*The Korean Journal of Applied Statistics* (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1147
**28**(6), 1147–1161

# A Divisive Clustering for Mixed Feature-Type Symbolic Data

Jaejik Kim[a,1]

[a]Department of Statistics, Sungkyunkwan University

(Received September 14, 2015; Revised November 2, 2015; Accepted November 3, 2015)

**Abstract**

Nowadays we are considering and analyzing not only classical data expressed by points in the $p$-dimensional Euclidean space but also new types of data such as signals, functions, images, and shapes, etc. Symbolic data also can be considered as one of those new types of data. Symbolic data can have various formats such as intervals, histograms, lists, tables, distributions, models, and the like. Up to date, symbolic data studies have mainly focused on individual formats of symbolic data. In this study, it is extended into datasets with both histogram and multimodal-valued data and a divisive clustering method for the mixed feature-type symbolic data is introduced and it is applied to the analysis of industrial accident data.

Keywords: mixed feature-type symbolic data, cluster analysis, industrial accident

## 1. 서론

오늘날 컴퓨터의 처리속도와 저장용량이 빠르게 늘어남에 따라 자료의 정기적인 수집 및 저장이 용이하게 되었고, 이로 인한 대용량 데이터셋(dataset)들의 출현은 일상적인 일이 되었다. 이렇게 대용량 데이터셋들은 주위에서 흔하게 찾아볼 수 있게 된 반면, 이를 분석하는 통계 기술의 발전은 데이터셋 크기가 증가하는 속도를 못 쫓아가고 있다. 이러한 대용량 데이터셋을 분석하기 위한 하나의 대안으로 심볼릭 데이터(symbolic data)가 Diday (1987)에 의해 제안되었다. 만일 우리가 개개의 관찰값에 관심이 있기 보다는 어떠한 그룹이나 범주에 관심이 있다고 가정해 보자. 예를 들어, 우리나라 산재보험의 보험요율에 대한 분석을 한다고 가정해보자. 우리나라 산재보험의 요율은 개개의 사업장별로 다르게 정해지는 것이 아니라 각 업종별로 정해진다. 즉, 같은 업종 안에 속한 모든 사업장들은 똑같은 요율을 적용받는다. 따라서, 요율에 대한 분석에 있어서 개개의 사업장은 우리의 관심이 아니고, 분석의 주된 관심은 요율이 정해지는 각 업종들이 된다. 이때 일반적인 분석은 각 업종의 관심이 되는 변수들에 대한 평균값이나 총합을 이용하는 경우가 많다. 하지만, 우리가 평균이나 총합과 같이 요약된 값을 이용하여 분석하게 되면, 각 업종 안에 속한 사업장들의 분포에 대한 정보를 잃게된다. 예를 들어, 어떤 두 업종의 평균 재해율이 똑같다고 가정해보자. 그러나, 한 업종은 대부분의 사업장이 비슷한 재해율을 보이는 반면 또 다른 업종은 사업장들의 재해율에 대한 분산이 크다고 한다면, 요율 산정에 있어서 두 업종은 다르게 취급되고 구분되어야만 한다. 만일 두 업종에 같은 요율을 부과하게 된다면 후자의 업종 내의 재

[1]Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-Gu, Seoul 03063, Korea. E-mail: jaejik@skku.edu

해율이 낮은 사업장들은 불만을 나타낼 가능성이 크다. 이 경우, 각 업종에 대해 요약된 값을 사용하는 대신 구간(interval)이나 히스토그램(histogram) 등을 사용하게 된다면 업종 내부의 변동에 대한 정보의 손실을 줄일 수 있을 것이다. 이러한 동기에서 출발한 것이 심볼릭 데이터이고, 구간, 히스토그램, 목록(list), 표(table), 분포, 모형 등의 형태를 갖는 자료들이 심볼릭 데이터의 범주 안에 속한다 (자세한 내용은 Billard와 Diday (2006), Bock과 Diday (2000) 참조).

대용량 데이터셋의 분석에 있어서 기본적인 분석 중 하나는 데이터셋 안의 복잡한 구조를 이해하고 정보를 추출하기 위해 사용되는 군집분석이다. 지금까지 심볼릭 데이터에 대한 군집분석 연구는 주로 구간과 목록의 형태를 갖는 자료들에 집중되어왔다. Gowda와 Diday (1991)와 Gowda와 Ravi (1995a)는 심볼릭 구간 데이터에 대한 병합적 군집분석방법(agglomerative clustering method)을 소개하였고, Gowda와 Ravi (1995b)와 Chavent (2000)는 구간과 히스토그램 데이터에 각각에 대한 분할적 군집분석방법(divisive clustering method)을 제안하였다. De Carvalho 등 (2006)과 De Carvalho와 Lechevallier (2009)는 구간 데이터에 대한 $K$-평균 군집분석 방법들을 개발하였다. Kim과 Billard (2011)는 히스토그램 데이터의 모든 변수들을 동시에 이용하여 최적의 분할을 찾는 군집분석 방법을 개발하였으며, Kim과 Billard (2012)는 멀티모달(multimodal) 데이터에 대해 한번에 하나의 변수만을 사용하여 최적의 분할을 찾는 방법을 제안하였다. 이렇게 지금까지 많은 연구들은 심볼릭 데이터의 개개의 형태별로 군집분석방법들을 개발하는데 주로 집중되어왔고, 심볼릭 데이터의 여러가지 형태들이 혼합된 데이터에 대한 연구는 De Souza와 De Carvalho (2007)와 De Carvalho와 De Souza (2010)가 $K$-평균 군집분석방법을 다루는데 그쳤다.

