디지털산업정보학회 논문지 제19권 제4호-2023년 12월
http://dx.doi.org/10.17662/ksdim.2023.19.4.001

분류나무를 활용한 군집분석의 입력특성 선택: 신용카드 고객세분화 사례

윤 한 성*

Classification Tree-Based Feature-Selective Clustering Analysis: Case of Credit Card Customer Segmentation

Yoon Hanseong

〈Abstract〉

Clustering analysis is used in various fields including customer segmentation and clustering methods such as k-means are actively applied in the credit card customer segmentation. In this paper, we summarized the input features selection method of k-means clustering for the case of the credit card customer segmentation problem, and evaluated its feasibility through the analysis results. By using the label values of k-means clustering results as target features of a decision tree classification, we composed a method for prioritizing input features using the information gain of the branch. It is not easy to determine effectiveness with the clustering effectiveness index, but in the case of the CH index, cluster effectiveness is improved evidently in the method presented in this paper compared to the case of randomly determining priorities. The suggested method can be used for effectiveness of actively used clustering analysis including k-means method.

Key Words : k-Means, Decision Tree Classification, Input Feature Selection, Customer Segmentation

Ⅰ. 서론

매력적인 상품과 서비스를 제공하는 경쟁사로의 빈번한 고객전환(customer switching) 및 심화하는 시장경쟁에 따라, 금융시장에서 마케팅 비용부담이 증가하고 있다[1]. 따라서 신용카드를 비롯한 금융서비스 분야의 기업이 고객유지를 위해 개별 고객 또는 고객 그룹들을 식별하여 각각에 차별화된 가치를 창출하여 제공하는 것이 중요한데, 이는 고객세분화(customer segmentation)의 필요성과 목적이 될 수 있다[2]. 또한 고객세분화는 경쟁사로 이탈하는 신용카드 고객의 평생가치 위험을 줄이는 마케팅 활동 또는 고객가치 분석의 첫 단계일 수 있다[2, 3].

데이터를 통한 신용카드 고객의 세분화는 대개 k-평균과 같은 군집분석으로 이루어지는데[4-6], 분석목적이나 필요에 따라 데이터의 일부 특성을 선별적으로 선택할 수 있다. 군집분석을 포함한 데이터마이닝 분야에서 전체 데이터가 아니라 특성선택을 통한 선별적 데이터의 분석으로 개선된 성능의 모델을 구성

* 경상대학교 경영대학 교수(단독저자)

할 수 있다[6]. 군집분석의 성능개선을 위한 특성선택은 여러 방식으로 이루어질 수 있으나[6, 7], 본 논문에서는 의사결정나무 분류를 활용하고자 한다.

의사결정나무 분류(decision tree classification)의 경우, 인공신경망과 보완적인 또는 결합한 형태의 성능개선 분석방식이 연구되고 있으나[8, 9] k-평균을 포함한 군집분석의 입력특성 선택을 위해 모델이 구성되는 연구사례는 찾기가 어렵다. 본 논문에서는 이러한 의사결정나무 분류를 군집분석의 특성선택에 활용하는 방안을 정리하고자 하며, 신용카드 고객세분화 사례에 적용하고 효과를 확인하고자 한다.

II. 이론적 배경

2.1 신용카드 고객세분화와 k-평균

고객세분화는 큰 고객그룹에 대해 사회적, 행위적 및 소비적 특성 등에 기반하여 특성값이 유사한 고객들끼리 구성된 여러 작은 고객그룹들로 분리하는 것을 의미한다[10]. 고객세분화를 통해 분리된 고객 그룹별 특징을 식별하는 것은 고객의 소비행위와 선호도를 명확히 하여 고객 그룹별 차별화된 상품과 서비스 제공을 지원할 수 있다[5, 11].

신용카드 분야의 고객세분화에서도 여러 특성들이 선택되어 활용된다. 예를 들어 고객의 가입정보, 신용카드 사용정보, 신용정보 등이거나[4], 고객의 이용서비스영역, 이용업종, 이용시간대 등이기도 하고[5], 이용업종, 이용시간대, 이용지역, 거래실적, 할부실적, 현금서비스실적, 인구통계학적 변수 등이 선택되기도 한다[7]. 그리고 신용카드 분야에서 활발히 이루어지는 고객세분화 방식으로 k-평균 분석이 폭넓게 활용된다[2, 4, 7, 12, 13].

k-평균 방식은 비지도 학습(unsupervised learning) 알고리즘으로서 목표특성(target feature)이 요구되지 않고 대량의 데이터를 빠르게 처리하는 장점이 있다. 군집의 수를 의미하는 k는 사전에 정하는 것이 필요하며, 적절한 k의 값은 흔히 실루엣(silhouette) 값 등을 통해 결정된다[14]. k-평균 방식은 분석할 전체 데이터 세트(data set)를 미리 정해진 k개의 군집으로 나누는 다음의 단계들로 처리되며[15], 군집분석의 결과인 k개의 군집에 대해 각 군집의 특징을 파악하는 과정이 추가로 이루어진다.

