The Journal of the Convergence on Culture Technology (JCCT)
Vol. 7, No. 3, pp.481-486, August 31, 2021. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2021.7.3.481
JCCT 2021-8-57

군집분석의 분할 유용도 점수의 영향 분석

Impact Analysis of Partition Utility Score in Cluster Analysis

이계성*

Gye Sung Lee*

요 약 기계학습 알고리즘은 기준 함수를 채택하여 데이터를 처리하고 학습 모델을 유도한다. 군집분석에서 사용하는 기준 함수는 어떤 형태로든지 선호성을 내포하게 되고 이를 통해 유사한 데이터끼리 묶어 준 후 이를 구성하는 변수와 값들을 특정하여 군집을 정의하게 된다. 군집분석에서 사용하는 카테고리 유용도와 분할 유용도 점수가 군집분석 결과물에 어떤 영향을 주는지를 파악하고 이들이 결과에 어떤 편향성으로 이어지는지를 분석한다. 본 연구는 군집분석에 사용되는 기준 함수의 특성에 따라 결과에 미치는 영향을 파악하기 위해 여러 데이터 세트를 이용해 실험하고 결과를 평가한다.

주요어 : 군집분석, 카테고리 유용도, 분할 유용도, 개념형성, 지식발견

Abstract Machine learning algorithms adopt criterion function as a key component to measure the quality of their model derived from data. Cluster analysis also uses this function to rate the clustering result. All the criterion functions have in general certain types of favoritism in producing high quality clusters. These clusters are then described by attributes and their values. Category utility and partition utility play an important role in cluster analysis. These are fully analyzed in this research particularly in terms of how they are related to the favoritism in the final results. In this research, several data sets are selected and analyzed to show how different results are induced from these criterion functions.

Key words : Cluster Analysis, Category Utility, Partition Utility, Concept Formation, Knowledge Discovery

Ⅰ. 서 론

기계학습 알고리즘은 기준 함수(criterion function) 또는 목적함수를 선정하여 학습모델을 유도해 낸다. 이 기준 함수의 선택은 학습 결과 자체의 질을 결정할 뿐만 아니라 그것의 활용 목적이나 방향을 결정하기 때문에 매우 중요한 부분이 아닐 수 없다[1-2]. 감독분류와 달리 군집화라 불리는 무감독분류는 클래스 레이블이 없는 데이터를 대상으로 군집화한다. 군집분석의 산출 결과물은 같은 그룹 내에서는 객체 유사도가 큰 데이터들을 군집시켜 모아야 하고 서로 다른 그룹들 사이에서는 객체 비 유사도 값이 큰 데이터들을 서로 분리해 최적의 구조를 갖게 만든다[3-5]. 기계학습은 크게 두 가지 활용 목적을 통해 구분될 수 있다. 첫째는 기계학습의 결과 모델을 새로운 데이터에 적용하여 예측하는 문제해결 지향적 접근방식이다[6]. 또 다른 접근 방향으로

*정회원, 단국대학교 소프트웨어학과 교수 (제1저자)
접수일: 2021년 6월 25일, 수정완료일: 2021년 7월 21일
게재확정일: 2021년 7월 30일

Received: June 25, 2021 / Revised: July 21, 2021
Accepted: July 30, 2021
*Corresponding Author: gslee@dankook.ac.kr
Dept. of Software, Dankook Univ, Korea

군집화를 통해 새로운 지식이나 패턴을 발견하거나 개념을 정립하는 목적으로 사용하는 접근방법이 있다[5]. 수집된 데이터 또는 관측 자료로부터 유용한 지식이나 개념을 생성하는 것이다. 이 두 가지 방향에서 기계학습은 서로 다른 알고리즘을 채택할 뿐만 아니라 알고리즘의 핵심인 기준 함수를 다르게 설정하게 된다. 학습의 결과물은 의사결정 트리, 네트워크 모델, 규칙, 또는 군집화 등으로 표현된다. 문제해결 지향적인 접근에서는 결과로 생성되는 학습 모델을 활용해 새로운 데이터에 대한 분류나 추론을 시행한다[6]. 이때 다양한 모델이 유도될 수 있는데 가장 높은 예측정확도를 갖는 모델을 선정하게 된다. 지식발견의 경우 자료에 대한 특정 패턴이나 속성값들로 이뤄진 모델을 통해 새로운 사실이나 개념을 발견하게 된다. 이런 접근방식에서는 군집의 동질성, 분리의 명확성과 차별적 표현성의 능력을 중시하게 된다[6-8].