본 연구에서는 히스토그램과 멀티모달 변수를 동시에 갖는 관찰값들로 구성된 데이터셋에 대한 계층 분할적 군집분석방법(hierarchical divisive clustering method)을 소개하고 이를 산업재해자료의 분석에 적용한다. 계층 분할적 군집분석방법은 주어진 자료의 분할을 구한다는 측면에서 $K$-평균 군집분석방법이나 PAM(Partitioning Around Medoids)과 비슷하다고 할 수 있으나, $K$-평균 군집분석방법과 PAM은 미리 정해진 그룹의 개수만큼 관찰값들을 나눔으로써 분할이 이루어지는 반면에 계층 분할적 군집분석은 전체 관찰값을 가지고 위에서부터 아래의 방향으로 진행되며, 관찰값들에 대한 두 개의 그룹으로의 연속적인 분할이 계층적으로 이루어지는 군집분석방법이다. 따라서, 계산의 속도에 있어서는 $K$-평균 군집분석과 PAM이 빠르나 계층 분할적 군집분석은 자료의 계층적 구조를 파악하는데 유리하다. 또한, 계층 분할적 군집분석방법은 계층 병합적 군집분석방법에 비해 분석의 초기단계에서 대상들이 잘못 분류될 가능성이 낮다는 장점이 있는 반면 속도가 느리다는 단점이 존재한다.

본 논문에서의 계층 분할적 군집분석방법은 각 분할이 이루어지는 기준이 될 수 있는 이진 질문(binary question)을 제공함으로써 자료 계층적 구조에 대한 이해와 해석을 돕는다. 또한, $n$개의 대상에 대한 모든 가능한 분할(partition)은 $2^{(n-1)}-1$개가 존재하는데 비해 본 연구에서는 히스토그램과 멀티모달 변수를 모두 포함하는 관찰값에 대해 단지 $p(n-1)$개의 분할만을 이용하여 최적의 분할을 찾는 방법을 제안한다. 2절에서는 히스토그램과 멀티모달 데이터를 정의하고 변환하는 방법을 소개한다. 3절에서는 히스토그램과 멀티모달 변수가 혼합된 데이터에 대한 분할적 군집분석방법을 제안하고, 4절에서는 그 군집분석방법을 적용하여 산업재해자료를 분석한다.

## 2. 심볼릭 변수와 그 관찰값들의 변환

본 연구에서 혼합형 심볼릭 데이터는 심볼릭 관찰값 $\mathbf{y}_i$의 $p$개의 변수 중 $q$개의 변수가 히스토그램 값을 갖는 변수이고 나머지 $p-q$개의 변수가 멀티모달 값을 갖는 변수인 자료를 의미한다. 일반적으로 심볼릭 변수 $Y_j,\ j=1,\ldots,p$에 대한 관찰값은 $\mathbf{y}_i=(y_{i1},\ldots,y_{ip})$이다. 여기서 개개의 심볼릭 값 $y_{ij}$는 $Y_j$가 히스토그램 변수라면 히스토그램을 값으로 갖고, $Y_j$가 멀티모달 변수라면 멀티모달인 형태의 값을

**Table 2.1.** An example of the transformation for histogram observations

| $Y_1$ | 원 히스토그램 관찰값 | 변환된 히스토그램 관찰값 |
|---|---|---|
| $y_{11}$ | {[0, 2), 0.2; [2, 4), 0.4; [4,6), 0.4} | {[0, 2), 0.2; [2, 4), 0.4; [4, 6), 0.4; [6, 8), 0; [8, 10), 0} |
| $y_{21}$ | {[3, 7), 0.8; [7, 10), 0.2} | {[0, 2), 0; [2, 4), 0.2; [4, 6), 0.4; [6, 8), 0.267; [8, 10), 0.133} |

갖는다.