(1) 데이터 세트에서 무작위로 선택한 k개의 개체를 각각 k개의 군집(cluster) 중심으로 선정

(2) 각 데이터 개체를 유클리디언(Euclidean) 거리가 가장 가까운 군집 중심에 할당

(3) 현재의 군집에 할당된 데이터 개체들에 대해 군집의 새로운 중심을 재계산

(4) 새로운 군집 중심이 수렴(convergence) 기준을 충족하지 않으면, 단계 (2)로 이동

k-평균은 기본적으로 수치형 특성의 유클리디언 거리로 계산되는 유사성(similarity)을 통해 이루어지며, 이론적인 명료성(simplicity)과 계산의 신속성(speed)의 장점을 가진 것으로 평가된다[16]. k-평균 방식은 목표특성(target feature)이 요구되지 않고 데이터 개체들이 가지는 특성값의 유사성으로 계산되므로, 군집화 성능은 분석에 사용되는 특성에 많이 의존하게 된다[6]. 이때 지나치게 많은 수의 특성은 군집분석의 성능에 부정적일 수가 있고, 범주형 특성이 포함되면 k-평균에 필요한 수치형 값으로 변환 및 유사성 산정에 어려움이 있다[17].

k-평균을 포함한 군집분석에 대한 결과의 평가는 흔히 군집유효성지수(clustering validity index)로 불리는 측정치를 활용한다[18]. 군집분석 결과인 여러 군집별로 동일 군집에 속한 개체 간의 유사도가 클수록 응집도가 높고, 다른 군집에 속한 개체 간의 유사도가 클수록 분리도가 낮은데, 대개 응집도와 분리도의 비율로써 군집유효성지수를 계산한다. 이때 응집도와 유사도의 계산방식 등에 따라 여러 군집유효성

지수가 존재한다. <표 1>의 DB지수(Davies-Bouldin score)와 CH지수(Calinski-Harabasz score)가 사용 빈도가 높고 우수하다고 알려져 있으며[18], 본 논문에서도 활용하기로 한다.

<표 1> DB지수 및 CH지수[18]

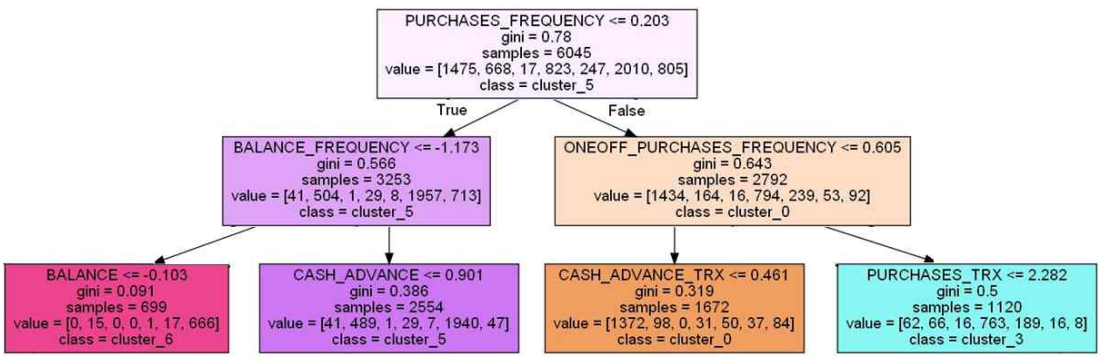
종류	계산식 및 설명
DB 지수	$DB=\frac{1}{K}\sum_{i=1}^{K}\max_{i=1,\cdots,K,i\neq j}\left\{\left(\sqrt{\frac{1}{n_i}\sum_{x\in c_i}d(x,z_i)^2}+\sqrt{\frac{1}{n_j}\sum_{y\in c_j}d(y,z_j)^2}\right)/d(z_i,z_j)\right\}$ (n_i: 군집 C_i의 개체 수)
	• {군집 내 중심과 개체 간 거리(응집도)의 합} ÷ {군집들의 중심점 간 거리(분리도)} • DB지수값이 작을수록 군집분석 결과가 우수
CH 지수	$CH=\sum_{i=1}^{K}n_i\cdot d(z_i,z_{tot})^2/(K-1)\cdot (N-K)/\sum_{i=1}^{K}\sum_{x\in c_i}d(x,z_i)^2$ • N: 전체 데이터의 개수 • K: 군집 수 • z_i: 군집 C_i의 중심점 • z_{tot}: 전체 데이터의 중심점
	• {군집 중심점과 전체 중심점 간 거리(분리도)} ÷ {군집 내 개체와 중심점 간 거리(응집도)} • CH지수값이 클수록 군집분석 결과가 우수

2.2 k-평균 입력특성 선택과 의사결정나무

분석모델의 입력특성(input feature) 선택은 데이터마이닝 분야의 중요한 주제 중 하나이며, 모델의 성능개선을 위해 가용한 전체 특성의 부분집합(subset)을 발견하는 차원축소(dimensionality reduction) 문제를 의미한다[6]. 낮은 차원의 데이터를 통한 과적합 위험의 감소, 복잡성 감소를 통한 모델의 일반화 수준 향상 등도 특성선택의 결과인 적은 수의 특성으로 이루어지는 분석의 장점일 수 있다[19, 20].

군집분석에서도 전체 데이터가 아니라 선별적 특성을 통해 분석 결과의 개선을 도모하는데[6], 군집분석을 통한 고객세분화에서 분야의 경험적 특성으로써 적절히 체계화된 입력특성의 선택[7]이 그 사례가 될 수 있다. 미리 정한 기준으로 특성을 평가하여 선택하거나 또는 가능한 모든 특성의 조합에 대해 군집화 영향을 평가하여 특성을 선택할 수도 있는데[6], 이러한 경우 분류분석에 비해 군집분석 결과의 정확도에 대한 객관적 측정방식이 상대적으로 부족하다는 지적이 있으며[6, 21] 선택할 특성의 효과적 탐색이 어려운 것으로 판단되기도 한다[21].