예측정확도 제고를 위한 방법에서는 일반화 과정을 거친다. 이는 과적합을 피하고 모델을 단순화하여 개념을 일반화한다. 군집화 학습에서는 단순화 과정도 중요하나 차별화하는 과정도 중요하다. 본 연구에서는 주어진 데이터 세트를 군집으로 분할하는 군집화 알고리즘을 선정하고 여기에서 사용하는 기준 함수를 통해 어떤 결과를 유도하는지 실험을 통해 확인해 본다. 각 데이터 세트에 대한 군집화 결과를 평가하여 기준 함수의 영향을 분석해 본다.

II. 카테고리 유용도

군집화를 위해 사용하는 대표적인 알고리즘의 하나로 Cobweb 시스템[8]이 있다. 여기에 사용되는 알고리즘은 점진적으로 개념을 학습하는 알고리즘으로 학습 데이터를 입력받아 이를 학습하여 개념 트리를 구성한다. 데이터가 하나씩 입력되면서 개념 트리가 지속해서 변형해 나가는 전형적인 점진적 학습 기법이다. Cobweb 알고리즘은 기준 평가함수로 아래의 카테고리 유용도 함수를 사용하여 학습한다[8-10].

$$CU_k = P(C_k)\sum_i\sum_j (P(A_i = V_{ij}|C_k)^2 - P(A_i = V_{ij})^2) \quad (1)$$

여기서 CU_k는 속성 A_i가 값 V_{ij}를 갖는 데이터에 대해 클러스터 C_k에 속할 확률값의 기댓값을 의미한다. 이는 클러스터 내부의 응집도를 의미하는 클러스터 내 유사도와 클러스터 간 비 유사도를 높여 군집화를 최적화하는 데 사용되는 함수이다. 수식의 맨 앞단에 있는 $P(C_k)$는 클러스터의 크기를 고려한 항으로 작은 규모의 클러스터로 조각나 파편 현상이 일어나지 않게 하여 적정 수의 클러스터와 적정 크기를 갖춘 균형 잡힌 군집화를 유도하는데 쓰인다[11,12]. 이 항이 없다면 단일 데이터로 이뤄진 클러스터들의 군집화가 가능하게 된다. $P(C_k)$를 통해 개념 트리의 복잡도를 현저히 줄여 학습한 개념을 단순화하고 일반화할 수 있다. 그러나 이 항은 클러스터의 크기가 클수록 큰 값을 갖게 되기 때문에 좀 더 큰 규모의 클러스터를 선호하는 편향성도 동시에 갖게 된다. Cobweb 계열의 후속 시스템인 Iterate[9,13]와 Reit[11,12] 군집화 시스템은 입력되는 데이터의 처리 순서에 따라 다양한 개념 트리가 발생하는 문제를 해결하고자 카테고리 일치도 함수를 활용한 재배치 알고리즘과 다음 식으로 표현한 분할 유용도(Partition Utility) 점수인 PU를 사용한다.

$$PU = \frac{1}{K}\sum_{j=1}^{K} CU_j \quad (2)$$

분할 유용도 점수는 K개의 클러스터를 갖는 군집화에서 모든 클러스터의 카테고리 유용도의 평균값을 산출하여 구해진다. 데이터의 재배치 과정을 거쳐 카테고리 유틸리티를 재계산한 후 이들의 평균값을 계산하고 이 값을 개선하면서 새로운 군집화를 구축해 나가는 것이다. 이 값이 클수록 균형 잡힌 군집화가 구성된 것으로 볼 수 있으므로 군집화 결과가 더 안정된 것으로 판단한다. 카테고리 유용도와 마찬가지로 분할 유용도 점수에도 K가 사용되는데 이 또한 적은 수의 클러스터 수를 선호하는 방향으로 군집화가 진행됨을 의미하고 큰 규모의 클러스터를 선호하는 편향성으로 학습 결과가 유도된다. 따라서 각 클러스터의 크기가 더 커지는 방향으로 군집화가 일어나게 된다. 데이터의 입력 순서에 따른 학습 결과 차이를 최소화하기 위해 분류 트리를 통해 얻어지는 일차 분할에 대해 카테고리 일치도 함수를 적용하여 데이터를 재배치한다. 이 과정은 일종의 클러스터 보정 과정으로 볼 수 있다. 데이터 d와 클러스터 k 사이의 일치도를 측정하는 카테고리 일치도는 다음과 같이 정의된다[13].