히스토그램 변수에 대한 값은 양적인 구간들과 각 구간에 대응하는 가중값(또는 상대도수)으로 구성된다. 멀티모달 변수에 대한 값은 히스토그램과 유사하게 질적인 항목들과 각 항목(item)에 대응하는 가중값으로 이루어진다. 먼저, 히스토그램 변수 $Y_j,\ j = 1, \ldots, p$에 대한 값 $y_{ij}$는 다음과 같이 정의된다:

$$y_{ij} = \{[a_{jk}, a_{j,k+1}),\ \pi_{ijk};\ k = 1, \ldots, t_j\}, \quad i = 1, \ldots, n,\ j = 1, \ldots, p, \tag{2.1}$$

여기서 히스토그램의 구간 $[a_{jk}, a_{j,k+1})$은 실선 $\mathcal{R}$의 부분집합이고 $\pi_{ijk}$는 $i$번째 관찰값, $j$번째 변수, $k$번째 구간에 대응하는 가중값이고 $\sum_{k=1}^{t_j} \pi_{ijk} = 1$이다. 위의 정의에서 주의할 점은 모든 관찰값들이 각 변수에 대해 똑같은 히스토그램 구간들을 갖는다는 것이다. 일반적으로 히스토그램 데이터는 여러가지 다른 조사나 자료로 부터 수집될 수 있기 때문에 각 히스토그램 관찰값들은 서로 다른 구간들을 갖는 것으로 가정된다. 하지만, 일반적으로 히스토그램 데이터는 각 구간들에 대해 자료가 균일하게 분포되어 있다고 가정 하기 때문에 히스토그램들은 모두 똑같은 구간을 갖도록 변환될 수 있다. 우선, 히스토그램 변수 $Y_j$에 대해 서로 다른 구간들을 갖는 $n$개의 히스토그램들 $\{y_{1j}, \ldots, y_{nj}\}$을 가정해보자. 이 $n$개의 히스토그램들의 전체 구간들 중에서 최소값과 최대값을 구한 후 그 최소값과 최대값을 적당한 길이의 구간들로 분할한다. 이 새로운 구간과 각 히스토그램들이 갖고 있는 기존의 구간들이 겹쳐지는 길이의 비율에 따라 각 구간의 가중값들을 분배하고 겹치지 않는 구간에 대해서는 0을 할당함으로써 식 (2.1)에서 정의된 것과 같이 각 변수에 대한 모든 히스토그램들이 똑같은 구간들을 갖도록 만들 수 있다. Table 2.1에 있는 두 개의 원 히스토그램 관찰값을 예로 들어보자. 히스토그램 변수 $Y_1$에 대한 두 개의 관찰값 $y_{11}$과 $y_{21}$는 히스토그램의 구간도 다르고 그 개수도 다르다. 이들이 같은 구간을 갖도록 만들기 위해 우선 구간의 최소값과 최대값을 구하면 각각 0과 10이고, 이것을 길이가 2인 구간들로 나누면 우리는 $\{[0,2), [2,4), [4,6), [6,8), [8,10)\}$과 같은 변환된 구간을 가질 수 있다. 관찰값 $y_{11}$은 원자료에 $[6,8)$과 $[8,10)$의 구간이 없으므로 그 구간에 대한 가중값은 0이 된다. 관찰값 $y_{21}$의 구간 $[3,7)$은 변환된 구간 $[2,4)$, $[4,6)$, $[6,8)$과 겹친다. 그 겹치는 구간의 길이에 비례하여 구간 $[3,7)$의 가중값 0.8을 그 세 개의 구간에 할당하면 각각 0.2, 0.4, 0.2가 되며, 변환된 구간 $[6,8)$은 다시 원구간 $[7,10)$과 겹치므로 같은 방식으로 그 가중값 0.2를 나누면 0.067이 변환된 구간 $[6,8)$에 할당된다. 따라서, 최종적인 변환된 구간 $[6,8)$의 가중값은 0.267 $(= 0.2 + 0.067)$이 된다. 이러한 과정을 거쳐서 Table 2.1에 보이는 변환된 히스토그램 관찰값을 얻을 수 있다 (더 자세한 과정은 Kim과 Billard (2013), De Carvalho와 De Souza (2010) 참조).

이러한 변환은 이론상으로는 큰 이점이 없지만 계산상으로는 상당한 편리함을 제공한다. 심볼릭 구간 데이터(symbolic interval-valued data)는 구간의 상한과 하한만으로 이루어지고, 이는 히스토그램 데이터에서 구간을 하나만 갖고 그 구간의 가중값이 1인 특별한 경우로 고려될 수 있다. 따라서, 구간 관찰값은 위에 소개된 변환을 통해 쉽게 히스토그램 관찰값으로 변환될 수 있다.