한편 의사결정나무는 분석할 집단에 대해 선택한 특성의 특정 값을 기준으로 분리하는데, 분리된 그룹별 개체들의 목표특성에 대한 불순도(impurity)가 최소화되도록 분지(splitting) 과정을 반복하여 트리(tree) 형태의 분류모형을 구성한다[9]. 의사결정나무의 구성에는 CART(Classification and Regression Trees), CHAID(CHi-square Automatic Interaction Detectio



<그림 1> 의사결정나무 분류 사례

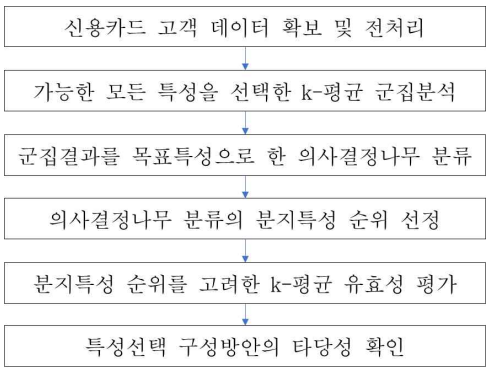
n), C4.5 등의 알고리즘이 활용된다. 의사결정나무 분류의 분지특성이나 말단노드로 분류된 레이블(label) 값이 인공신경망의 입력특성으로 활용되기도 하며[8, 22, 23], 의사결정나무 분지의 정보이득(inforrmation gain) 효과가 큰 특성이 인공신경망의 입력특성으로 선택되기도 한다[9]. 즉, <그림 1>과 같이 깊이가 3인 의사결정나무에서 상위에 존재하여 정보이득이 큰 PURCHASES_FREQUENCY, BALANCE_FREQUENCY, ONEOFF_PURCHASES_FREQUENCY 등의 순서로 입력특성을 선택하여 인공신경망의 예측력을 개선한다는 것이다. 이와 같은 의사결정나무 분지를 활용한 입력특성 선택의 방식을 k-평균과 같은 군집분석에서 이용할 수 있는 방안을 본 논문에서 정리하고자 한다.

Ⅲ. 연구 범위 및 내용

k-평균과 같은 군집분석의 입력특성 선택을 위해 의사결정나무의 분지특성을 활용하는데 있어서, 고려할 수 있는 사항으로 군집분석과 의사결정나무 분류분석에서 서로 공유하는 방향성을 가지도록 하는 것이다. k-평균의 분석방향은 유사성 기반의 k개의 군집화로 두고, 의사결정나무 분류는 k개의 군집분석 결과를 분류의 목표특성으로 하는 것을 두 분석방식의 방향성 공유로 생각할 수 있다.

의사결정나무 분류에서 먼저 선택되는 분지변수는 정보이득이 크므로 분류효과가 크다고 평가할 수 있고, 분지변수로 선택되지 못하거나 늦은 순서의 깊이(depth)에서 선택되는 특성은 전체 데이터에 대한 분류효과가 적은 것으로 판단할 수 있다. 이 두 가지 경우에서 첫 번째의 경우는 특성을 선택하는 우선순위로써 활용할 수 있고, 두 번째는 모든 특성이 선택된 상황에서 비효율적인 특성으로 선택하여 제외하는 순서로 고려할 수 있다.

이와 같이 분석방향의 공유, 특성의 선택 또는 제외하는 측면들을 고려하여 본 논문의 연구범위를 <그림 2>와 같이 단계별 내용으로 구성하였다. 첫 번째 단계에서는 신용카드 고객세분화에 활용할 가치가 있는 데이터를 확인 및 수집, 그리고 필요한 전처리를 수행한다. 본 논문에서 활용할 모형은 의사결정나무와 k-평균 모형이므로, 결측 데이터의 처리나 데이터값의 정규화 등이 필요하다. 두 번째 단계에서는 적절한 군집의 수 k를 결정하고 이에 따른 k-평균 분석을 한다. 분석결과인 군집 레이블(label)을 추가한 고객 데이터로써 세 번째 단계인 의사결정나무 분류를 처리한다. 이때 목표특성으로 군집 레이블 값을 활용한다. 네 번째 단계에서는 의사결정나무의 분지특성 선택과 분지에 따른 불순도(impurity)에 따라 k-평균 입력특성 순위를 정하는 기준을 제안한다. 다음 단계에서는 기준에 따라 선택한 분지특성으로 k-군집의 처리 및 군집결과 유효성 평가를 진행하고, 최종 단계에서는 앞 단계의 k-평균 군집의 입력특성 선택 방식에 대해 타당성을 확인하기로 한다.



<그림 2> 연구 범위 및 내용

Ⅳ. 의사결정나무 분류를 활용한 k-평균 입력특성 선택

4.1 데이터 및 의사결정나무 구성

고객세분화를 위한 k-평균은 고객 간 수치형 속성 데이터로 계산되는 거리유사도로써 분석되므로, 이에 필요한 고객 데이터는 기본적으로 수치형 값을 가진다. 고객세분화를 위한 k-평균의 입력특성 선택에 활용할 의사결정나무는 분류의 목표값이 필요하며, 본 논문에서는 k-평균과 의사결정나무 분류가 공유하는 분석방향으로 k개의 군집결과를 의사결정나무 분류의 목표특성으로 구성하였다. 이를 위해서는 의사결정나무의 구성에 앞서 k-평균 군집화가 필요한데, 이를 포함한 의사결정나무의 구성은 다음 순서에 따라 이루어졌다.

