으로 분석할 시에는 변수명만 써준다.

```
#GLM
glm(factor(반응변수)~설명변수1+설명변수2, data=자료)
#GLMM
#Charpentier(2015) 참고
#CART
library(rpart)
모형이름<-rpart(factor(반응변수)~설명변수1+설명변수2, data=자료)
library(rattle) #나무모형그림을 위한 것.
plot(모형이름)
#MARS
library(earth)
earth(factor(반응변수)~설명변수1+설명변수2, data=자료)
#ENSEMBLE METHODS
library(ipred) #bagging
bagging(factor(반응변수)~설명변수1+설명변수2, data=자료)
library(gbm); library(adabag) #boosting
```

gbm(factor(반응변수)~설명변수1+설명변수2, data=자료)

library(randomForest) #randomforest

randomForest(factor(반응변수)~설명변수1+설명변수2, data=자료)

#NEURAL NETWORKS

library(dummy) #데이터를 가변수화 시킴.

set.seed(숫자) # 이 함수를 먼저 써주고 모형을 만들면 다음 시행에도 같은 #모형 추정해줌.

library(nnet) # 은닉층 2층 이상 불가.

library(neuralnet)

nnet(factor(반응변수)~설명변수1+설명변수2, data=자료)

neuralnet(factor(반응변수)~설명변수1+설명변수2, data=자료)

보험연구원(KIRI) 발간물 안내

※ 2017년부터 기존의 연구보고서, 정책보고서, 경영보고서, 조사보고서가 연구보고서로 통합되었습니다.

■ 연구보고서

2017-1 보험산업 미래 / 김석영·윤성훈·이선주 2017.2
2017-2 자동차보험 과실상계제도 개선방안 / 전용식·채원영 2017.2
2017-3 상호협정 관련 입법정책 연구 / 정호열 2017.2
2017-4 저소득층 노후소득 보장을 위한 공사연계연금 연구 / 정원석·강성호·마지혜 2017.3
2017-5 자영업자를 위한 사적소득보상체계 개선방안 / 류건식·강성호·김동겸 2017.3
2017-6 우리나라 사회안전망 개선을 위한 현안 과제 / 이태열·최장훈·김유미 2017.4
2017-7 일본의 보험회사 도산처리제도 및 사례 / 정봉은 2017.5
2017-8 보험회사 업무위탁 관련 제도 개선방안 / 이승준·정인영 2017.5
2017-9 부채시가평가제도와 생명보험회사의 자본관리 / 조영현·이혜은 2017.8
2017-10 효율적 의료비 지출을 통한 국민건강보험의 보장성 강화 방안 / 김대환 2017.8
2017-11 인슈어테크 혁명: 현황 점검 및 과제 고찰 / 박소정·박지윤 2017.8
2017-12 생산물 배상책임보험 역할 제고 방안 / 이기형·이규성 2017.9
2017-13 보험금청구권과 소멸시효 / 권영준 2017.9
2017-14 2017년 보험소비자 설문조사 / 동향분석실 2017.10
2017-15 2018년도 보험산업 전망과 과제 / 동향분석실 2017.11
2017-16 퇴직연금 환경변화와 연금세제 개편 방향 / 강성호·류건식·김동겸 2017.12
2017-17 자동차보험 한방진료 현황과 개선방안 / 송윤아·이소양 2017.12
2017-18 베이비부머 세대의 노후소득 / 최장훈·이태열·김미화 2017.12
2017-19 연금세제 효과연구 / 정원석·이선주 2017.12
2017-20 주요국의 지진보험 운영 현황 및 시사점 / 최창희·한성원 2017.12
2017-21 사적연금의 장기연금수령 유도방안 / 김세중·김유미 2017.12
2017-22 누적전망이론을 이용한 생명보험과 연금의 유보가격 측정 연구 / 지홍민 2017.12
2018-1 보증연장 서비스 규제 방안 / 백영화·박정희 2018.1
2018-2 건강생활서비스 공·사 협력 방안 / 조용운·오승연·김동겸 2018.2
2018-3 퇴직연금 가입자교육 개선 방안 / 류건식·강성호·이상우 2018.2