히스토그램 변수의 값이 양적인 구간을 갖는다면 멀티모달 값은 질적인 항목을 갖는다. 일반적으로 멀티모달 값들은 정부기관이나 통계기관에서 발간하는 통계표(statistical table)에서 쉽게 찾아볼 수 있다. 물론, 전통적인 데이터셋에서도 멀티모달 자료들을 수집할 수 있다. 예를 들어 신용카드 회사는 고객의 모든 신용카드 결제들을 데이터베이스(database)에 기록하는데, 이 기록들을 각 고객에 대해 신용

**Table 2.2.** An example of the transformation for multimodal observations

| $Y_1$ | 원 멀티모달 관찰값 | 변환된 멀티모달 관찰값 |
|---|---|---|
| $y_{11}$ | {음식, 0.2; 교육, 0.4; 교통, 0.4} | {음식, 0.2; 교육, 0.4; 교통, 0.4; 의류, 0} |
| $y_{12}$ | {교통, 0.8; 의류, 0.2} | {음식, 0; 교육, 0; 교통, 0.8; 의류, 0.2} |

카드를 결제한 분야별로 분류하고 이를 전체 사용금액에 대한 비율로써 나타낸다고 가정해보자. 구체적인 예로 어떤 고객이 {음식, 0.3; 교육, 0.4; 교통 0.1; 의류, 0.2}와 같이 신용카드 사용처에 대한 비율을 가지고 있다면, 이는 하나의 멀티모달 값이 될 것이다. 여기서 멀티모달 변수 $Y$는 '신용카드 사용처'가 될 것이고, 그 고객은 심볼릭 데이터의 대상(object)이 되며, 멀티모달 변수 $Y$가 갖는 항목들은 {음식, 교육, 교통, 의류}가 된다. 따라서, 멀티모달 변수 $Y_j$에 대한 값 $y_{ij}$는 다음과 같이 정의할 수 있다:

$$y_{ij} = \{b_{jk},\ \pi_{ijk};\ k = 1, \ldots, t_j\}, \quad i = 1, \ldots, n,\ j = 1, \ldots, p, \tag{2.2}$$

여기서 $b_{jk}$는 항목 또는 범주를 나타내고, $\pi_{ijk}$는 $i$번째 관찰값, $j$번째 변수, $k$번째 항목에 대응하는 가중값이고 $\sum_{k=1}^{t_j} \pi_{ijk} = 1$이다. 히스토그램 변수와 마찬가지로 멀티모달 변수 역시 각 멀티모달 값들이 서로 다른 항목들의 집합을 갖는 것으로 가정하는 것이 일반적이다. 그러나, 이 또한 식 (2.2)에서 정의된 바와 같이 모든 멀티모달 값들이 같은 항목의 집합을 갖도록 쉽게 변환할 수 있다. 변수 $Y_j$가 멀티모달 변수이고 $n$개의 멀티모달 관찰값 $\{y_{1j}, \ldots, y_{nj}\}$이 있다고 가정해보자. 이 경우 $n$개의 멀티모달 관찰값들이 갖는 모든 항목들의 전체집합을 $\{b_{jk},\ k = 1, \ldots, t_j\}$라고 놓고 각 멀티모달 값에서 없는 항목에는 0의 가중값을 할당하면 모든 관찰값들이 똑같은 항목들을 갖도록 변환될 수 있다. 예를 들어 Table 2.2에 나타난 두 개의 원 멀티모달 관찰값을 가정해보자. 먼저 두 관찰값에 나타난 항목들의 전체집합을 구하면 {음식, 교육, 교통, 의류}이며, 원 관찰값 $y_{11}$은 의류 항목이 없기 때문에 변환된 관찰값에서 의류 항목에 대한 가중값이 0이 된다. 마찬가지로 관찰값 $y_{21}$는 전체집합에 있는 음식과 교육 항목이 없으므로 그 항목들에 대한 가중값으로 0이 할당된다.

심볼릭 멀티(목록) 데이터(symbolic multi-valued data 또는 list-valued data)는 가중값 없이 항목들만의 집합으로 이루어진 자료이고, 이 또한 멀티모달 데이터의 특수한 형태이다. 예를 들어 새의 색깔을 심볼릭 멀티 데이터로 표현할 때 까마귀는 검정색 하나의 항목만을 갖고 까치는 검정과 흰색 두 개의 항목을 갖고 있다. 이를 멀티모달 데이터로 변환하면 까마귀는 {검정색, 1; 흰색, 0}이 되고 까치는 {검정색, 0.5; 흰색 0.5}가 된다. 즉, 각 관찰값의 항목의 개수만큼 균등하게 가중값을 배분함으로써 멀티 변수의 값은 멀티모달 변수의 값으로 변환될 수 있다.

구간 데이터는 히스토그램의 특수한 형태이고 멀티 데이터는 멀티모달 데이터의 특수한 형태이기 때문에 결과적으로 3절에서 제안되는 방법은 구간, 목록, 히스토그램, 멀티모달 변수들의 어떠한 조합을 갖는 데이터셋이라도 적용가능하다.