(1) 실루엣(silhouette) 또는 엘보우(elbow) 방법 등을 통하여 군집의 수(k)를 결정하고[14], 개별 군집에 포함된 고객별 데이터에 해당 군집의 레이블 값을 포함시킨다.

(2) 개별 군집을 의미하는 군집 레이블 값을 목표변수로 하여 의사결정나무를 구성한다. 이때 분지의 깊이를 충분히 하여, 가능한 한 모든 특성이 분지특성으로 선택될 수 있도록 한다. 의사결정나무는 <그림 1>의 사례와 같이 구성될 수 있는데, <그림 1>의 각 노드에서 분류되는 군집의 레이블 값(cluster_0, cluster_3 등)을 확인할 수 있다.

4.2 k-평균 군집의 입력특성 선택

군집 레이블 값을 목표특성으로 하여 정보이득을 최대화하도록 분지가 이루어지는 의사결정나무의 분지특성은 다음의 특성을 가진다고 할 수 있다. 첫 번째로 의사결정나무 분지에 보다 늦게 선택되어 의사결정나무에 위치하는 깊이(depth)가 깊은 특성일수록 의사결정나무의 전체 불순도를 낮추는 효과가 작다. 이는 의사결정나무 분류에서 분지특성이 가지는 일반적인 특성이다. 두 번째로 늦게 선택되는 분지특성은 개체 수가 상대적으로 적은 하위의 노드에서만 정보이득효과를 가지는 과적합(overfitting)의 가능성이 커진다는 것이다. 이러한 분지특성은 전체 데이터에 대한 분류효과가 크지 않다고 볼 수 있다.

이상의 두 가지 특성을 고려하여, 군집 레이블 값을 목표특성으로 하는 의사결정나무의 분지특성으로부터 k-평균 군집화의 입력특성 선택기준을 구성할 수 있다. 본 논문에서는 전체 특성으로부터 군집유효성 효과가 적을 것으로 고려되는 특성을 제외하는 방식으로 다음과 같이 구성하였다.

(1) 상위 계층의 분지특성을 먼저 선택하고, 동일한 깊이의 분지특성들에 대해서는 불순도가 낮은 그룹의 분지특성에 우선순위를 부여하여 선택
(2) 이미 선택된 분지특성은 무시하고, 모든 특성이 선택되도록 깊이를 낮춰가며 (1)을 반복
(3) 모든 특성에 우선순위가 부여되면, 위 (1) 및 (2)를 멈춤.
(4) 우선순위가 낮은 특성을 제외해가면서, 군집효율성을 고려하여 k-평균의 입력특성 선택

<그림 1>의 의사결정나무에 위 과정을 적용하면, PURCHASES_FREQUENCY→BALANCE_FREQUENCY→ONEOFF_PURCHASES_FREQUENCY→BALANCE→CASH_ADVANCE_TRX→CASH_ADVANCE→PURCHASES_TRX로 7개의 특성에 대해 우선순위가 만들어진다. 우선순위가 낮아서 군집유효성 효과가 작을 것으로 판단되는 PURCHASES_TRX, CASH_ADVANCE, CASH_ADVANCE_TRX, BALANCE 등의 순서로 전체 7개의 특성으로부터 1개씩 빼가면서, 나머지 특성들에 대하여 군집효율성을 고려하여 입력특성으로 선택할 수 있을 것이다. 이때 군집분석의 최소 입력특성의 수(N)가 필요할 수 있으며, 만일 N=4이라면 PURCHASES_TRX→CASH_ADVANCE→C

ASH_ADVANCE_TRX까지의 순서로 각 특성이 제외된 나머지 4개까지 입력특성으로 선택될 수 있을 것이다.

4.3 입력특성 선택 및 k-군집 유효성

k-평균 군집분석의 결과에 대해 군집유효성지수로써 평가할 수 있다[18]. 분류분석에 비해 군집분석 결과의 정확도에 대한 객관적 측정방식이 부족하다는 지적은 있으나[6, 21], 본 논문에서는 제안되고 있는 기존의 여러 군집유효성 지수중에서 DB지수 및 CH지수를 함께 활용하기로 한다. 동일한 군집에 속한 개체 간의 응집도와 타 군집에 속한 개체 간의 분리도의 비율 개념으로써 계산되는 DB지수 및 CH지수는 서로 응집도와 분리도가 개념적으로 역수관계에 있다. 따라서 DB지수는 작을수록, CH지수는 클수록 군집결과가 더 우수하게 평가된다.

앞서 의사결정나무 분지에 따른 우선순위로 포함되는 특성 값에 대한 k-평균 군집결과를 DB지수 및 CH지수의 측정치로 평가하여 최종적으로 k-평균 입력특성 선택여부를 판단할 수 있다. 그리고 최소 입력특성의 수(N)가 정해져 있다면, N개 이상의 입력특성을 군집유효성지수를 참고하여 선택할 수 있을 것이다.