$$CM_{dk} = P(C_k) \sum_{i,j \in \{A_i\}_d} (P(A_i = V_{ij} | C_k)^2 - P(A_i = V_{ij})^2) \quad (3)$$

이미 배치된 데이터는 재배치 과정을 통해 새로운 위치로 이동할 수 있다. 각 데이터에 대해 더 나은 일치도를 갖는 클러스터가 있다면 기존에 속해 있는 클러스터에서 더 일치하는 클러스터로 이동하는 것이 가능하다. 분할 유용도를 활용한 군집화는 분할 유용도 점수가 증가하는 한 계속 반복될 것이다. 카테고리 유용도(CU), 분할 유용도(PU), 카테고리 일치도(CM)를 이용한 군집화 알고리즘[12]이 아래에 표시되어 있다.

1. PU =0
2. CU로 분류 트리 생성하고 초기 클러스터 생성
3. CM으로 자료의 재배치 실행
4. PU계산, PU 개선이 없으면 종료
5. 중심정렬로 자료 정렬
6. 2번으로 이동

III. 데이터 세트

분할 유용도에 의한 군집화 알고리즘을 적용하여 군집화할 데이터 세트는 UCI 자료 사이트[14]로부터 수집하였고 본 연구에서는 4개의 데이터 세트를 사용한다. 일반적으로 군집화를 위한 데이터 세트는 클래스 레이블이 없는 데이터를 대상으로 한다. 본 연구에서는 군집화의 결과 분석에서 같은 클래스의 데이터 개체 간의 군집 내 응집도를 분석하고 서로 다른 클래스 정보를 갖는 데이터의 혼재도를 분석하기 위해 클래스 레이블을 포함하는 데이터 세트를 선정하였다. 먼저 콩(soybean)의 질병에 관련된 데이터 세트는 총 47개 데이터로 이뤄져 있으며 35개의 속성으로 이뤄져 있다. 35개 속성값에 대해 네 가지($D_1 \sim D_4$) 질병 중 하나를 특정하게 된다. D_1에서 D_3까지는 각 질병에 대해 10개의 데이터를 포함하고 일련번호가 1~10, 11~20, 21~30을 갖는다. 마지막 D_4에는 17개(31~47번)의 데이터가 포함되어 있다.

두 번째 데이터 세트는 붓꽃(iris) 군집화에 대한 데이터로 150개의 자료로 이뤄져 있으며 4개의 속성값을 갖는다. 붓꽃 데이터 세트는 3개 종류의 붓꽃 50개씩으로 이뤄져 있다: setosa, virginica, versicolor. 세 번째 데이터 세트는 풍선에 관한 데이터로 총 6개 속성에 대하여 풍선이 불려 진(inflated, defalated) 상태를 결정하는 12개의 데이터로 이뤄져 있다. 마지막으로 체스 엔드 게임 데이터 세트가 있다. 총 6개의 속성값에 대하여 18개의 클래스 값 중 하나를 갖는다. 클래스 값은 무승부와, 0번에서 16번의 이동을 통해 화이트 플레이어가 승리하는 경우로 나눠진다. 총 28,056개의 데이터를 갖는다.

카테고리 유용도를 사용하여 군집화 과정을 거치는 경우 군집화가 항상 동일한 결과로 이어지지 않는다. 카테고리 유용도를 이용해 분류 트리를 형성하고 카테고리 유용도 값의 피크 값의 위치를 조사하여 초기 분할의 클러스터를 결정한다. 이때 제공되는 데이터 개체는 랜덤하게 선택되어 처리되고 초기 분류 트리가 다양하게 구성되며 이로부터 얻어지는 초기 분할도 서로 다른 결과로 이어진다. 그러나 다양한 분할은 재배치 과정을 거치면서 특정 군집화로 수렴하게 되고 분할 유용도 점수가 크게 변하지 않으면 군집화를 종료하게 되며 최종 결과를 최적의 군집화로 결정한다.