## 3. 계층 분할적 군집분석

병합적 군집분석방법은 상대적으로 분할적 군집분석방법에 비해 속도가 빠르다는 장점을 가지지만 초기 단계에서 대상을 잘못 분류를 했을 경우 끝까지 그것을 수정할 수 없다는 단점을 가진다. 이에 반해 분할적 군집분석은 병합적 군집분석에 비해 대상을 잘못 분류할 가능성은 낮지만, 분류의 속도가 느리다는 단점이 있다. 따라서, 분할적 군집분석방법의 핵심은 얼마나 빨리 최적의 분할(partition)을 찾을 수 있는가에 달려있다. 즉, 만일 우리가 $n$개의 관찰값들을 가지고 있다면 이 관찰값들에 대한 모든 가능한 조합의 분할의 개수는 $2^{n-1} - 1$로 $n$이 증가하에 따라 모든 가능한 분할의 개수는 기하급수적으로 증가한다. Kim과 Billard (2012)는 $p$개의 멀티모달 변수에 대해 $p(n-1)$개의 분할만을 고려하는 방법을 제

안했는데, 이를 이용하여 본 연구에서는 총 $p$개의 혼합형 심볼릭 변수들을 갖는 대상들의 최적의 분할을 찾기 위해 변수의 형태에 상관없이 똑같이 $p(n-1)$개의 분할만을 검색하는 방법을 소개한다.

먼저 군집분석을 하기 위해서는 일반적으로 관찰값들간의 거리 또는 비유사성에 대한 척도가 필요하다. 심볼릭 데이터는 전통적인 데이터와는 달리 각 변수에 대한 개개의 값이 히스토그램이나 멀티모달 값으로써 이들은 내부적으로 변동을 갖는다. 따라서, 이러한 점을 고려한 다양한 비유사성 척도들이 존재한다. 히스토그램 데이터에 대해 Cha와 Srihari (2002)는 히스토그램 사이의 상관관계를 고려한 척도를 개발했으며, Irpino와 Verde (2006)는 Wasserstein 거리에 근거한 히스토그램의 거리 척도를 제안하였다. Kim과 Billard (2013)는 Gowda와 Diday (1991), Ichino와 Yaguchi (1994), De Carvalho (1994, 1998)가 제안한 구간 데이터에 대한 척도들을 히스토그램의 평균과 분산을 이용하여 히스토그램에 대한 비유사성 척도들로 확장하였다. 또한, Kim과 Billard (2012)는 멀티모달 데이터에 대한 비유사성 척도를 소개하였고, 이를 종합하여 Billard와 Kim (2013)은 혼합형 심볼릭 데이터에 대한 척도를 제안하였다. 본 연구에서는 4절의 자료분석을 위해 히스토그램과 멀티모달 데이터가 혼합된 형태의 심볼릭 데이터에 대한 Euclidean 제곱 거리(squared Euclidean distance)를 사용한다. 혼합형 심볼릭 관찰값 $\mathbf{y}_{i_1}$과 $\mathbf{y}_{i_2}$에 대한 Euclidean 제곱 거리는 다음과 같다:

$$D(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \sum_{j=1}^{p} \sum_{k=1}^{t_j} \left(\pi_{i_1 jk} - \pi_{i_2 jk}\right)^2 . \tag{3.1}$$

식 (3.1)의 Euclidean 제곱 거리는 0과 1 사이의 값만을 갖는 가중값 $\pi_{ijk}$에 의해 정의되므로 각 변수의 단위와 무관한 척도(normalized measure)이다.

계층 분할적 군집분석(divisive hierarchical cluster analysis)은 관찰값들의 전체집합 $\Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$으로 부터 시작하여 각 단계에서 하나의 그룹이 두 개의 그룹으로 나누어지면서 분할이 진행된다. 즉, $r$번째 단계에서의 분할 $P_r = \{C_u^r,\ u = 1, \ldots, r\}$이고 $r$개의 군집 $C_u^r$를 갖는다. 각 단계에서 하나의 군집이 선택되고 그 군집이 두 개의 군집으로 나누어진다. 이제 문제는 어떠한 군집이 어떠한 두 개의 군집으로 나누어질 것인가를 결정하는 것이다. 이를 위해 본 연구에서는 지도학습방법(supervised learning method)의 하나인 분류와 회귀나무(Classification And Regression Tree; CART) 모형의 방식을 이용한다. 나누어질 군집과 그 군집이 나누어진 두 개의 군집을 찾기 위해 한 번에 하나의 변수를 검색하는 방식으로 최적의 분할을 찾기를 시도한다.

그러기 위해서는 먼저 최적의 분할을 결정하는 기준이 필요하다. 잘 분류된 군집들은 군집 내부의 변동은 작고 군집간의 변동은 커야 한다. 본 연구에서는 그 기준으로써 군집 내부의 변동을 최소화하는 것을 사용한다.