Ⅴ. 사례 데이터를 통한 적용 및 평가

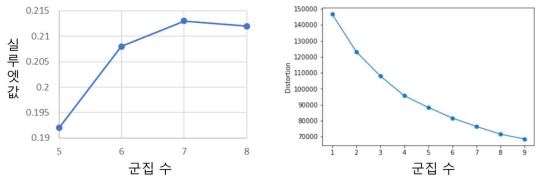
5.1 사례 데이터 및 의사결정나무 구성

신용카드 고객의 세분화 사례연구[24]에서 활용되는 데이터를 본 논문에서도 활용하기로 한다. 이 데이터는 캐글 사이트(www.kaggle.com)를 통해 공개되고 있으며, <표 2>와 같은 특성으로 요약된 6개월간의 신용카드 지불거래 내역을 8,950명의 고객을 대상으로 포함하고 있다. CUST_ID는 본 논문의 분석내용과 관계가 적으므로 제외하고, 결측치가 포함된 경우를 제거한 8,636명의 데이터 값을 표준화 스케일링(scaling)한 후 분석하기로 한다.

실루엣Silhouette) 값 및 엘보우(Elobow) 방식의 도표를 <그림 3>과 같이 확인하여, 군집 수 k=7로 결정할 수 있다. 다음으로 표준화 스케일링된 8,636명의 데이터를 대상으로 k-평균으로 군집화한 후, 7개의 군집을 의미하는 레이블 값 0, 1,…,6을 목표특성으로 하는 의사결정나무를 구성할 수 있다. 이때 군집화를 위한 입력특성의 선택에 가능한 한 모든 특성이 포함될 수 있도록, 의사결정나무의 깊이(depth)에 제한을 두지 않고 <그림 4>와 같이 구성할 수 있다. 의사결정나무의 일부를 확대하여, 각 노드의 분류값(cluster_5, cluster_6 등) 등을 포함한 구성형태를 시각적으로 확인할 수 있다.

<표 2> 사례 데이터에 포함된 특성

특성 이름	내용
CUST_ID	신용카드 보유자 ID
BALANCE	일일 평균 잔고의 월 평균값
PURCHASES	지난 1년간 총 구매금액
ONEOFF_PURCHASES	일시불 구매 총금액
INSTALLMENTS_PURCHASES	할부 구매 총금액
CASH_ADVANCE	현금선불 총 금액
BALANCE_FREQUENCY	잔고가 존재한 월의 비율
PURCHASES_FREQUENCY	구매가 발생한 월의 비율
ONEOFF_PURCHASES_FREQUENCY	일시불 구매가 발생한 월의 비율
PURCHASES_INSTALLMENTS_FREQUENCY	할부 구매가 발생한 월의 비율
CASH_ADVANCE_ FREQUENCY	선불 구매가 발생한 월의 비율
PURCHASES_TRX	평균 구매금액
CASH_ADVANCE_TRX	현금선불 평균 구매금액
CREDIT_LIMIT	신용구매 한도
PAYMENTS	12개월간 청구서에 의한 지불금액
MINIMUM_PAYMENTS	12개월간 최소 청구 금액
PRC_FULL_PAYMENT	청구금액을 완불한 월의 비율
TENURE	카드 보유 개월 수



<그림 3> 실루엣 및 엘보우 처리결과

<표 3> 우선순위별 입력특성

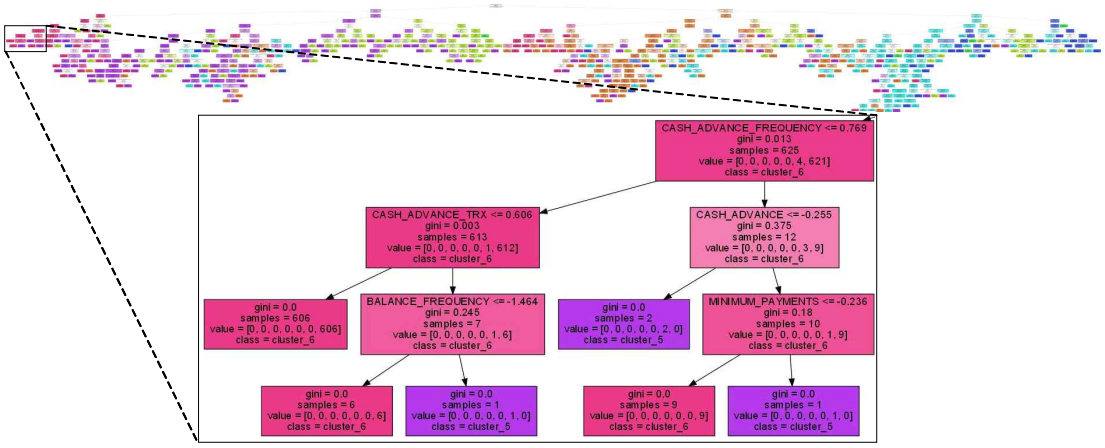
순위	특성 (깊이, 지니지수)
1	PURCHASES_FREQUENCY (1, 0.780)
2	BALANCE_FREQUENCY (2, 0.566),
3	ONEOFF_PURCHASES_FREQUENCY (2, 0.643)
4	BALANCE (3, 0.091)
5	CASH_ADVANCE_TRX(3, 0.319)
6	CASH_ADVANCE(3, 0.386)
7	PURCHASES_TRX(3, 0.500)
8	CASH_ADVANCE_ FREQUENCY(4, 0.224)
9	PURCHASES (4, 0.350)
10	INSTALLMENTS_PURCHASES (5, 0.31)
11	MINIMUM_PAYMENTS (5, 0.500)
12	CREDIT_LIMIT (5, 0.500)
13	PURCHASES_INSTALLMENTS_FREQUENCY (6, 0.278)
14	PAYMENTS (7, 0.133)
15	TENURE (7, 0.32)
16	ONEOFF_PURCHASES (7, 0.523)
17	PRC_FULL_PAYMENT (8, 0.48)

4.2절에서 정한 입력특성 우선순위에 따라, 전체 깊이가 20인 <그림 4>의 CART방식으로 구한 의사결정나무로부터 특성을 <표 3>과 같이 선택할 수 있었다. 깊이 8까지 입력특성 선택을 진행했을 때, 17개 특성이 모두 선택되었음을 알 수 있다. PURCHASES_FREQUENCY는 최상위노드(root node)의 특성이며, PRC_FULL_PAYMENT는 깊이 8의 노드에서 선택된 분지특성이다.