IV. 실험 및 결과

분할 유용도에 의한 군집화 알고리즘을 앞장에서 소개한 4개의 데이터 세트에 적용하고 이를 통해 생산된 군집화 결과를 분석한다. 서로 다른 결과로 이어지는 것은 군집화 과정의 특징 중 하나이다. 초기 시드 선택에 따라 다른 결과로 이어지는 경우가 있고 데이터 처리 순서에도 영향을 받는다. 이들 결과에 영향을 미치는 카테고리 유용도와 분할 유용도 관점에서 각 데이터 세트의 결과물을 분석한다.

1. 콩 질병 군집분석

군집화 알고리즘을 콩 질병 데이터에 적용하였을 때 4개의 질병 그룹으로 군집화되는 것을 쉽게 예상할 수 있다. 실제 4개로 군집화되는 결과도 있지만 3개로 군집화되는 경우도 종종 발생한다. 4개로 군집화될 때는 정확하게 D_1에서 D_4로 군집화됨을 확인할 수 있었는데 이때의 분할 유용도 점수가 1.389로 나왔다. 3개 그룹으로 군집화되는 경우 유사한 속성값을 공유하는 D_3와 D_4가 합쳐서 하나의 클러스터를 형성한다. 이때의 분할 유용도 점수는 4개의 클러스터의 경우보다 큰 1.468이 된다. 군집화 알고리즘은 결과적으로 3 클러스터 군집화를 최종 결과로 선택하게 된다. 4개의 클러스

터로 이뤄진 군집화가 더 직관적으로 이해될 수 있어 이 결과를 수용하는데 큰 이의가 없다고 본다. 분류 유용도 관점에서는 큰 규모의 클러스터를 선호하는 편향성이 적용되었다고 볼 수 있어 이 카테고리 유용도와 분할 유용도의 영향력이 결정적으로 작용하였다고 볼 수 있다. 특히 이 군집화의 목적이 지식 발견이나 개념 형성에 있다면 큰 규모의 클러스터를 선호하는 대신 다수 개의 클러스터를 갖는 군집화가 선호될 수도 있을 것이다. 즉, 4개 클러스터의 장점은 각 질병에 대해 속성값을 파악할 수 있고 각 질병의 특징이나 패턴을 발견할 수 있다. 또한 질병 간 속성 값을 차별화할 수 있게 되므로 지식 발견 관점에서는 좀 더 활용 가치가 높은 군집화라 볼 수 있다.

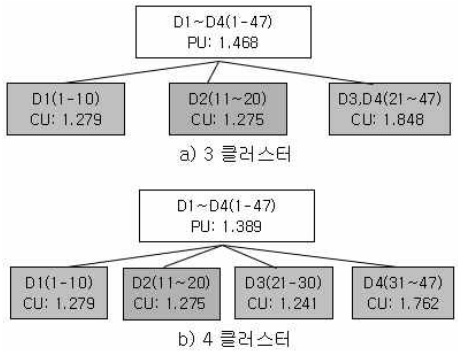

그림 1. 콩 질병 군집분석
Figure 1. Soybean Disease Clustering

3 클러스터 군집화에 대해서는 단순화를 통한 일반화가 더 잘 되었다고 해석할 수 있고, D_3와 D_4가 유사성을 공유하여 통합된 것으로 해석할 수 있다. D_3와 D_4는 좀 더 세심하게 구분할 필요가 있다고 판단되면 이를 다시 군집화하여 나누는 것도 하나의 방법이 될 수 있다.

2. 붓꽃 군집분석

붓꽃 군집화의 경우 3개의 클래스 레이블로 나뉘어 있으므로 3개의 클러스터를 예상해 볼 수 있는데 실제 군집화 결과에서는 3개 클러스터의 군집화가 가장 많이 발생하지만 2개 또는 4개의 클러스터로 군집화되기도 한다. 3개의 클래스 각각 50개의 데이터가 포함되어 있는데 setosa의 경우 독자적인 특성이 뚜렷하여 항상 별도의 클러스터를 형성하나 나머지 2종류의 붓꽃에서는 공유되는 속성값들이 혼재되어 있어 군집화 결과에서도 혼재되는 경우가 자주 발생한다. 혼재되는 정도를 나타내는 혼재도를 다음과 같이 정의한다. 클러스터를 대표하는 클래스 레이블을 클러스터의 대표클래스 집단으로 설정하고 이에 일치하지 않은 데이터를 이상치(O_i)로 분리하여 이들의 합을 전체 데이터 크기(T)로 나눠 군집분석 결과의 혼재도(H)를 다음과 같이 정의한다.

$$H = \sum O_i / T \tag{4}$$

혼재도 H가 크면 서로 다른 클래스 레이블이 클러스터에 혼재되어 있어 클러스터의 대표성을 명확하게 표현하기 힘들어 클래스를 특정하기 어렵게 될 것이다.

표 1. 붓꽃 군집화
Table 1. Irish Clustering

구분	클러스터수	PU	H(%)
군집화1	3	0.462	11.3
군집화2	3	0.453	14.7
군집화3	3	0.457	5.3
군집화4	4	0.383	4.0

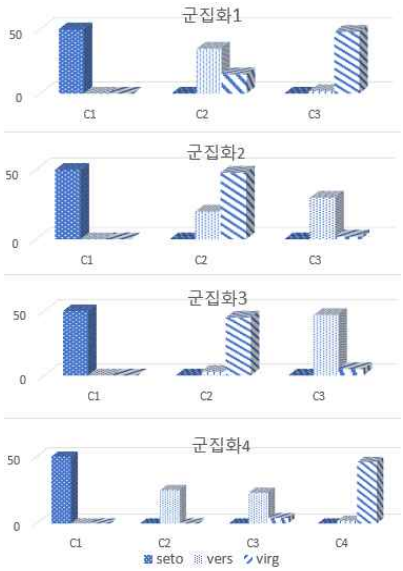

그림 2. 붓꽃 군집화 분포
Figure 2. Iris Clustering Distribution

표 1은 붓꽃 군집화 결과 중 대표적인 4개 군집화 결과를 보여준다. 첫 3개의 군집화는 3개의 클러스터로 군집화하는 결과를 보여주고 있으며 4개의 클러스터로 군집화하는 결과는 군집화 4에 나타나 있다. 첫 3개의 군집화는 상대적으로 큰 분할 유용도 점수를 가지고 있어 본 군집화 알고리즘이 선호하는 결과이다. 반면 4개의 클러스터를 갖고 분할하는 군집화 4에서는 분할 유

용도 값이 상대적으로 낮은 것을 볼 수 있다. 3개 클러스터로 분할하는 군집화의 클래스 분포를 그림 2가 잘 보여주고 있다. 군집화 1, 2, 3의 클러스터 C_2에는 versicolor와 virginica가 섞여 있는 것을 볼 수 있고 이는 테이블 2의 혼재도를 통해 확인할 수 있다. 낮은 분할 유용도를 가짐에도 혼재도가 가장 낮은 군집화 4에서는 비록 클러스터의 수가 많기는 하지만 클래스들이 상대적으로 잘 분할되었다고 판단할 수 있다. 콩 질병 군집화의 경우와 마찬가지로 클러스터의 크기가 큰 것을 선호하는 기준 함수의 선호성으로 인해 발생하는 결과로 판단된다. 콩 질병의 경우와 다른 점은 동일 클래스로 이뤄진 집단이 2개의 집단으로 나뉘는 점이다. 동종의 붓꽃이지만 특징상 분리될 수 있어 이를 통해 차별화되는 속성값들을 특정할 수 있다.

3. 풍선 군집화

풍선 군집화의 결과는 3개의 클러스터로 군집화되는 경우와 5개로 군집화되는 경우로 나눠진다[12]. 3개 클러스터의 경우 분할 유용도 점수가 0.445이며 5개로 군집화되는 경우 0.418 보다 크다. 앞선 2개의 데이터 세트의 결과와 달리 작은 수의 클러스터를 갖는 군집화의 분할 유용도 점수가 더 높았다. 이 경우 카테고리 유용도와 분할 유용도가 적절히 작용하는 경우로 볼 수 있다.

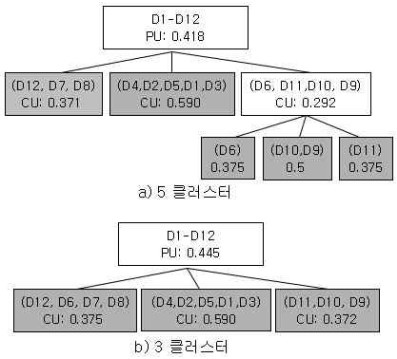

그림 3. 풍선 자료 군집화
Figure 3. Balloon Data Clustering

데이터 세트의 크기가 다른 2개의 데이터 세트와 비교하여 상대적으로 작아 초기 클러스터의 크기가 작게 된다. 클러스터의 크기가 작았을 때 파편화 현상을 피하기 어려운 점이 있다. 카테고리 유용도가 적용될 때 응집도를 수치화하는 조건부 확률값이 상대적으로 큰 값을 산출하게 되어 클러스터 크기를 반영하는 $P(C_k)$의 영향력이 상대적으로 제한될 수 있어 발생하는 현상이라고 볼 수 있기 때문이다. 본 데이터 세트에 대해서는 카테고리 유용도가 이러한 파편화 현상을 막는 목적으로 잘 동작하고 있음을 확인할 수 있다. 그림 3 a)의 5 클러스터 군집화에서 가장 우측 하단의 3 클러스터는 동일 클래스 레이블 deflated를 갖는 군집을 세부 분할한 것으로 과도한 분할이 발생한다고 볼 수 있다. 따라서 이를 제거하는 방향으로 군집화가 이뤄지는 것이 바람직하고 실제 군집화 알고리즘에서는 3개의 클러스터를 갖는 군집화를 선택하게 된다.

4. 체스 자료 군집화

체스 자료 군집화의 결과는 3개의 클러스터로 군집화되는 경우와 4개로 군집화되는 경우로 나눠진다. 우선 대용량의 자료를 통해 초기 분류 트리를 완성하는 것은 매우 불합리한 방식이기 때문에 적정 규모의 깊이를 갖는 초기 분류 트리를 구성하도록 알고리즘을 수정한다. 카테고리 유용도 점수가 개선되지 않을 때는 즉시 분류 트리의 확장을 멈춘다. 분류트리 확장도 너비 우선 방식으로 확장해 나가 트리의 깊이 제한에 이르면 확장을 멈춘다. 모든 자료를 사용하지 않는 방식을 택해 트리의 규모가 과도하게 확장되는 것을 막아 관리할 수 있는 트리를 유지하는 것이 중요하다. 체스 데이터의 군집화 결과는 표 2와 표 3에 표시되어 있다. 앞에 설명한 데이터 세트들과 유사한 형태를 보인다. 3개의 클러스터 군집화의 경우 분할 유용도 점수가 0.212로 4개 클러스터 군집화의 0.188보다 큰 값을 갖게 되고 3개

표 2. 체스 군집화 1 (단위 %)
Table 2. Chess Clustering 1

class	C1	C2	C3	C4
Draw	19.8	7.2	31.0	42.0
0	0	0	81.5	18.5
1	0	0	84.6	15.4
2	0	0	82.5	17.5
3	0	0	39.5	60.5
4	0	0	0.5	99.5
5	0	0	8.3	91.7
6	3.7	0	9.1	87.2
나머지	20.1	7.12	31.6	41.2

표 3. 체스 군집화 2 (단위 %)
Table 3. Chess Clustering 2

class	C1	C2	C3
나머지	39.5	34.0	26.4
15	92.1	0	7.9
16	100.0	0	0

클러스터를 선호하게 된다.

체스 데이터 세트에 대한 실험 결과는 앞선 데이터 세트의 결과와 다른 특징을 포함하고 있다. 표 2와 표 3의 수치는 해당 클러스터에 포함된 클래스 분포를 보여준다. 표의 각 값은 해당 이동 수의 분포를 비율로 표시한 것이다. 표 2의 결과는 6회 이하 소수의 이동으로 화이트 플레이어가 이기는 클래스를 잘 구분하는 반면에 두 번째 테이블인 3 클러스터 군집화는 많은 수의 이동을 통해 승리하는 경우의 클래스를 명확하게 구분하는 결과를 갖게 되었다. 나머지로 표현된 클래스에 대해서는 표에 분포 비율이 의미하는 것처럼 여러 클래스 데이터가 혼재되어 있어 대표 클래스 레이블을 특정할 수 없는 경우이다. 이들을 모아 분포 비율을 합산한 결과다. 앞선 3개의 데이터 세트의 결과와 달리 클러스터의 수가 군집화의 분리 특성을 다르게 보여주는 결과이다.

V. 결 론

기계학습의 목적이 문제해결 능력의 증진을 구하는 것이라고 한다면 학습 방법은 과적합을 피하면서 단순화하는 것이 일반적인 접근방법이다. 군집분석에서는 지식이나 패턴을 발견하거나 새로운 개념을 생성하는 것에 초점을 둔다. 군집분석에 초점을 둔 본 연구에서는 단순화를 선호하는 편향성이 다소 과도한 단순화로 이어지고 이를 통해 서로 다른 클래스의 데이터가 혼재될 수 있다는 점을 여러 데이터 세트에 대한 실험을 통해 밝혔다. 기준 함수인 카테고리 유용도와 분할 유용도 점수의 활용이 군집화 결과에 다양한 영향을 미친다는 것을 확인하였다. 단순화된 학습 모델과 좀 더 세분된 학습 모델 간 선택의 문제가 대두될 때 어떤 학습 모델을 선택해야 하는지는 모델의 활용 목적에 견줘 결정해야 할 필요가 있다. 또 학습 모델 결정에 있어 해당 분야의 전문가 참여와 분석을 통해 적당한 모델의 선택이 필요할 수도 있다. 혼재도가 높은 클러스터의 경우 이를 좀 더 세분하여 군집화 트리로 확장하는 것도 대안으로 제시될 수 있다.

References

[1] I.H. Witten, E. Frank, M Hall, and C. Palestro, “Data Mining: Practical Machine Learning Tools and Techniques,” Elsevier Science & Technology, pp. 9-33, 2017.

[2] T. Mitchell, Machine Learning, McGraw-Hill Education, 1997.

[3] P. Berkhin, “A Survey of Clustering Data Mining Techniques,” Grouping Multidimensional Data, Springer, Berlin, Heidelberg, pp. 25-71, 2006. DOI: https://doi.org/10.1007/3-540-28349-8_2

[4] V. Kumar, N. Rathee, “Knowledge Discovery from Database using an Integration of Clustering and Classification,” International Journal of Advanced Computer Science and Application, Vol. 2. No.3., pp. 29-33, March 2011.

[5] Wikipedia, “Conceptual Clustering,” https://en.wikipedia.org/wiki/Conceptual_clustering

[6] P. Domingos, “The Role of Occam’s Razor in Knowledge Discovery,” Data Mining and Knowledge Discovery, pp. 409-425, Kluwer Academic Publishers, 1999.

[7] U. Luxburg, “Clustering Stability: An Overview,” Foundations and Trends in Machine Learning,” Vol. 2, No. 3, pp. 235-274, 2010, DOI: 10.1561/2200000008

[8] D. Fisher, “Knowledge acquisition via incremental conceptual clustering,” Machine Learning Vol.2, pp. 139-172, 1987. DOI: https://doi.org/10.1007/BF00114265

[9] D. Fisher, “Iterative Optimization and Simplification of Hierarchical Clustering,” Journal of AI Research, Vol. 4, pp. 147-179, 1996. DOI: https://doi.org/10.1613/jair.276

[10]V. Kanageswari and A.Pethalakshmi, "A Novel Approach of Clustering Using COBWEB". International Journal of Information Technology (IJIT), Vol. 3 No. 3, pp 37-42, Jun 2017. DOI: https://doi.org/10.33144/24545414

[11]G.S. Lee, “A Study on Simplification of Machine Learning Model,” The Journal of IIBC, Vol. 16., No. 4., pp. 147-152, Aug 2016. DOI: https://dx.doi.org/10.7236/IIIBC.2015.15.5

[12]G.S. Lee, “The effect of Bias in Data Set for Conceptual Clustering Algorithms,” International Journal of Advanced Smart Convergernce, Vol. 8 No.3, pp. 46-52, 2019. DOI: https://dx.doi.org/10.7236/IJASC.2019.8.3.46

[13]G. Biswas, J.B. Weinberg, and D. Fisher, “ITERATE: A Conceptual Clustering Algorithm for Data Mining,” IEEE Tr. on Systems, Man and Cybernetics, Vol. 28, Part C No. 2. 1998. DOI:https://doi.org/10.1109/5326.669556

[14]UCI repository, https://archive.ics.uci.edu