**정의 3.1** 주어진 $r$번째 분할 $P_r = \{C_u^r,\ u = 1, \ldots, r\}$에 대해 *총 군집내 변동* $W(P_r)$은 다음과 같이 정의된다.

$$W(P_r) = \sum_{u=1}^{r} I\left(C_u^r\right), \tag{3.2}$$

여기서 $I(C_u^r)$는 군집 $C_u^r$에 대한 군집내 변동으로 군집 $C_u^r = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$에 대해 다음과 같다:

$$I(C_u^r) = \frac{1}{n_u} \sum_{i_1=1}^{n_u} \sum_{i_1 < i_2} D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}), \tag{3.3}$$

여기서 $D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$는 두 심볼릭 관찰값 $\mathbf{y}_{i_1}$과 $\mathbf{y}_{i_2}$ 간의 거리 또는 비유사성 척도에 제곱한 값이다.

결국, 우리의 목적은 각 단계에서 $W(P_r)$을 최소화하는 분할을 찾는 것이다. $W(P_r)$의 계산은 전체 관찰값들이 포함된 계산을 요구한다. 이는 계산상 비효율적이므로, $W(P_r)$의 최소화 문제는 두 개의 군

집으로 분할될 군집의 군집내 변동에서 분할된 두 개의 군집의 군집내 변동의 합의 차를 최대화하는 문제로 대체될 수 있다. 즉, 분할될 군집을 $C_u$이라고 하고 $C_u$가 두 개로 분할되어서 나온 군집을 $C_{u1}$과 $C_{u2}$라고 하면, 결과한 $(r+1)$번째 분할의 총 군집내 변동은 $W(P_{r+1}) = W(P_r) - \{I(C_u) - I(C_{u1}) - I(C_{u2})\}$이고, 그러므로 $W(P_{r+1})$을 최소화하는 문제는 식 (3.4)를 최대화하는 문제와 같다.

$$d_u = I(C_u) - I(C_{u1}) - I(C_{u2}) \tag{3.4}$$

따라서, 모든 관찰값들이 포함된 총 군집내 변동을 계산할 필요없이 각 군집에 대해 식 (3.4)의 $d_u$를 계산하여 최대값을 갖는 분할될 군집 $C_u$와 분할된 군집 $(C_{u1}, C_{u2})$을 찾을 수 있다.

이제 문제는 어떻게 식 (3.4)를 최대화하는 분할을 효율적으로 찾을것인가로 귀결된다. 이 문제를 해결하기 위해 한 번에 하나의 변수를 검색하는 방식을 이용한다. 먼저, 히스토그램 변수의 경우에는 각 히스토그램에 대한 중앙값을 이용하여 두 개의 군집으로 나뉘어지는 점을 찾는다. 즉, 어떤 하나의 군집 $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$으로부터 분할되는 두 개의 새로운 군집 $(C_{u1}, C_{u2})$는 히스토그램 변수 $Y_j$와 각 변수에 대한 히스토그램의 중앙값을 기준으로 나누어지는 점 $s$에 의해 결정된다. 주어진 히스토그램 변수 $Y_j$와 분할점 $s$에 의해 결정되는 두 개의 군집 $(C_{u1}(j,s), C_{u2}(j,s))$을 정의하면 다음과 같다:

$$C_{u1}(j,s) = \{\mathbf{y}_i | \text{Med}(y_{ij}) \le s\} \quad \text{and} \quad C_{u2}(j,s) = \{\mathbf{y}_i | \text{Med}(y_{ij}) > s\}, \tag{3.5}$$

여기서 $\text{Med}(y_{ij})$는 $j$번째 변수에 대한 $i$번째 히스토그램의 중앙값이고 이는 히스토그램의 각 구간 내에서는 자료가 균일하게 분포되어있다는 가정하에서 누적 가중값이 0.5가 되는 점을 뜻한다. 즉, $\text{Med}(y_{ij})$는

$$\text{Med}(y_{ij}) = a_{jk'} + (a_{j,k'+1} - a_{jk'}) \left( \frac{0.5 - \sum_{l=1}^{k'-1} \pi_{ijl}}{\pi_{ij,k'}} \right), \tag{3.6}$$

여기서 $k' = \min\{k | \sum_{l=1}^{k} \pi_{ijl} > 0.5\}$. 결론적으로 군집분석의 각 단계에서 식 (3.4)를 최대화시키는 나누어질 군집에 해당하는 첨자 $u$, 히스토그램 변수에 해당하는 첨자 $j$, 그리고 나누어지는 점 $s$를 결정해야 한다.

멀티모달 변수의 경우 Kim과 Billard (2012)에서 제안된 방법을 이용하여 식 (3.4)를 최대화시키는 분할을 찾는다. 각 멀티모달 변수는 식 (2.2)에서 보여지듯이 여러 개의 항목들 $b_{jk},\ k = 1, \ldots, t_j$를 가지고 있다. 만일 각 항목을 변수로써 취급한다면 최적의 분할을 찾기위해 모든 항목들을 고려해야할 것이고 이것은 군집분석의 속도가 항목의 개수 $t_j$에 의존함을 의미한다. 즉, $t_j$가 크면 클수록 군집분석의 속도가 느려진다. 이를 피하기 위해 Kim과 Billard (2012)는 각 멀티모달 변수에 대한 대표 항목을 선택하는 방법을 제안하였다. 즉, 모든 항목을 고려하는 대신에 각 변수에 대해 하나의 대표 항목을 정해서 그 대표 항목만을 검색하여 최적의 분할을 찾는다. 대표 항목은 어떤 한 항목과 그 항목을 제외한 나머지 항목들 간의 연관성을 나타내는 척도로써 정해진다.