2018-4 IFRS 9과 보험회사의 ALM 및 자산배분 / 조영현·이혜은 2018.2
2018-5 보험상품 변천과 개발 방향 / 김석영·김세영·이선주 2018.2
2018-6 계리적 관점에서 본 실손의료보험 개선 방안 / 조재린·정성희 2018.3
2018-7 국내 보험회사의 금융겸업 현황과 시사점 / 전용식·이혜은 2018.3
2018-8 장애인의 위험보장 강화 방안 / 오승연·김석영·이선주 2018.4
2018-9 주요국 공·사 건강보험 연계 체계 분석 / 정성희·이태열·김유미 2018.4
2018-10 정신질환 위험보장 강화 방안 / 이정택·임태준·김동겸 2018.4
2018-11 기초서류 준수의무 위반 시 과징금 부과기준 개선방안 / 황현아·백영화·권오경 2018.9
2018-12 2018년 보험소비자 설문조사 / 동향분석실 2018.9
2018-13 상속법의 관점에서 본 생명보험 / 최준규 2018.9
2018-14 호주 퇴직연금제도 현황과 시사점 / 이경희 2018.9
2018-15 빅데이터 기반의 사이버위험 측정 방법 및 사이버사고 예측모형 연구 / 이진무 2018.9

■ 연구보고서(구)

2008-1 보험회사의 리스크 중심 경영전략에 관한 연구 / 최영목·장동식·김동겸 2008.1
2008-2 한국 보험시장과 공정거래법 / 정호열 2008.6
2008-3 확정급여형 퇴직연금의 자산운용 / 류건식·이경희·김동겸 2008.3
2009-1 보험설계사의 특성분석과 고능률화 방안 / 안철경·권오경 2009.1
2009-2 자동차사고의 사회적 비용 최소화 방안 / 기승도 2009.2
2009-3 우리나라 가계부채 문제의 진단과 평가 / 유경원·이혜은 2009.3
2009-4 사적연금의 노후소득보장 기능제고 방안 / 류건식·이창우·김동겸 2009.3
2009-5 일반화선형모형(GLM)을 이용한 자동차보험 요율상대도 산출 방법 연구 / 기승도·김대환 2009.8
2009-6 주행거리에 연동한 자동차보험제도 연구 / 기승도·김대환·김혜란 2010.1
2010-1 우리나라 가계 금융자산 축적 부진의 원인과 시사점 / 유경원·이혜은 2010.4
2010-2 생명보험 상품별 해지율 추정 및 예측 모형 / 황진태·이경희 2010.5
2010-3 보험회사 자산관리서비스 사업모형 검토 / 진 익·김동겸 2010.7

■ 정책보고서(구)

2008-2 환경오염리스크관리를 위한 보험제도 활용방안 / 이기형 2008.3

2008-3 금융상품의 정의 및 분류에 관한 연구 / 유지호·최 원 2008.3
2008-4 2009년도 보험산업 전망과 과제 / 이진면·이태열·신종협·황진태·유진아·김세환·이정환·박정희·김세중·최이섭 2008.11
2009-1 현 금융위기 진단과 위기극복을 위한 정책제언 / 진 익·이민환·유경원·최영목·최형선·최 원·이경아·이혜은 2009.2
2009-2 퇴직연금의 급여 지급 방식 다양화 방안 / 이경희 2009.3
2009-3 보험분쟁의 재판외적 해결 활성화 방안 / 오영수·김경환·이종욱 2009.3
2009-4 2010년도 보험산업 전망과 과제 / 이진면·황진태·변혜원·이경희·이정환·박정희·김세중·최이섭 2009.12
2009-5 금융상품판매전문회사의 도입이 보험회사에 미치는 영향 / 안철경·변혜원·권오경 2010.1
2010-1 보험사기 영향요인과 방지방안 / 송윤아 2010.3
2010-2 2011년도 보험산업 전망과 과제 / 이진면·김대환·이경희·이정환·최 원·김세중·최이섭 2010.12
2011-1 금융소비자 보호 체계 개선방안 / 오영수·안철경·변혜원·최영목·최형선·김경환·이상우·박정희·김미화 2010.4
2011-2 일반공제사업 규제의 합리화 방안 / 오영수·김경환·박정희 2011.7
2011-3 퇴직연금 적립금의 연금전환 유도방안 / 이경희 2011.5
2011-4 저출산·고령화와 금융의 역할 / 윤성훈·류건식·오영수·조용운·진 익·유진아·변혜원 2011.7
2011-5 소비자 보호를 위한 보험유통채널 개선방안 / 안철경·이경희 2011.11
2011-6 2012년도 보험산업 전망과 과제 / 윤성훈·황진태·이정환·최 원·김세중·오병국 2011.12
2012-1 인적사고 보험금의 지급방식 다양화 방안 / 조재린·이기형·정인영 2012.8
2012-2 보험산업 진입 및 퇴출에 관한 연구 / 이기형·변혜원·정인영 2012.10
2012-3 금융위기 이후 보험규제 변화 및 시사점 / 임준환·유진아·이경아 2012.11
2012-4 소비자중심의 변액연금보험 개선방안 연구: 공시 및 상품설계 개선을 중심으로 / 이기형·임준환·김해식·이경희·조영현·정인영 2012.12
2013-1 생명보험의 자살면책기간이 자살에 미치는 영향 / 이창우·윤상호 2013.1
2013-2 퇴직연금 지배구조체계 개선방안 / 류건식·김대환·이상우 2013.1
2013-3 2013년도 보험산업 전망과 과제 / 윤성훈·전용식·이정환·최 원·김세중·채원영 2013.2
2013-4 사회안전망 체제 개편과 보험산업 역할 / 진 익·오병국·이성은 2013.3
2013-5 보험지주회사 감독체계 개선방안 연구 / 이승준·김해식·조재린 2013.5