5.2 입력특성의 선택과 군집유효성

군집결과의 군집 레이블 값을 목표특성으로 한 <그림 4>의 의사결정나무로부터 우선순위에 따라 구한 <표 3>의 특성들이 다음 두 경우에 효과를 가지는지 확인하고자 한다. 첫 번째는 <표 3>의 17개 특성 중에서 우선순위에 따라 포함되어야 할 최적의 입력특성을 선택하는 것이다. 두 번째는 우선순위로 정한 특성들이 k-평균 입력특성으로서 가지는 군집유효성의 개선효과 여부이다. 이를 위해 다음과 같이 군집유효성지수를 비교하는 방식을 진행하였다.

(1) <표 3>의 전체 특성으로부터 시작하여 후순위 특성을 1개씩 제외해가며, k-평균 군집화 결과의 군집유효성지수를 계산하여 비교한다. 이를 통해



<그림 4> 군집 레이블값이 목표특성으로 구성된 의사결정나무

최선의 군집유효성지수를 보이는 특성집합을 선택할 수 있다.

(2) 전체 17개 특성에서 시작하여, 랜덤(random)하게 정한 순서로 특성을 1개씩 제외하여 구한 k-평균 군집화의 군집유효성지수를 위 (1)의 경우와 비교한다. 이를 통해 <표 3>의 경우가 다른 입력특성 대안보다 군집유효성이 개선되는지 확인이 가능하다.

위 (1)의 방식으로 전체 17개 특성으로부터 시작하여 1개의 특성(PURCHASES_FREQUENCY)까지 줄어들수록, k-평균 군집결과에 대한 두 가지 군집유효성지수가 모두 개선되어 가는 <그림 5>의 경향을 보인다. 1개 특성에 대한 군집결과의 CH지수는 아주 큰 값(190,300.9)이어서 그림에 표시되지 않았다. DB지수와 CH지수는 각각 작거나 클수록 군집유효성이 우수하게 평가되는데, <그림 5>에서 1개의 특성만으로 군집화할 때 군집유효성이 높게 나타난다. 최소 입력특성의 수 N=8이라면(제외된 특성의 수가 9이하), DB지수로 평가하는 경우 <표 3>의 CREDIT_LIMIT 이하의 후순위 특성은 제외할 때 <그림 5>의 (1)에서 1.136의 DB지수값으로서 군집유효성이 우수하다. 반면, CH지수의 경우에도 DB지수와 같이 특성의 수가 적을수록 군집유효성이 대체로 좋게 나타난다. 미세하지만, 4개 순위까지의 특성이 제외된 경우가 3개 또는 5개 순위까지 제외된 경우보다 우수하게 나타난다.

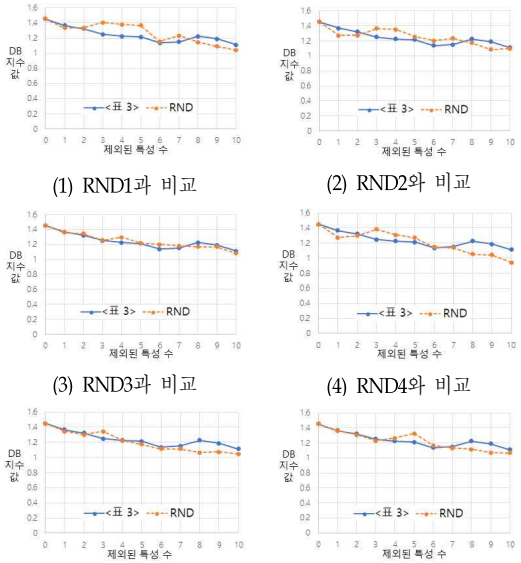
<그림 5>와 같이 군집유효성지수를 위해 데이터의 특성을 과도하게 줄이는 것이 비현실적이므로, 전체 17개의 특성 중에서 제외가능한 특성의 수를 10개까지(최소 입력특성의 수 N=7) 고려하였다. <표 3>의 우선순위로 선택한 특성들로부터 후순위인 17순위의 특성부터 하나씩 제외한 데이터들에 대해 k-평균 군집화를 하였다. 동시에 랜덤하게 선택한 같은 수의 특성(랜덤 특성)으로 구성된 데이터에 대하여, k-평균 군집화를 하고 군집유효성지수를 계산하여 비교하였다. 비교의 객관성을 위해 다수의 랜덤하게 구성한 특성집합을 구성하여 비교하였으며, <표 3>의 특성집합과 랜덤하게 선택한 6개의 특성집합 'RND1' ~ 'RND6'의 경우에 대하여 k-평균 군집결과의 DB지수와 CH지수를 비교한 결과는 각각 <그림 6>과 <그림 7>과 같다.