**정의 3.2** 군집 $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$에 있는 관찰값들의 멀티모달 변수 $Y_j$에 대해 항목 $b_{jk}$와 나머지 항목들 $b_{jk'},\ k' = 1, \ldots, t_j,\ k \neq k'$ 사이의 연관성 척도 $\psi_{jk}^u$는 다음과 같이 정의된다:

$$\psi_{jk}^u = \sum_{k'=1, k' \neq k} \left| \left\{ \sum_{i=1}^{n_u} \pi_{ijk} \pi_{ijk'} \right\} \left\{ \sum_{i=1}^{n_u} (1-\pi_{ijk}) \left(1-\pi_{ijk'}\right) \right\} - \left\{ \sum_{i=1}^{n_u} \pi_{ijk} \left(1-\pi_{ijk'}\right) \right\} \left\{ \sum_{i=1}^{n_u} (1-\pi_{ijk}) \pi_{ijk'} \right\} \right|,$$
$$u = 1, \ldots, r,\ k = 1, \ldots, t_j. \tag{3.7}$$

식 (3.7)의 연관성 척도는 항목들 사이의 독립성 가정하에서 $2 \times 2$표에서 두 항목 간의 관계의 척도를

근거로 한다. 일반적으로 심볼릭 데이터 연구에서는 하나의 멀티모달 값에서는 항목 간의 연관성에 대한 정보가 없기 때문에 항목들 간의 독립성을 가정한다. 결론적으로 군집 $C_u$에 대해 식 (3.7)의 $\psi_{jk}^u$ 값이 가장 큰 항목이 $j$번째 멀티모달 변수의 대표 항목으로 선정되고 이 항목을 이용하여 최적의 분할을 찾기를 시도한다. 따라서, 멀티모달 변수에서 어떤 하나의 군집 $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$로부터 분할된 두 개의 새로운 군집 $(C_{u1}, C_{u2})$은 멀티모달 변수 $Y_j$와 그 멀티모달 변수에 해당하는 대표 항목의 가중값을 기준으로 나뉘어지는 점 $m$에 의해 결정된다. 주어진 멀티모달 변수 첨자 $j$와 분할점 $m$에 의해 결정되는 두 개의 군집 $(C_{u1}(j,m), C_{u2}(j,m))$을 정의하면 다음과 같다:

$$C_{u1}(j,m) = \{\mathbf{y}_i | \pi_{ijk^*} \le m\} \quad \text{and} \quad C_{u2}(j,m) = \{\mathbf{y}_i | \pi_{ijk^*} > m\}, \tag{3.8}$$

여기서 $\pi_{ijk^*}$는 군집 $C_u$에서 멀티모달 변수 $Y_j$에 대한 대표 항목 $b_{jk^*}$에 대응하는 가중값이다. 히스토그램 변수처럼 멀티모달 변수에 대해서도 군집분석의 각 단계에서 식 (3.4)를 최대화시키는 나누어질 군집에 해당하는 첨자 $u$, 멀티모달 변수에 해당하는 첨자 $j$, 그리고 분할점 $m$을 결정한다.

결과적으로 히스토그램 변수와 멀티모달 변수를 갖는 혼합형 심볼릭 데이터에 대한 계층 분할적 군집분석의 알고리즘은 다음과 같다:

Step 1. $P_1 \equiv \Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$이고 $r = 1$로 놓는다.

Step 2. $r$번째 단계에서 분할 $P_r = \{C_u^r,\ u = 1, \ldots, r\}$이고, 군집 $C_u^r = \{\mathbf{y}_1^u, \ldots, \mathbf{y}_{n_u}^u\}$이고 $\sum_{u=1}^r n_u = n$이다.

Step 3. 군집 $C_u^r,\ u = 1, \ldots, r$과 각 심볼릭 변수 $Y_j,\ j = 1, \ldots, p$에 대해서 반복하라.

(1) 만일 심볼릭 변수 $Y_j$가 히스토그램 변수라면,

(a) 식 (3.6)을 이용하여 히스토그램 변수 $Y_j$에 대한 $i$번째 히스토그램 값의 중앙값 $\mathrm{Med}(y_{ij}^u),\ i = 1, \ldots, n_u$를 계산한다.

(b) 계산된 중앙값의 크기를 이용하여 오름차순으로 군집 $C_u^r$에 속한 심볼릭 관찰값을 정렬한다.

(2) 만일 심볼릭 변수 $Y_j$가 멀티모달 변수라면,

(a) 멀티모달 변수 $Y_j$에 대한 항목 $b_{jk},\ k = 1, \ldots, t_j$ 각각에 대해 식 (3.7)를 이용하여 연관성 척도 $\psi_{jk}^u,\ \ k = 1, \ldots, t_j$를 계산한 후, 가장 큰 연관성 척도값을 갖는 항목을 변수 $Y_j$에 대한 대표 항목으로 정한다.

(b) 대표항목을 $b_{jk^*}$라고 하면 그에 대응하는 멀티모달의 가중값 $\pi_{ijk^*},\ i = 1, \ldots, n_u$를 이용하여 오름차순으로 군집 $C_u^r$에 속한 심볼릭 관찰값을 정렬한다.

(3) 정렬된 관찰값을 $\mathbf{y}_{(1)}^u, \ldots, \mathbf{y}_{(n_u)}^u$라고 하면, $C_u^r$로부터 분할된 두 개의 군집 $C_{u1}^{rjl} = \{\mathbf{y}_{(1)}^u, \ldots, \mathbf{y}_{(l)}^u\}$이고 $C_{u2}^{rjl} = \{\mathbf{y}_{(l+1)}^u, \ldots, \mathbf{y}_{(n_u)}^u\}$, $l = 1, \ldots, n_u - 1$이 된다 (즉, 변수 $Y_j$에 대해 $(n_u - 1)$개의 분할 $(C_{u1}^{rjl}, C_{u2}^{rjl})$이 존재).

(4) $(n_u - 1)$개의 분할 $(C_{u1}^{rjl}, C_{u2}^{rjl})$ 각각에 대해 식 (3.4)를 계산한다. 즉,

$$d_u^{jl} = I\left(C_u^r\right) - I\left(C_{u1}^{rjl}\right) - I\left(C_{u2}^{rjl}\right), \quad l = 1, \ldots, n_u - 1. \tag{3.9}$$

Step 4. $r$번째 단계에서의 모든 군집들 $C_u^r,\ u = 1, \ldots, r$과 모든 변수들 $Y_j,\ j = 1, \ldots, p$에 대해 $d_u^{jl}$의 최대값을 구한다.

$$\max_{u,j,l} \left\{ d_u^{jl},\ u = 1, \ldots, r,\ j = 1, \ldots, p,\ l = 1, \ldots, n_u \right\}. \tag{3.10}$$

그 최대값에 해당하는 $u\,j,\,l$로써 분할되어지는 군집 $C_u^r$과 분할된 두 개의 새로운 군집 ($C_{u1}^{rjl}$, $C_{u2}^{rjl}$)을 결정한다.

Step 5. 식 (3.10)에 대응되는 $u$, $j$, $l$에 의해 두 개의 군집으로 분할되어지는 점(cut point)을 구한다.

(1) 만일 $j$에 해당하는 변수가 히스토그램 변수라면 분할점 $v_r$은 나누어지는 경계에 있는 두 심볼릭 관찰값 $\mathbf{y}_{(l)}^u$와 $\mathbf{y}_{(l+1)}^u$의 변수 $Y_j$에 해당하는 두 히스토그램 $y_{(l)j}^u$와 $y_{(l+1)j}^u$의 평균 히스토그램 $y_{(l,l+1)j}^u$의 중앙값이 된다. 즉, $v_r = \mathrm{Med}(y_{(l,l+1)j}^u)$이고, 여기서 $y_{(l,l+1)j}^u = \{[a_{jk}, a_{j,k+1}),\ (\pi_{(l)jk} + \pi_{(l+1)jk})/2;\ k = 1, \ldots, t_j\}$이다.

(2) 만일 $j$에 해당하는 변수가 멀티모달 변수라면 분할점 $v_r$은 나누어지는 경계에 있는 두 심볼릭 관찰값 $\mathbf{y}_{(l)}^u$와 $\mathbf{y}_{(l+1)}^u$의 변수 $Y_j$의 대표항목 $b_{jk^*}$에 대응하는 가중값들의 평균이 된다. 즉, $v_r = (\pi_{(l)jk^*} + \pi_{(l+1)jk^*})/2$이다.

Step 6. Step 2–5를 $r = n$ 또는 $r = R$이 될 때까지 반복한다. 여기서 $R$은 미리 정한 군집의 개수이다.

분할적 군집분석에서는 $r$번째 단계에서 총 $\sum_{u=1}^{r}(2^{n_u-1} - 1)$개의 분할들이 존재한다. 하지만, 제안된 알고리즘은 히스토그램과 멀티모달 변수 각각의 개수에 상관없이 군집분석의 각 단계에서 $\sum_{u=1}^{r} p(n_u - 1)$개의 분할들만 고려하여 최적의 분할 찾기를 시도한다.

분할점(cut point)은 각 단계에서 어떠한 변수에 의해 그리고 무엇을 기준으로 군집이 분할되었는지에 대한 정보를 제공함으로써 군집분석의 결과에 대한 해석을 돕는다. