디지털산업정보학회 논문지 제20권 제2호 - 2024년 6월
http://dx.doi.org/10.17662/ksdim.2024.20.2.001

특성중요도를 활용한 분류나무의 입력특성 선택효과 : 신용카드 고객이탈 사례

윤 한 성*

Feature Selection Effect of Classification Tree Using Feature Importance : Case of Credit Card Customer Churn Prediction

Yoon Hanseong

〈Abstract〉

For the purpose of predicting credit card customer churn accurately through data analysis, a model can be constructed with various machine learning algorithms, including decision tree. And feature importance has been utilized in selecting better input features that can improve performance of data analysis models for several application areas. In this paper, a method of utilizing feature importance calculated from the MDI method and its effects are investigated in the credit card customer churn prediction problem with classification trees. Compared with several random feature selections from case data, a set of input features selected from higher value of feature importance shows higher predictive power. It can be an efficient method for classifying and choosing input features necessary for improving prediction performance. The method organized in this paper can be an alternative to the selection of input features using feature importance in composing and using classification trees, including credit card customer churn prediction.

Key Words : Classification Tree, Feature Importance, Credit Card Customer, Churn Prediction

Ⅰ. 서론

경쟁이 심화되는 은행분야의 주요 사업영역인 신용카드는 안정성(reliability), 낮은 이자율, 높은 신용한도(credit limit) 등의 금융 편의성을 통해 고객에게 매력을 제공한다[1]. 그런데 발급되는 신용카드 종류 및 수의 증가, 상품과 서비스의 균질화 등으로 빈번해지는 기존 신용카드 고객의 이탈은 중요한 과제로 다루어지고 있으며, 고객 이탈률(churn rate)의 감소는 은행의 이익향상과 밀접한 관계가 있다. 은행이 누적된 고객행동 데이터에 기반하여 효과적인 고객이탈 예측모형을 구성하는 것은 신용카드 고객의 이탈식별 및 고객유지 전략수립을

* 경상대학교 경영대학 교수(단독저자)

위한 신용카드 비즈니스의 중요한 도전과제로 받아들여진다[1, 2].

데이터분석을 통한 신용카드 고객이탈의 예측에는 분류문제에 활용되는 의사결정나무를 비롯하여 다양한 머신러닝 알고리즘으로 모형을 활용할 수 있다[3]. 그리고 이 예측모형(predictive model)의 입력변수로 선택되는 각 특성들의 예측성능에 미치는 영향 정도를 변수중요도(variable importance)로 불리기도 하는 특성중요도(feature importance)로써 측정하기도 한다[4]. 특성중요도는 여러 측정방식으로 계산되는데, 예측모형의 종류에 따라 적절한 방식을 선택할 수 있다. 또한 여러 응용분야에 대해 모형의 성능을 높일 수 있는 최선의 입력특성 선택에 특성중요도를 이용하기도 한다[5-7].

본 논문에서는 분류 의사결정나무, 즉 분류나무로 구성되는 신용카드 고객이탈 예측모형의 특성선택에 특성중요도의 활용방식과 효과를 정리하였다. 정리한 방식을 사례 데이터에 적용하여 모형을 구성하고 가능한 효과를 확인하였다. 본 논문의 방식은 신용카드 고객이탈 예측을 비롯한 분류나무 모형의 구성에 참고하거나 효과적인 대안이 될 수 있을 것이다.

Ⅱ. 이론적 배경

2.1 신용카드 고객이탈 예측과 머신러닝

타 분야와 마찬가지로 은행의 금융상품에서도 제품중심에서 고객중심의 마케팅 방향으로 발전되고 있으며[1], 데이터중심의 산업으로도 간주되는 은행이 잠재적인 이탈고객을 식별하고 고객유지 전략의 수립을 지원할 수 있도록 누적된 고객행동 데이터로부터 효과적인 예측모형을 구성하는 것이 중요하게 다루어진다[2]. 데이터분석을 통한 신용카드 고객이탈의 예측에는 분류문제에 적용가능한 다양한 머신러닝 방식이 활용된다. 이 방식들에는 의사결정나무 분류 또는 로지스틱 회귀, SVM (support vector machine)이나 ANN(artificial neural network), RF(random forest) 및 XGBoost를 포함한 여러 앙상블(ensamble) 알고리즘 등이 열거될 수 있다[8].

신용카드 고객이탈 모형에서 선택되는 입력특성은 주어진 학습데이터, 분석목적이나 방향 등에 따라 다양한 기준과 방식으로 선택된다. 예를 들어 빈도, 시간, 극단(extreme) 정도를 기준으로 하여 이를 나타내는 특성이 선택되기도 하고[9], 분야 전문가의 의견에 따라 특성이 선별되기도 하며[10], 확보된 데이터로부터 기본적인 인구통계학 특성 및 이용금액에 관련된 정보가 활용되기도 한다[11].

2.2 특성중요도와 입력특성 선택

신용카드 고객이탈을 포함한 여러 예측을 위한 모형에서 입력치로 선택되는 다수의 특성(feature)은 각 특성별로 모형의 예측에 독특하게 영향을 끼침으로써 전체 예측결과가 계산된다. 특성중요도는 각 특성별로 계산되며, 모형에 끼치는 각 특성의 영향 정도를 의미한다고 할 수 있다. 특성중요도의 측정방식은 크게 필터(filter) 방식, MDA(mean decrease in accuracy) 방식, MDI(mean decrease in impurity) 방식 등이 있다[4, 12].

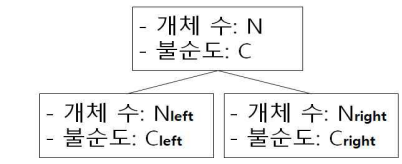
필터 방식은 목표변수(target variable)와 가지는 상관계수 또는 선형회귀계수의 절대값 등으로 특성중요도를 파악하며, 특성중요도의 순위(ranking)를 통해 특성선택(feature selection)으로 이어진다[13]. MDA 방식은 퍼뮤테이션(permutation) 방식으로 특정 한 개의 특성만을 재배열한 데이터에 대한 모형구성 및 정확도 측정을 반복하여, 원래 데이터의 경우에 확인된 정확도와 반복하여 측정한 평균치의 차이를 통해 특성중요도를 계산한다. MDI 방식은 의사결정나무를 구성하여 특성별로 분지변수로 선택되는 경우의 불순도 감소값으로써 특성중요도를 측정한다.

모형에 필요한 입력특성을 선택하는 경우, 필터(filter) 방식 또는 래퍼(wrapper) 방식을 따를 수 있다[13]. 필터

방식에서는 특성중요도 크기의 순위에 따라 정해진 기준을 적용하여 입력특성을 선택할 수 있다. 래퍼 방식은 모든 가능한 입력특성의 조합에 대해 모형의 최고 성능을 보이는 조합을 탐색하여 선택하는 방식이다.

2.3 분류나무와 특성중요도

분류나무는 불순도를 최소화하는 기준으로 노드의 분지(splitting)를 반복하여 구성되는데[14], 분류나무를 구성하는 데이터에 대해서 MDI방식의 특성중요도가 계산된다[15]. 즉, 분류나무에서 분지(splitting) 전후의 불순도(impurity) 감소분을 활용하여 특성중요도가 구해지며, 불순도는 분류나무의 구성에 적용되는 알고리즘에 따라 독특한 방식으로 계산된다. 분류나무 구성의 알고리즘인 CART, C4.5, CHAID 등은 불순도 계산에 각각 지니지수(gini index), 엔트로피(entropy), 카이(Chi) 제곱 통계량을 활용한다[16]. 다진분류가 가능한 C4.5는 과적합의 지적이 있고, CHAID는 명목형 입력변수만이 가능한 제한이 있다. 따라서 본 논문에서는 CART를 통해 분류나무를 구성하기로 한다. 분류나무의 특정 노드에서 분지가 되는 <그림 1>의 사례는 분지전 상위노드와 분지후 2개의 하위노드(left, right)를 나타내며, 각 노드에 포함되는 개체 수와 불순도를 통해 상위노드의 노드별 노드중요도가 계산되고, 계산식에 따라서 특성i의 특성중요도가 구해진다.



- 노드중요도 = $N \cdot C - (N_{left} \cdot C_{left} + N_{right} \cdot C_{right})$
- 특성i의 특성중요도
  = Σi(특성i를 통해 분지된 노드중요도) ÷
  (모든 노드의 노드중요도 합)

〈그림 1〉 분류나무의 분지사례와 특성중요도

<그림 1>의 계산식을 보면, 특성별 특성중요도의 크기는 해당 특성이 분지특성으로 선택되는 횟수와 해당 노드의 노드중요도 크기에 비례하는 값을 가진다. 여기서 노드중요도는 해당 노드의 분지에 따른 정보이득(information gain)[14]을 의미한다. <그림 1>의 특성중요도 값은 분류문제에서 개체들을 보다 명확히, 즉 불순도를 효과적으로 줄이는 특성을 찾고자 하는 의도가 고려되었다고 볼 수 있다. 여러 분류나무로 구성되는 랜덤포레스트(random forest)에서도 세부 분류나무 각각에서 구한 특성중요도 값을 특성별로 합산하여, 분류나무의 수로 나눈 평균값을 특성중요도로 활용하기도 한다[15]. 본 논문에서도 분류나무에 기반하여 신용카드 이탈고객 예측모형을 구성하므로, MDI 방식의 특성중요도를 활용하기로 한다.

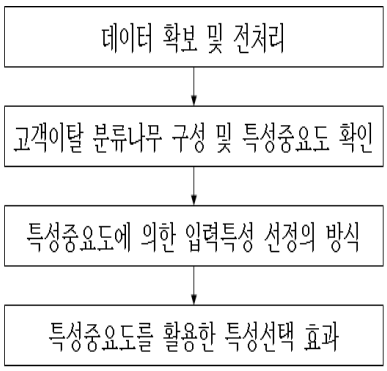
신용카드 고객세분화를 위한 군집분석의 군집유효성을 위해, 분류나무에서 먼저 선택되는 분지특성들을 군집분석의 특성으로 활용하기도 한다[17]. 이와 같은 분류나무의 분지특성 우선순위가 아니라, 본 논문에서는 <그림 1>과 같이 계산되는 특성중요도의 순서에 따르기로 한다. 또한, 군집분석의 분석 데이터는 고객별 특징의 표현에 적절한 특성들로 구성되는데 반해, 본 논문에서는 분류나무 구성이 가능하도록 고객이탈 여부 및 고객이탈에 영향을 미치는 특성들로 구성된다.

Ⅲ. 연구 범위 및 내용

의사결정나무의 분류예측 기능, 즉 일반적인 분류나무를 활용하는 경우 모형의 성능개선을 위해 고려할 수 있는 사항은 (1) 분류나무의 최대깊이(max. depth), (2) 노드별 개체 수의 최소치, (3) 입력특성 선택 등이다. 여기서 (1)과 (2)는 C4.5 또는 CART와 같은 알고리즘 내부에서 처리되는 기능인데 반해, (3)은 기본적인 분류나무 알고리즘과는 별개로 처리되는 영역으로 특성중요도 등을 활용할 수 있다.

CART와 같은 알고리즘을 통해 위 (1)의 '분류나무 최대깊이'의 선택 및 효과의 측정은 용이하지만, (2)의 '노드별 개체 수의 최소치'는 선택의 가변성이 커서 적절한 선택 및 효과의 확인이 쉽지 않다. 따라서 본 논문에서는 특성중요도에 의한 '입력특성 선택'에 대하여 '분류나무 최대깊이'가 가지는 예측성능과 함께, 신용카드 고객이탈 예측모형을 중심으로 래퍼 방식에 따라 랜덤한 경우의 입력특성 조합들과 예측성능을 비교하여 효과를 평가하고 활용방안을 정리하기로 한다.

이를 위해 본 논문의 내용을 <그림 2>와 같이 단계별로 구성하였다. 첫 번째 단계에서는 신용카드 고객이탈 예측모형에 적절한 데이터를 확보 및 필요한 전처리를 수행한다. 두 번째 단계에서는 분류나무를 구성하고, 분류나무의 깊이에 따른 최선의 분류성능을 확인한다. 그리고 구성된 분류나무로부터 특성중요도를 확인한다. 세 번째 단계에서는 특성중요도를 통한 분류나무의 입력특성 선정 및 활용방식을 정리하였다. 마지막 단계에서는 입력특성 선택에 의한 분류나무 구성방식의 효과를 확인하도록 한다. 이와 같은 단계를 통해, 신용카드 이탈고객 예측에 있어서 특성중요도에 의한 입력특성 선택의 활용방식과 효과를 정리하도록 한다.



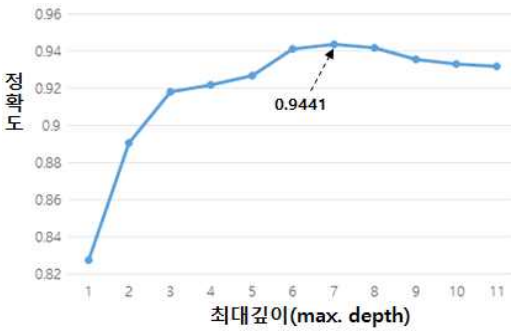
〈그림 2〉 연구의 범위 및 내용

Ⅳ. 특성중요도를 활용한 분류나무의 입력특성 선정

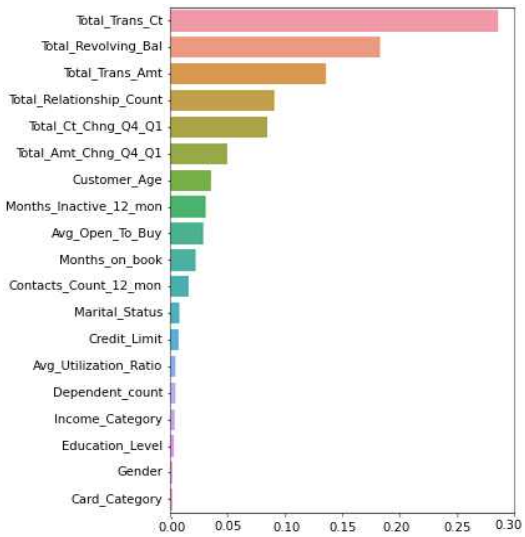
4.1 분류나무 구성 및 특성중요도 확인

깊이(depth)에 제한을 두지 않고 분류나무를 구성하면, 말단노드(terminal node)의 불순도가 0이 되도록 분지가 이루어진다. 이는 과적합(overfitting)으로 인한 모형의 성능감소를 초래하게 된다. 따라서 최선의 깊이가 중요하며, 대개 시행착오를 통해 결정된 최대깊이로써 분류나무가 구성된다. 최대깊이의 변화에 따른 분류나무에서, 평가용 데이터로써 구한 정확도의 변화를 <그림 3>의 사례에서 볼 수 있다. <그림 3>의 경우 깊이 7에서 가장 높은 정확도(0.9441)를 보이며, 최대깊이를 늘여도 정확도는 개선되지 않는다.