2013-6 2014년도 보험산업 전망과 과제 / 윤성훈·전용식·최 원·김세중·채원영 2013.12
2014-1 보험시장 경쟁정책 투명성 제고방안 / 이승준·강민규·이해랑 2014.3
2014-2 국내 보험회사 지급여력규제 평가 및 개선방안 / 조재린·김해식·김석영 2014.3
2014-3 공·사 사회안전망의 효율적인 역할 제고 방안 / 이태열·강성호·김유미 2014.4
2014-4 2015년도 보험산업 전망과 과제 / 윤성훈·김석영·김진억·최 원·채원영·이아름·이해랑 2014.11
2014-5 의료보장체계 합리화를 위한 공·사건강보험 협력방안 / 조용운·김경환·김미화 2014.12
2015-1 보험회사 재무건전성 규제 - IFRS와 RBC 연계방안 / 김해식·조재린·이경아 2015.2
2015-2 2016년도 보험산업 전망과 과제 / 윤성훈·김석영·김진억·최 원·채원영·이아름·이해랑 2015.11
2016-1 정년연장의 노후소득 개선 효과와 개인연금의 정책방향 / 강성호·정봉은·김유미 2016.2
2016-2 국민건강보험 보장률 인상 정책 평가: DSGE 접근법 / 임태준·이정택·김혜란 2016.11
2016-3 2017년도 보험산업 전망과 과제 / 동향분석실 2016.12

■ 경영보고서(구)

2009-1 기업휴지보험 활성화 방안 연구 / 이기형·한상용 2009.3
2009-2 자산관리서비스 활성화 방안 / 진 익 2009.3
2009-3 탄소시장 및 녹색보험 활성화 방안 / 진 익·유시용·이경아 2009.3
2009-4 생명보험회사의 지속가능성장에 관한 연구 / 최영목·최 원 2009.6
2010-1 독립판매채널의 성장과 생명보험회사의 대응 / 안철경·권오경 2010.2
2010-2 보험회사의 윤리경영 운영실태 및 개선방안 / 오영수·김경환 2010.2
2010-3 보험회사의 퇴직연금사업 운영전략 / 류건식·이창우·이상우 2010.3
2010-4(1) 보험환경변화에 따른 보험산업 성장방안 / 산업연구실·정책연구실·동향분석실 2010.6
2010-4(2) 종합금융서비스를 활용한 보험산업 성장방안 / 금융제도실·재무연구실 2010.6
2010-5 변액보험 보증리스크관리연구 / 권용재·장동식·서성민 2010.4

2010-6 RBC 내부모형 도입 방안 / 김해식·최영목·김소연·장동식·서성민 2010.10
2010-7 금융보증보험 가격결정모형 / 최영수 2010.7
2011-1 보험회사의 비대면채널 활용방안 / 안철경·변혜원·서성민 2011.1
2011-2 보증보험의 특성과 리스크 평가 / 최영목·김소연·김동겸 2011.2
2011-3 충성도를 고려한 자동차보험 마케팅전략 연구 / 기승도·황진태 2011.3
2011-4 보험회사의 상조서비스 기여방안 / 황진태·기승도·권오경 2011.5
2011-5 사기성클레임에 대한 최적조사방안 / 송윤아·정인영 2011.6
2011-6 민영의료보험의 보험리스크관리방안 / 조용운·황진태·김미화 2011.8
2011-7 보험회사의 개인형 퇴직연금 운영방안 / 류건식·김대환·이상우 2011.9
2011-8 퇴직연금시장의 환경변화에 따른 확정기여형 퇴직연금 운영방안 / 김대환·류건식·이상우 2011.10
2012-1 국내 생명보험회사의 기업공개 평가와 시사점 / 조영현·전용식·이혜은 2012.7
2012-2 보험산업 비전 2020 : ⓝ sure 4.0 / 진 익·김동겸·김혜란 2012.7
2012-3 현금흐름방식 보험료 산출의 시행과 과제 / 김해식·김석영·김세영·이혜은 2012.9
2012-4 보험회사의 장수리스크 발생원인과 관리방안 / 김대환·류건식·김동겸 2012.9
2012-5 은퇴가구의 경제형태 분석 / 유경원 2012.9
2012-6 보험회사의 날씨리스크 인수 활성화 방안: 지수형 날씨보험을 중심으로 / 조재린·황진태·권용재·채원영 2012.10
2013-1 자동차보험시장의 가격경쟁이 손해율에 미치는 영향과 시사점 / 전용식·채원영 2013.3
2013-2 중국 자동차보험 시장점유율 확대방안 연구 / 기승도·조용운·이소양 2013.5
2016-1 뉴 노멀 시대의 보험회사 경영전략 / 임준환·정봉은·황인창·이혜은·김혜란·정승연 2016.4
2016-2 금융보증보험 잠재 시장 연구: 지방자치단체 자금조달 시장을 중심으로 / 최창희·황인창·이경아 2016.5
2016-3 퇴직연금시장 환경변화와 보험회사 대응방안 / 류건식·강성호·김동겸 2016.5

■ 조사보고서(구)

2008-1 보험회사 글로벌화를 위한 해외보험시장 조사 / 양성문·김진억·지재원·박정희·김세중 2008.2
2008-2 노인장기요양보험 제도 도입에 대응한 장기간병보험 운영 방안 / 오영수 2008.3
2008-3 2008년 보험소비자 설문조사 / 안철경·기승도·이상우 2008.4
2008-4 주요국의 보험상품 판매권유 규제 / 이상우 2008.3
2009-1 2009년 보험소비자 설문조사 / 안철경·이상우·권오경 2009.3
2009-2 Solvency II의 리스크 평가모형 및 측정 방법 연구 / 장동식 2009.3
2009-3 이슬람 보험시장 진출방안 / 이진면·이정환·최이섭·정중영·최태영 2009.3
2009-4 미국 생명보험 정산거래의 현황과 시사점 / 김해식 2009.3
2009-5 헤지펀드 운용전략 활용방안 / 진 익·김상수·김종훈·변귀영·유시용 2009.3
2009-6 복합금융 그룹의 리스크와 감독 / 이민환·전선애·최 원 2009.4
2009-7 보험산업 글로벌화를 위한 정책적 지원방안 / 서대교·오영수·김영진 2009.4
2009-8 구조화금융 관점에서 본 금융위기 분석 및 시사점 / 임준환·이민환·윤건용·최 원 2009.7
2009-9 보험리스크 측정 및 평가 방법에 관한 연구 / 조용운·김세환·김세중 2009.7
2009-10 생명보험계약의 효력상실·해약분석 / 류건식·장동식 2009.8
2010-1 과거 금융위기 사례분석을 통한 최근 글로벌 금융위기 전망 / 신종협·최형선·최 원 2010.3
2010-2 금융산업의 영업행위 규제 개선방안 / 서대교·김미화 2010.3
2010-3 주요국의 민영건강보험의 운영체계와 시사점 / 이창우·이상우 2010.4
2010-4 2010년 보험소비자 설문조사 / 변혜원·박정희 2010.4
2010-5 산재보험의 운영체계에 대한 연구 / 송윤아 2010.5
2010-6 보험산업 내 공정거래규제 조화방안 / 이승준·이종욱 2010.5
2010-7 보험종류별 진료수가 차등적용 개선방안 / 조용운·서대교·김미화 2010.4
2010-8 보험회사의 금리위험 대응전략 / 진 익·김해식·유진아·김동겸 2011.1
2010-9 퇴직연금 규제체계 및 정책방향 / 류건식·이창우·이상우 2010.7
2011-1 생명보험설계사 활동실태 및 만족도 분석 / 안철경·황진태·서성민 2011.6
2011-2 2011년 보험소비자 설문조사 / 김대환·최 원 2011.5
2011-3 보험회사 녹색금융 참여방안 / 진 익·김해식·김혜란 2011.7
2011-4 의료시장 변화에 따른 민영실손의료보험의 대응 / 이창우·이기형 2011.8

2011-5 아세안 주요국의 보험시장 규제제도 연구 / 조용운·변혜원·이승준·김경환·오병국 2011.11
2012-1 2012년 보험소비자 설문조사 / 황진태·전용식·윤상호·기승도·이상우·최 원 2012.6
2012-2 일본의 퇴직연금제도 운영체계 특징과 시사점 / 이상우·오병국 2012.12
2012-3 솔벤시 Ⅱ의 보고 및 공시 체계와 시사점 / 장동식·김경환 2012.12
2013-1 2013년 보험소비자 설문조사 / 전용식·황진태·변혜원·정원석·박선영·이상우·최 원 2013.8
2013-2 건강보험 진료비 전망 및 활용방안 / 조용운·황진태·조재린 2013.9
2013-3 소비자 신뢰 제고와 보험상품 정보공시 개선방안 / 김해식·변혜원·황진태 2013.12
2013-4 보험회사의 사회적 책임 이행에 관한 연구 / 변혜원·조영현 2013.12
2014-1 주택연금 연계 간병보험제도 도입 방안 / 박선영·권오경 2014.3
2014-2 소득수준을 고려한 개인연금 세제 효율화방안: 보험료 납입단계의 세제방식 중심으로 / 정원석·강성호·이상우 2014.4
2014-3 보험규제에 관한 주요국의 법제연구: 모집채널, 행위 규제 등을 중심으로 / 한기정·최준규 2014.4
2014-4 보험산업 환경변화와 판매채널 전략 연구 / 황진태·박선영·권오경 2014.4
2014-5 거시경제 환경변화의 보험산업 파급효과 분석 / 전성주·전용식 2014.5
2014-6 국내경제의 일본식 장기부진 가능성 검토 / 전용식·윤성훈·채원영 2014.5
2014-7 건강생활관리서비스 사업모형 연구 / 조용운·오승연·김미화 2014.7
2014-8 보험개인정보 보호법제 개선방안 / 김경환·강민규·이해랑 2014.8
2014-9 2014년 보험소비자 설문조사 / 전용식·변혜원·정원석·박선영·오승연·이상우·최 원 2014.8
2014-10 보험회사 수익구조 진단 및 개선방안 / 김석영·김세중·김혜란 2014.11
2014-11 국내 보험회사의 해외사업 평가와 제언 / 전용식·조영현·채원영 2014.12
2015-1 보험민원 해결 프로세스 선진화 방안 / 박선영·권오경 2015.1
2015-2 재무건전성 규제 강화와 생명보험회사의 자본관리 / 조영현·조재린·김혜란 2015.2
2015-3 국내 배상책임보험 시장 성장 저해 요인 분석 – 대인사고 손해배상액 산정기준을 중심으로 – / 최창희·정인영 2015.3
2015-4 보험산업 신뢰도 제고 방안 / 이태열·황진태·이선주 2015.3
2015-5 2015년 보험소비자 설문조사 / 동향분석실 2015.8
2015-6 인구 및 가구구조 변화가 보험 수요에 미치는 영향 / 오승연·김유미 2015.8

2016-1 경영환경 변화와 주요 해외 보험회사의 대응 전략 / 전용식·조영현 2016.2
2016-2 시스템리스크를 고려한 복합금융그룹 감독방안 / 이승준·민세진 2016.3
2016-3 저성장 시대 보험회사의 비용관리 / 김해식·김세중·김현경 2016.4
2016-4 자동차보험 해외사업 경영성과 분석과 시사점 / 전용식·송윤아·채원영 2016.4
2016-5 금융·보험세제연구: 집합투자기구, 보험 그리고 연금세제를 중심으로 / 정원석·임 준·김유미 2016.5
2016-6 가용자본 산출 방식에 따른국내 보험회사 지급여력 비교 / 조재린·황인창·이경아 2016.5
2016-7 해외 사례를 통해 본 중·소형 보험회사의 생존전략 / 이태열·김해식·김현경 2016.5
2016-8 생명보험회사의 연금상품 다양화 방안: 종신소득 보장기능을 중심으로 / 김세중·김혜란 2016.6
2016-9 2016년 보험소비자 설문조사 / 동향분석실 2016.8
2016-10 자율주행자동차 보험제도 연구 / 이기형·김혜란 2016.9

■ 조사자료집

2014-1 보험시장 자유화에 따른 보험산업 환경변화 / 최 원·김세중 2014.6
2014-2 주요국 내부자본적정성 평가 및 관리 제도 연구 – Own Risk and Solvency Assessment – / 장동식·이정환 2014.8
2015-1 고령층 대상 보험시장 현황과 해외사례 / 강성호·정원석·김동겸 2015.1
2015-2 경증치매자 보호를 위한 보험사의 치매신탁 도입방안 / 정봉은·이선주 2015.2
2015-3 소비자 금융이해력 강화 방안: 보험 및 연금 / 변혜원·이해랑 2015.4
2015-4 글로벌 금융위기 이후 세계경제의 구조적 변화 / 박대근·박춘원·이항용 2015.5
2015-5 노후소득보장을 위한 주택연금 활성화 방안 / 전성주·박선영·김유미 2015.5
2015-6 고령화에 대응한 생애자산관리 서비스 활성화 방안 / 정원석·김미화 2015.5
2015-7 일반 손해보험 요율제도 개선방안 연구 / 김석영·김혜란 2015.12
2018-1 변액연금 최저보증 및 사업비 부과 현황 조사 / 김세환 2018.2

■ 연차보고서

제 1 호 2008년 연차보고서 / 보험연구원 2009.4
제 2 호 2009년 연차보고서 / 보험연구원 2010.3
제 3 호 2010년 연차보고서 / 보험연구원 2011.3
제 4 호 2011년 연차보고서 / 보험연구원 2012.3
제 5 호 2012년 연차보고서 / 보험연구원 2013.3
제 6 호 2013년 연차보고서 / 보험연구원 2013.12
제 7 호 2014년 연차보고서 / 보험연구원 2014.12
제 8 호 2015년 연차보고서 / 보험연구원 2015.12
제 9 호 2016년 연차보고서 / 보험연구원 2017.1
제 10 호 2017년 연차보고서 / 보험연구원 2018.1

■ 영문발간물

제 7 호 Korean Insurance Industry 2008 / KIRI, 2008.9
제 8 호 Korean Insurance Industry 2009 / KIRI, 2009.9
제 9 호 Korean Insurance Industry 2010 / KIRI, 2010.8
제10호 Korean Insurance Industry 2011 / KIRI, 2011.10
제11호 Korean Insurance Industry 2012 / KIRI, 2012.11
제12호 Korean Insurance Industry 2013 / KIRI, 2013.12
제13호 Korean Insurance Industry 2014 / KIRI, 2014.8
제14호 Korean Insurance Industry 2015 / KIRI, 2015.8
제15호 Korean Insurance Industry 2016 / KIRI, 2016.8
제16호 Korean Insurance Industry 2017 / KIRI, 2017.8
제 7 호 Korean Insurance Industry Trend 2Q FY2013 / KIRI, 2014.2
제 8 호 Korean Insurance Industry Trend 3Q FY2013 / KIRI, 2014.5
제 9 호 Korean Insurance Industry Trend 1Q FY2014 / KIRI, 2014.8
제10호 Korean Insurance Industry Trend 2Q FY2014 / KIRI, 2014.10
제11호 Korean Insurance Industry Trend 3Q FY2014 / KIRI, 2015.2
제12호 Korean Insurance Industry Trend 4Q FY2014 / KIRI, 2015.4
제13호 Korean Insurance Industry Trend 1Q FY2015 / KIRI, 2015.8
제14호 Korean Insurance Industry Trend 2Q FY2015 / KIRI, 2015.11
제15호 Korean Insurance Industry Trend 3Q FY2015 / KIRI, 2016.2
제16호 Korean Insurance Industry Trend 4Q FY2015/ KIRI, 2016.6

제17호 Korean Insurance Industry Trend 1Q FY2016/ KIRI, 2016.9
제18호 Korean Insurance Industry Trend 2Q FY2016/ KIRI, 2016.12
제19호 Korean Insurance Industry Trend 3Q FY2016/ KIRI, 2017.2
제20호 Korean Insurance Industry Trend 4Q FY2016/ KIRI, 2017.5
제21호 Korean Insurance Industry Trend 1Q FY2017/ KIRI, 2017.9
제22호 Korean Insurance Industry Trend 2Q FY2017/ KIRI, 2017.11

■ CEO Report

2008-1 자동차보험 물적담보 손해율 관리 방안 / 기승도 2008.6
2008-2 보험산업 소액지급결제시스템 참여 관련 주요 이슈 / 이태열 2008.6
2008-3 FY2008 수입보험료 전망 / 동향분석실 2008.8
2008-4 퇴직급여보장법 개정안의 영향과 보험회사 대응과제 / 류건식·서성민 2008.12
2009-1 FY2009 보험산업 수정전망과 대응과제 / 동향분석실 2009.2
2009-2 퇴직연금 예금보험요율 적용의 타당성 검토 / 류건식·김동겸 2009.3
2009-3 퇴직연금 사업자 관련규제의 적정성 검토 / 류건식·이상우 2009.6
2009-4 퇴직연금 가입 및 인식실태 조사 / 류건식·이상우 2009.10
2010-1 복수사용자 퇴직연금제도의 도입 및 보험회사의 대응과제 / 김대환·이상우·김혜란 2010.4
2010-2 FY2010 수입보험료 전망 / 동향분석실 2010.6
2010-3 보험소비자 보호의 경영전략적 접근 / 오영수 2010.7
2010-4 장기손해보험 보험사기 방지를 위한 보험금 지급심사제도 개선 / 김대환·이기형 2010.9
2010-5 퇴직금 중간정산의 문제점과 개선과제 / 류건식·이상우 2010.9
2010-6 우리나라 신용카드시장의 특징 및 개선논의 / 최형선 2010.11
2011-1 G20 정상회의의 금융규제 논의 내용 및 보험산업에 대한 시사점 / 김동겸 2011.2
2011-2 영국의 공동계정 운영체계 / 최형선·김동겸 2011.3
2011-3 FY2011 수입보험료 전망 / 동향분석실 2011.7
2011-4 근퇴법 개정에 따른 퇴직연금 운영방안과 과제 / 김대환·류건식 2011.8
2012-1 FY2012 수입보험료 전망 / 동향분석실 2012.8
2012-2 건강생활서비스법 제정(안)에 대한 검토 / 조용운·이상우 2012.11
2012-3 보험연구원 명사초청 보험발전 간담회 토론 내용 / 윤성훈·전용식·전성주·

채원영 2012.12
2012-4 새정부의 보험산업 정책(Ⅰ): 정책공약집을 중심으로 / 이기형·정인영 2012.12
2013-1 새정부의 보험산업 정책(Ⅱ): 국민건강보험 본인부담경감제 정책에 대한 평가 / 김대환·이상우 2013.1
2013-2 새정부의 보험산업 정책(Ⅲ): 제18대 대통령직인수위원회 제안 국정과제를 중심으로 / 이승준 2013.3
2013-3 FY2013 수입보험료 수정 전망 / 동향분석실 2013.7
2013-4 유럽 복합금융그룹의 보험사업 매각 원인과 시사점 / 전용식·윤성훈 2013.7
2014-1 2014년 수입보험료 수정 전망 / 동향분석실 2014.6
2014-2 인구구조 변화가 보험계약규모에 미치는 영향 분석 / 김석영·김세중 2014.6
2014-3 『보험 혁신 및 건전화 방안』의 주요 내용과 시사점 / 이태열·조재린·황진태·송윤아 2014.7
2014-4 아베노믹스 평가와 시사점 / 임준환·황인창·이혜은 2014.10
2015-1 연말정산 논란을 통해 본 소득세제 개선 방향 / 강성호·류건식·정원석 2015.2
2015-2 2015년 수입보험료 수정 전망 / 동향분석실 2015.6
2015-3 보험산업 경쟁력 제고 방안 및 이의 영향 / 김석영 2015.10
2016-1 금융규제 운영규정 제정 의미와 시사점 / 김석영 2016.1
2016-3 2016년 수입보험료 수정 전망 / 동향분석실 2016.7
2016-4 EU Solvency Ⅱ 경과조치의 의미와 시사점 / 황인창·조재린 2016.7
2016-5 비급여 진료비 관련 최근 논의 동향과 시사점 / 정성희·이태열 2016.9
2017-1 보험부채 시가평가와 보험산업의 과제 / 김해식 2017.2
2017-2 2017년 수입보험료 수정 전망 / 동향분석실 2017.7
2017-3 1인 1 퇴직연금시대의 보험회사 IRP 전략 / 류건식·이태열 2017.7
2018-1 2018년 수입보험료 수정 전망 / 동향분석실 2018.7
2018-2 북한 보험산업의 이해와 대응 / 안철경·정인영 2018.7