(1) 특성 제외에 따른 DB지수의 변화

(2)특성 제외에 따른 CH지수의 변화

<그림 5> 특성의 제외에 따른 군집유효성지수

DB지수로 비교한 <그림 6>에서, <표 3>의 특성과 랜덤특성 간에는 군집유효성에 있어서 일관적인 차이가 있다고 보기 어렵다. 그러나 CH지수로 비교한 <그림 7>을 보면, <표 3>의 특성과 랜덤특성 간에는 군집유효성 차이가 뚜렷하며, 모든 랜덤특성에 대해서 <표 3>의 특성이 우수한 군집유효성을 보인다. 이와 같이, 군집유효성지수의 종류에 따라 의사결정나무를 이용한 입력특성 선택방식의 군집유효성 개선효과의 측정에 차이가 나타난다.

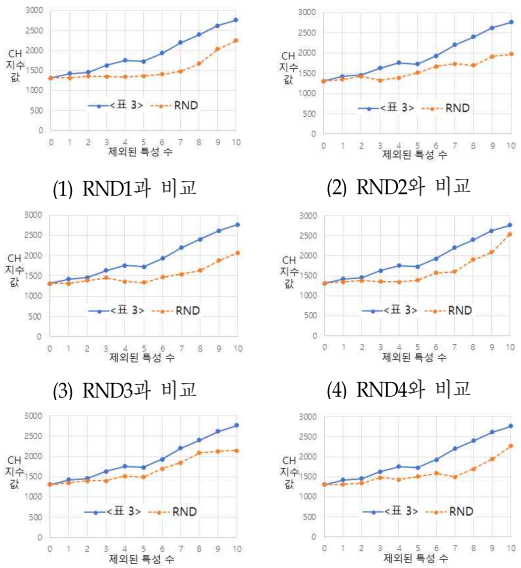


(1) RND1과 비교 (2) RND2와 비교

(3) RND3과 비교 (4) RND4와 비교

(5) RND5와 비교 (6) RND6과 비교

<그림 6> 선택한 입력특성(<표 3>)과 랜덤특성 간의 DB지수 비교



(1) RND1과 비교 (2) RND2와 비교

(3) RND3과 비교 (4) RND4와 비교

(5) RND5와 비교 (6) RND6과 비교

<그림 7> 선택한 입력특성(<표 3>)과 랜덤특성 간의 CH지수 비교

5.3 입력특성 구성방안의 타당성

의사결정나무를 활용한 k-평균 입력특성 방안을 제안하고, 신용카드 고객세분화의 데이터를 통해 적용 및 평가한 결과로써 다음 사항을 확인할 수 있다.

(1) k-평균 군집 레이블 값을 목표변수로 하는 의사결정나무로부터 우선순위로 구한 입력특성에서, 우선순위가 낮은 입력특성부터 순서대로 제외하면 군집유사성지수가 개선되는 결과를 가져온다. 그런데 입력특성의 수가 적어지면 랜덤 특성에서도 DB지수나 CH지수가 개선되는 경향이 있으므로, 의사결정나무를 통해 정해진 입력특성의 우선순위대로 반드시 군집유효성이 개선된다고 확신하기 어렵다.

(2) 의사결정나무의 분지특성 우선순위로 선택한 입력특성과 랜덤하게 선택한 입력특성에 대해 비교하면, CH지수로 평가하는 경우 군집유효성의 분명한 개선이 있는 것으로 확인된다. 그런데 DB지수의 경우 군집유효성의 차이가 분명히 확인되지 않는다. 따라서 군집유효성지수에 따라 개선효과가 다르게 평가될 수 있다.

VI. 결론 및 토의

본 논문에서는 최근 여러 분야[25, 26]에서 활발히 응용되는 의사결정나무 분류를 활용한 k-평균 군집화의 입력특성 선택방안을 정리하고, 신용카드 고객세분화 문제에 적용하여 타당성을 평가하였다. 세부적인 정리내용으로는 의사결정나무와 k-평균이 분석의 방향성을 공유하도록 k-평균의 군집결과 레이블 값을 의사결정나무 분지의 목적특성으로 구성하였으며, 의사결정나무의 정보이득을 고려한 분지특성으로써 k-평균의 입력특성을 선택하는 방안을 정리하였다. 군집유효성지수를 통한 평가에서 본 논문의 입력특성 우선순위가 반드시 군집유효성을 개선하는 것으로 판단하기는 어려우나, CH지수로 평가하는 경우 랜덤한 우선순위로 정한 입력특성보다는 본 논문에서 제시한 방식의 군집유효성 개선효과가 충분히 나타나

는 것으로 판단된다. 그런데, DB지수로는 군집유효성 개선효과를 확인하기가 어려웠다.

본 논문에서 시사하는 바로는, 첫 번째로 k-평균을 비롯한 군집분석에서 입력특성 선택방안의 대안이 될 수 있는 방안을 제시하였다는 점이다. 효과적인 군집분석을 위해서는 군집방식의 선택에 못지않게 입력특성의 선택이 중요하기 때문이다. 두 번째로는 k-평균과 같은 군집분석의 타당성 평가에서 군집유효성지수의 활용에 어려움이 있다는 것이다. 분류분석에 사용되는 정확도나 정밀도 등과 같은 뚜렷한 성과측정치와 달리, 군집분석의 군집유효성지수는 지수별로 응집도와 분리도의 계산식이 상이하여 효과적이고 일관적인 평가가 쉽지 않다는 점이다.