분류나무를 통한 특성중요도 계산(<그림 1>)에서 각 노드에 포함된 개체 수가 가중치 역할을 하므로, 상위노드에서 분지특성으로 선택될수록 해당 특성의 특성중요도가 큰 값을 가질 가능성이 크다. <그림 4>는 최대깊이의 제한을 두지 않은 분류나무에서 19개의 입력특성에 대한 특성중요도를 크기 순으로 나열한 것인데, 상위에 위치하는 'Total_Trans_CT'의 특성중요도가 가장 큰 값을 가진다.



〈그림 3〉 최대깊이에 따른 분류나무 정확도 사례



〈그림 4〉 분류나무의 특성중요도 사례

4.2 특성중요도에 의한 입력특성 선택

입력특성의 선택에 래퍼 방식을 적용하면, 전체 특성의 수(n)로부터 i개로 이루어진 입력특성의 가능한 조합의 수는 $\sum_{i=1,n}({}_nC_i)$가 되어 무수히 많아질 수 있다. 따라서 최선의 또는 최적의 입력특성 조합을 발견하기가 쉽지 않으며, 효과적이고 효율적인 입력특성의 선택방식이 요구된다고 볼 수 있다. 본 논문에서는 특성중요도를 활용한 입력특성의 선택 및 효과평가를 다음의 단계로 진행하였다.

⑴ 전체 입력특성(n개)으로 구성한 분류나무에서, 가장 높은 성능(A')의 깊이 D'를 확인한다.

⑵ 전체 입력특성으로써 최대깊이 제한이 없이 구성한 분류나무로부터 특성중요도를 구한다. 특성중요도 값이 큰 순서로 특성을 1개씩 추가하여 구성한 1개부터 n개까지의 특성으로 구성된 특성조합을 구하고, 각 특성조합으로써 구성한 분류나무에 대해 깊이에 따른 최선의 예측성능을 확인한다.

⑶ 위 ⑵에서 구한 예측성능이 (A' - θ)보다 큰 경우의 특성조합으로부터 최대값을 가지는 특성조합을 선택할 수 있으며, 기준치 θ는 0 이상의 작은 값으로 정한다.

⑷ 위 ⑵의 n개 특성의 순서를 랜덤하게 섞은(shuffle) 후, 같은 작업을 m회 반복(기존에 출현한 순서와 동일한 경우는 제외)한다. 적절한 반복횟수 m회를 통해, 특성중요도에 따른 입력특성 조합이 모형의 예측성능을 높이는 적절한 대안임을 확인하고, 그 효과를 평가한다.

위 ⑵의 경우만 보면, 특성을 1개씩 추가하는 필터 방식을 부분적으로 적용한다고 볼 수도 있다. 그렇지만 위 처리과정은 1~n개의 특성을 각각 포함한 입력특성의 조합을 고려하므로, 전체적으로 기본적인 래퍼 방식을 따르는 형식이라고 볼 수 있다. 또한 기준치와 단순한 비교를 통해 입력특성이 정해지는 필터 방식보다, 래퍼 방식을 통해서 특성중요도 효과 및 입력특성 선택의 평가 및 근거가 보다 합리적인 것으로 판단된다.

Ⅴ. 사례 데이터를 통한 적용 및 평가

5.1 사례 데이터 및 분류나무의 구성

신용카드 고객이탈 예측모형의 연구[3]에서 활용된 사례 데이터를 선택하였다. 이 데이터는 캐글 사이트(www.kaggle.com)에 공개되고 있으며, 신용카드 고객에 대해 <표 1>처럼 2015년~2017년간의 인구통계학적 정보, 신용계좌 및 거래내역 정보, 목표특성 등의 21개 특성으로 구성된다. 전체 10,127건의 데이터에서 1,627건이 이탈고객이며, 나머지 8,500건이 유지고객에 해당한다. 'Attrition_Flag'가 목표특성(target feature)이 되며, 고객의 이탈 또는 유지를 의미하는 값을 가진다. 고객ID인 'CLIENTNUM'은 이탈여부 파악에 필요하지 않으므로, 이를 제외한 19개 특성을 입력특성으로 활용할 수 있다.

〈표 1〉 사례 데이터에 포함된 특성

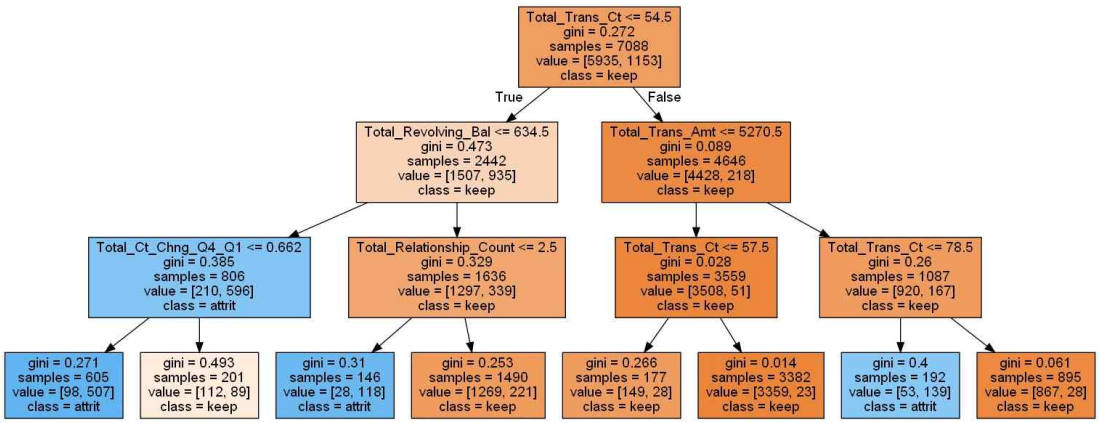
특성 이름	내용
CLIENTNUM	신용카드 보유자(고객) ID
Customer_Age	고객의 나이
Gender	고객의 성별
Dependent_count	부양가족 수
Education_Level	학력
Marital_Status	결혼상태
Income_Category	수입등급
Card_Category	신용카드 등급
Months_on_book	가입기간(개월 수)
Total_Relationship_Count	총 관계거래 건수
Months_Inactive	거래 미발생 개월 수
Contacts_Count	접촉(연락) 건수
Credit_Limit	현재 신용한도
Total_Revolving_Bal	총 지불연장 금액
Avg_Open_To_Buy	평균 신용한도
Total_Amt_Chng_Q4_Q1	거래금액의 변화율
Total_Trans_Amt	총 거래금액
Total_Trans_Ct	총 거래건수
Total_Ct_Chng_Q4_Q1	거래건수의 변화율
Avg_Util_Ratio	평균 이용율
Attrition_Flag	이탈 표시

준비된 데이터에 대해 7:3의 비율로 랜덤하게 나누어 훈련용 및 평가용 데이터에 할당하고, 훈련용 데이터로써 CART방식의 분류나무를 구성할 수 있다. 최대깊이를 증가시키면서 분류나무를 구성할 수 있는데, 최대깊이가 3인 분류나무는 <그림 5>와 같다.

5.2 특성중요도에 의한 입력특성 선택

준비한 데이터에 대해, 4.2절에서 정리한 (1)~(4)단계에 따라 진행하기로 한다. 먼저 (1)단계에서 <표 1>로부터 선택한 19개의 입력특성(n=19) 모두에 대해 최대깊이를 1부터 증가 시켜가면서 분류나무를 구성하였다. 여기서 분류나무의 가장 보편적인 예측성능이기도 한 정확도를 각 분류나무에서 평가용 데이터로부터 구할 수 있다. 최대깊이가 3인 <그림 5>의 분류나무는 정확도가 <그림 6>에서 0.9181이며, 최대깊이가 7(D')일 때에 정확도의 최대값(A')은 0.9441이 된다.

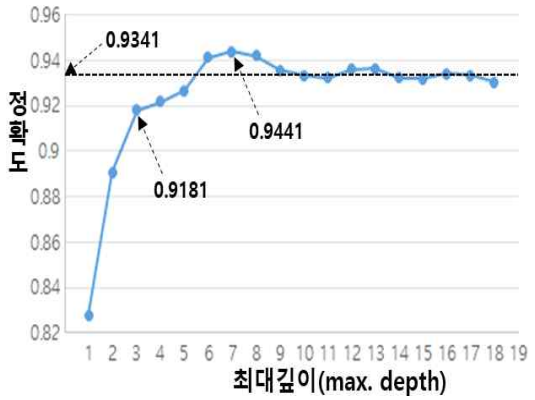
(2)단계에서는 최대깊이의 제한이 없이 구성한 분류나무로부터 구한 특성중요도의 크기순으로 특성 1개씩 추가하여 특성조합을 먼저 구성한다. 사례의 데이터로부터 구한 특성중요도 크기순의 특성은 <그림 4>와 같이 나열된다. 특성중요도 크기순의 Total_Trans_Ct, Total_Revolving_Bal, Total_Trans_Amt, … , Education_Level, Gender, Card_Category에 대하여, (Total_Trans_Ct), (Total_Trans_Ct, Total_Revolving_Bal), (Total_Trans_Ct, Total_Revolving_Bal, Total_Trans_Amt), … , (Total_Trans_Ct, T



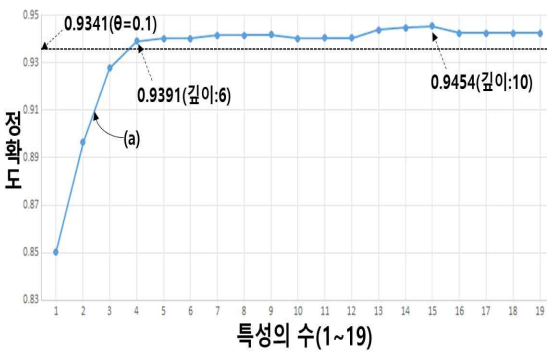
〈그림 5〉 최대깊이 3으로 구성한 분류나무 사례

otal_Revolving_Bal, … , Education_Level, Gender, Card_Category) 등과 같이 19개의 특성조합을 구할 수 있다. 각 특성조합에 대하여 최대깊이를 늘여가면서 분류나무를 구성하고, 예측성능을 정확도로써 확인하여 <그림 7>의 (a)로 나타내었다.

⑶단계에서는 앞에서 구한 정확도 A'(0.9441)에 대해 기준치 θ를 0.01로 정할 수 있다. 그러면 <그림 7>의 정확도가 0.9341(=0.9441 - θ)보다 큰 경우에서는, <그림 4>의 'Dependent_count'까지 포함된 입력특성의 수가 15개이며 최대깊이가 10일 때에 정확도는 0.9454의 최대값을 가진다. 이 값은 (1)단계에서 구한 정확도 A'에 비해, 정확도의 증가는 0.0013(=0.9454-0.9441)에 불과하다.

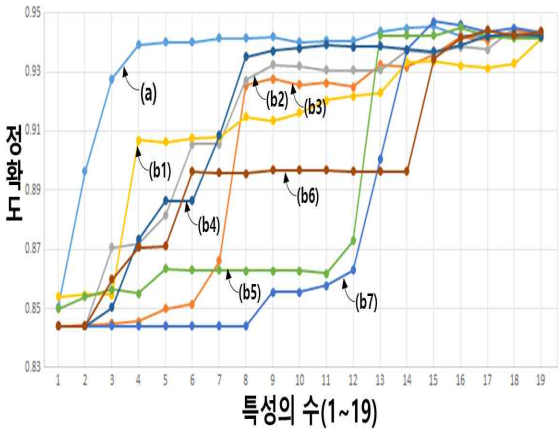


〈그림 6〉 최대깊이에 따른 분류나무 정확도 사례



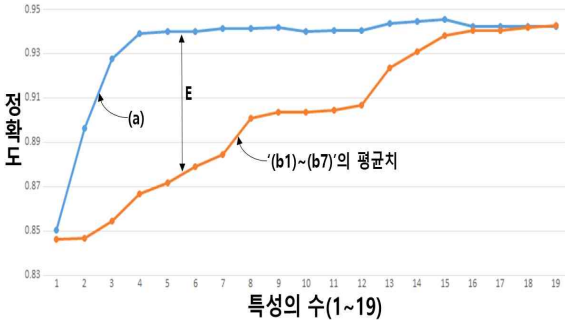
〈그림 7〉 특성중요도 순의 입력특성 수와 정확도

⑷단계에서는 반복횟수 m=7로 하고, 각 반복 수행에서는 19개의 특성을 랜덤(random)한 순서로 정렬하였다. 여기서 1개씩의 특성을 선택하여 (2)단계와 같은 방식의 분류나무 19개를 구성하였으며, 각각에서 깊이의 변화에 따라 가장 높은 정확도를 확인하였다. 그리고 <그림 7>의 특성중요도 순서와 겹치지 않고 반복되지 않은 순서가 되도록, 7회 수행하여 <그림 8>의 (b1)~(b7)과 같이 정확도를 확인하였다. 7회 랜덤순서를 반복하여 구성한 특성조합에서는, 특성중요도의 순서에 의한 경우보다 모든 경우에서 정확도가 훨씬 낮고 등락폭도 크다.



〈그림 8〉 입력특성 추가순서와 정확도

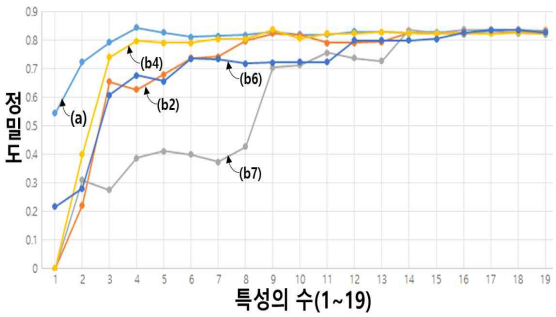
<그림 7>에서 θ값(0.1)을 모형선택의 기준치로 본다면, 특성중요도 순의 분류나무에서 Total_Relationship_Count까지 포함된 특성의 수가 4개만 되어도 우수한 정확도(0.9391)를 보인다. 특성의 수가 15개 이하에서는, 특성중요도의 순서대로 입력특성을 구성하는 방식이 <그림 9>의 'E' 만큼 정확도 개선효과를 가져오는 것으로 볼 수 있다. <그림 9>는 <그림 8>의 (b1)~(b7)에 대해 특성의 수별 정확도의 평균값을 구하여, 특성중요도 순서에 의한 (a)와 비교한 내용이다. 그리고 16개 이상의 입력특성이 포함되면, 랜덤한 순서의 입력특성 구성에 비해 정확도에서 차이가 거의 없으므로 전체 특성에서 나머지 4개를 제외하여도 모형의 예측성능에는 무방해 보인다.



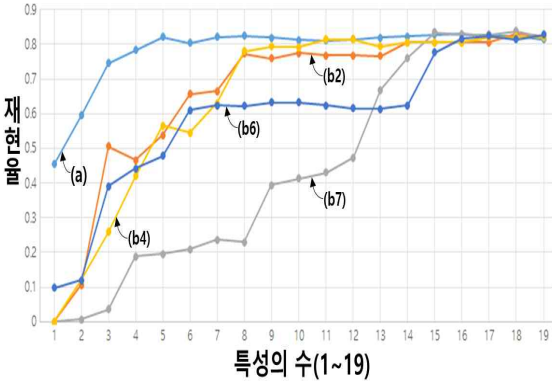
〈그림 9〉 분류나무 구성에 따른 정확도의 차이

한편 19개 전체 특성에 대해 최대깊이 7로 구성한 경우와 특성중요도 순으로 구성한 경우, 정확도는 각각 0.9441(<그림 6>)과 0.9454(<그림 7>)로서 차이가 미미하다. 이는 특성중요도에 의한 입력특성 선택이 최대깊이 조정을 통한 분류나무와 비교하여, 정확도를 획기적으로 개선하기는 쉽지 않다는 점을 보여준다. 그러나 높은 정확도를 제공하는 입력특성의 효과적인 선택과 이를 통한 효율적이고 간결한 모형의 구성에 좋은 대안일 수 있다.

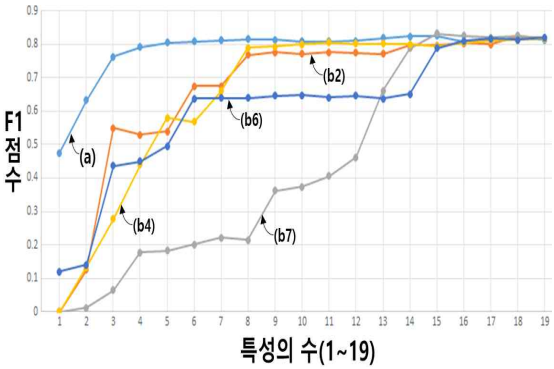
정확도 이외에, 분류나무의 성능평가에 활용되는 정확도(precision), 재현율(recall) 및 F1-점수에 대해 <그림 8>과 같은 형식으로 특성중요도의 순서에 의한 분류나무(a)와 랜덤한 순서로 구성한 4건의 분류나무를 비교하였다. 정확도의 경우와 마찬가지로 <그림 10>~<그림 12>에서 특성중요도 순서에 의한 분류나무가 우수한 예측성능을 보이며, 또한 16개 이상의 입력특성들에 대해서는 예측성능이 거의 같아지는 형태를 보인다.



〈그림 10〉 최대깊이와 분류나무 정밀도 사례



〈그림 11〉 최대깊이와 분류나무 재현율 사례



〈그림 12〉 최대깊이와 분류나무 F1-점수 사례

Ⅵ. 결론 및 토의

빅데이터 분야의 활용이 확대되어 가고 있고, 성능의 향상을 위한 적용방식도 다양해지고 있다[14, 18, 19]. 머신러닝의 예측모형 및 응용범위에 적절한 입력특성의 선택에 특성중요도를 고려하는 것도 이러한 일환의 하나인 것으로 볼 수 있다. 본 논문에서는 신용카드 고객이탈 예측을 위한 분류나무에서 MDI 방식의 특성중요도를 활용하여 입력특성의 선택방식을 구성해보았다. 구성한 방식에 대하여 랜덤하게 선택한 입력특성의 분류나무 및 일반 CART방식의 분류나무와 예측성능의 비교를 통해 효과를 확인하였다.

사례 데이터로부터 다수의 랜덤한 특성선택과 비교하여, MDI방식의 특성중요도 순서를 통해 우수한 예측력을 보이는 특성들을 확인하여 선택할 수 있었다. 또한 예측성능의 최대화에 필요한 입력특성들을 구분하여 선택하는데 효율적인 방안이 될 수 있을 것으로 본다. 그런데 일반적인 CART 방식의 최대깊이를 조정한 분류나무에 비해 예측성능을 개선하는 효과는 크지 않은 것으로 확인되었다. 이와 같은 방식은 데이터 확보가 쉽지 않은 실무적 여건에서 입력특성의 효과적인 선별적 선택에 도움이 될 수 있다. 또한 모델의 적절한 예측성능에 필요한 입력특성 조합을 효과적으로 선택할 수 있는 이론적인 수단으로도 활용이 가능할 수 있다.