■ Insurance Business Report

26호 퇴직연금 중심의 근로자 노후소득보장 과제 / 류건식·김동겸 2008.2
27호 보험부채의 리스크마진 측정 및 적용 사례 / 이경희 2008.6
28호 일본 금융상품판매법의 주요내용과 보험산업에 대한 영향 / 이기형 2008.6
29호 보험회사의 노인장기요양 사업 진출 방안 / 오영수 2008.6

30호 교차모집제도의 활용의향 분석 / 안철경·권오경 2008.7
31호 퇴직연금 국제회계기준의 도입영향과 대응과제 / 류건식·김동겸 2008.7
32호 보험회사의 헤지펀드 활용방안 / 진 익 2008.7
33호 연금보험의 확대와 보험회사의 대응과제 / 이경희·서성민 2008.9

■ 간행물

○ 보험동향 / 연 4회
○ 보험금융연구 / 연 4회

※ 2008년 이전 발간물은 보험연구원 홈페이지(http://www.kiri.or.kr)에서 확인하시기 바랍니다.

『도서회원 가입안내』

회원 및 제공자료

	법인회원	특별회원	개인회원
연회비	₩300,000원	₩150,000원	₩150,000원
제공자료	- 연구보고서 - 기타보고서 - 연속간행물 · 보험금융연구 · 보험동향 · KIRI 포커스 모음집 · KIRI 이슈 모음집 · KOREA INSURANCE INDUSTRY	- 연구보고서 - 기타보고서 - 연속간행물 · 보험금융연구 · 보험동향 · KIRI 포커스 모음집 · KIRI 이슈 모음집 · KOREA INSURANCE INDUSTRY	- 연구보고서 - 기타보고서 - 연속간행물 · 보험금융연구 · 보험동향 · KIRI 포커스 모음집 · KIRI 이슈 모음집 · KOREA INSURANCE INDUSTRY
	- 영문연차보고서	-	-

※ 특별회원 가입대상 : 도서관 및 독서진흥법에 의하여 설립된 공공도서관 및 대학도서관

가입문의

보험연구원 도서회원 담당
전화 : (02) 3775 - 9080　팩스 : (02) 3775 - 9102

회비납입방법

- 무통장입금 : 국민은행 (400401 - 01 - 125198)
  예금주 : 보험연구원

가입절차

보험연구원 홈페이지(www.kiri.or.kr)에 접속 후 도서회원가입신청서를 작성·등록 후
회비입금을 하시면 확인 후 1년간 회원자격이 주어집니다.

자료구입처

서울 : 보험연구원 자료실 (02-3775-9115 / cbyun@kiri.or.kr)

저 자 약 력

이 항 석

University of Iowa 보험계리학 박사
성균관대학교 보험계리학 교수
(E-mail : hangsuck@skku.edu)

연구보고서 2018-16

빅데이터 분석에 의한 요율산정 방법 비교
: 실손의료보험 적용 사례

발행일	2018년 9월
발행인	한 기 정
발행처	**보 험 연 구 원** 서울특별시 영등포구 국제금융로 6길 38 화재보험협회빌딩 대표전화 : (02) 3775-9000
조판및 인 쇄	고려씨엔피

ISBN 979-11-85691-93-0 94320
979-11-85691-50-3 (세트) 정가 10,000원