본 논문의 한계점으로 지적될 수 있는 사항은 제시한 방식의 실험적 적용 및 평가가 신용카드 고객세분화의 사례 데이터에 국한되어 이루어졌다는 것이다. 따라서 이론적 또는 경험적 측면에서 일반화하기에는 다소 부족한 면이 있으며, 이에 대한 추가적인 향후 연구가 필요한 것으로 사료된다. 또한 군집유효성 평가에서 군집유효성지수의 값이 보다 일관적일 수 있도록 개선된 측정지수의 연구가 필요한 것으로 보여진다.

참고문헌

[1] 이명식, “신용카드시장에서 지각된 서비스 가치와 혁신성이 고객의 브랜드참여 및 브랜드 로열티에 미치는 영향,” 신용카드리뷰, 제13권, 제2호, 2019, pp.36-60.

[2] Martins, M.C. and Cardoso, M., “Cross-validation of Segments of Credit Card Holders,” Journal of Retailing and Consumer Services, Vol.19, 2012, pp.629–636.

[3] Epstein, M.J., “Managing Customer Profitability,” Journal of Accountancy, Vol.206, No.6, 2008, pp.54–59.

[4] 진서훈 · 안상욱, “신용카드업에서 데이터마이닝의 활용: 고객행동기반의 고객세분화,” 한국통계학회 2004년도 학술발표논문집, 2004, pp.171–174.

[5] 진서훈, “데이터마이닝에 의한 고객세분화 개발,” 응용통계연구, 제18권, 제3호, 2005, pp.555-565.

[6] Maldonado, S., Carrizosa, E. and Weber, R., “Kernel Penalized K-means: A Ffeature Selection Method Based on Kernel K-means,” Information Sciences, Vol.322, 2015, pp.150–160.

[7] 박진수 · 장남식 · 황유섭, “S카드사의 가맹점 분류체계 정비를 통한 고객세분화 전략,” Information Systems Review, 제10권, 제3호, 2008, pp.89-109.

[8] 강진웅 · 금기정 · 손승녀, “의사결정나무와 신경망 모형 결합에 의한 운전자 우회결정요인 분석,” 한국도로학회논문집, 제13권, 제3호, 2011, pp.167-176.

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

[10] Li, Y et al., “Customer Segmentation Using K-Means Clustering And The Adaptive Particle Swarm Optimization Algorithm,” Applied Soft Computing, Vol.113, 2021, 107924.

[11] Calvo-Porral, C. and Levy-Mangin, J.P., “From Foodies to Cherry-Pickers: A Clustered-Based Segmentation of Specialty Food Retail Customers,” Journal of Retail Consumer Services, Vol.43, 2018, pp.278–284.

[12] Das, P., Das, D.K. and Dey, S., “A Modified Bee Colony Optimization And Its Hybridization with k-Means for An Application to Data

Clustering," Applied Soft Computing, Vol.70, 2018, pp.590-603.
[13] Xie, H. et al., "Improving K-Means Clustering with Enhanced Firefly Algorithms," Applied Soft Computing, Vol.84, 2019, 105763.
[14] Rousseeuw, P.J., "Silhouettes: A Graphical Aid to The Interpretation And Validation of Cluster Analysis," Journal of Computer Applied Mathematics, Vol.20, 1987, pp.53-65.
[15] MacQueen, J., "Some Methods for Classification And Analysis of Multivariate Observations," Fifth Berkley Symposium on Mathematical Statistics and Probability, Vol.1. No.1, 1967, pp.281-297.
[16] Cordeiro R. and Makarenkov, V., "On K-Means Iterations And Gaussian Clusters," Neurocomputing, Vol.553, 2023, 126547.
[17] 김민 · 전주혁 · 우경구 · 김명호, "범주형 속성 기반 군집화를 위한 새로운 유사 측도," 정보과학회 논문지, 제37권, 제2호, 2010, pp.71-81.
[18] 이수현 · 정영선 · 김재윤, "경영사례를 이용한 군집화 유효성지수의 성능비교," 한국경영과학회지, 제41권, 제2호, 2016, pp.17-33.
[19] Guyon, I. et al., "Feature Extraction," Foundations and Applications, Springer, 2006.
[20] Maldonado, S., Weber, R. and Basak, J., "Kernel-Penalized SVM for Feature Selection," Information Sciences, Vol.181, 2011, pp.115-128.
[21] Dash, M. et al., "Feature Selection for Clustering - A Filter Solution," IEEE International Conference on Data Mining - Proceedings, 2002, pp.115-122.
[22] 이극노 · 이홍철, "이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구," 한국지능정보시스템학회논문지, 제9권, 제1호, 2003, pp.139-155.
[23] 서광규 · 안범준, "하이브리드 의사결정나무와 인공신경망 모델을 이용한 방문학습지사의 고객세분화," 한국산학기술학회논문지, 제7권, 제3호, 2006, pp.518-523.
[24] Raj, S. et al., "Customer Segmentation Using Credit Card Data Analysis," IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications, 2023, pp.383-388.
[25] 김동형, "이미지 보간을 위한 의사결정나무 분류기법의 적용 및 구현," 디지털산업정보학회 논문지, 제16권, 제1호, 2020, pp.55-65.
[26] 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 : 2023년 11월 14일
수정접수일 : 2023년 11월 23일
게재확정일 : 2023년 12월 01일