본 논문에서 정리한 방식은 신용카드 고객이탈 예측을 비롯한 분류나무의 구성과 활용에 있어서, 특성중요도를 활용한 입력특성의 선택에 참고와 대안이 될 수 있을 것으로 생각된다. 본 논문에서는 한정된 사례 데이터에 적용한 효과를 확인하였으므로, 일반화하기에는 성급한 측면이 있다. 이에 대해서는 다양한 적용을 통한 경험적 타당성의 확보 또는 수치적인 타당성 분석 등의 추가적인 연구가 필요할 것으로 보인다.

참고문헌

[1] Larivi'ere, B., & Van den Poel, D. “Investigating The Role of Product Features in Preventing Customer Churn by Using Survival Analysis and Choice Modeling: The Case of Financial Services,” Expert Systems with Applications, Vol.77, No.2, 2004, pp.277–285.

[2] Devriendt, F. et al., “Why You Should Stop Predicting Customer Churn And Start Using Uplift Models,” Information Sciences, Vol.548, 2021, pp.497–515.

[3] Rao, C. et al., “Imbalanced Customer Churn Classification Using A New Multi-Strategy Collaborative Processing Method,” Expert Systems With Applications, Vol.247, 2024, 123251.

[4] Oh, Sejong, “Predictive Case-Based Feature Importance And Interaction,” Information Sciences, Vol.593, 2022, pp.155–176.

[5] Wang, X. et al., “Aircraft Taxi Time Prediction: Feature Importance And Their Implications,” Transportation Research Part C, Vol.124, 2021, 102892.

[6] Kanyongo, W. and Ezugwu, A.E., “Feature Selection And Importance of Predictors of Non-Communicable Diseases Medication Adherence from Machine Learning Research Perspectives,” Informatics in Medicine Unlocked, Vol.38, 2023, 101232.

[7] Heidari, M. et al., "Forward Propagation Dropout in Deep Neural Networks Using Jensen–Shannon And Random Forest Feature Importance Ranking," Neural Networks, Vol.165, 2023, pp.238–247.

[8] Tekouabou, S. C. K. et al., “Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing And Ensemble-Based Methods,” Mathematics, Vol.10, No.14, 2022, 2379.

[9] Nie, Guangli et al., “Credit Card Churn Forecasting by Logistic Regression And Decision Tree,” Expert Systems with Applications, Vol.38, 2011, pp.15273–15285.

[10] Lin, Chiun-Sin et al., “Combined Rough Set Theory And Flow Network Graph to Predict Customer Churn in Credit Card Accounts,” Expert Systems with Applications, Vol.38, 2011, pp.8–15.

[11] 이건창·정남호·신경식, “신용카드 시장에서 데이터 마이닝을 이용한 이탈고객 분석,” 지능정보연구,

Vol.8, No.2, 2002, pp.15-35.

[12] 윤태균·이관수, “의료진단 및 중요 검사 항목 결정 지원 시스템을 위한 랜덤 포레스트 알고리즘 적용,” 전기학회논문지, 제57권, 제6호, 2008, pp.1058-1062.

[13] Xiang, F. et al., “Ensemble Learning-Based Stability Improvement Method for Feature Selection Towards Performance Prediction,” Journal of Manufacturing Systems, Vol.74, 2024, pp.55 - 67.

[14] 윤한성, “의사결정나무를 활용한 신경망 모형의 입력특성 선택: 주택가격 추정 사례,” 디지털산업정보학회 논문지, 제19권, 제1호, 2023, pp.109-118.

[15] Dunne, Robert et al., “Thresholding Gini Variable Importance with A Single-Trained Random Forest: An Empirical Bayes Approach,” Computational and Structural Biotechnology Journal, Vol.21, 2023, pp.4354 - 4360.

[16] Chanmee, Sirichanya and Kesorn, Kraisak, “Semantic Decision Trees: A New Learning System for The ID3-Based Algorithm Using A Knowledge Base,” Advanced Engineering Informatics, Vol.58, 2023, 102156.

[17] 윤한성, “분류나무를 활용한 군집분석의 입력특성 선택: 신용카드 고객세분화 사례,” 디지털산업정보학회 논문지, 제19권, 제4호, 2023, pp.1-11.

[18] 김동형, “이미지 보간을 위한 의사결정나무 분류 기법의 적용 및 구현,” 디지털산업정보학회 논문지, 제16권, 제1호, 2020, pp.55-65.

[19] 정병호, “빅데이터 분류 기법에 따른 벤처 기업의 성장 단계별 차이 분석,” 디지털산업정보학회 논문지, 제15권, 제4호, 2019, pp.197-212.

■ 저자소개 ■

윤 한 성
(Yoon Hanseong)

2001년 3월~현재
경상대학교 경영대학 교수
1998년 8월 한국과학기술원
테크노경영대학원(공학박사)
1987년 8월 한국과학기술원
산업공학과(공학석사)
1985년 2월 서울대학교 산업공학과(공학사)

관심분야 : 디지털비즈니스, 공급망관리, 데이터분석 등
E-mail : hsyun@gnu.ac.kr

논문접수일 : 2024년 3월 25일
수정접수일 : 2024년 4월 10일
게재확정일 : 2024년 4월 28일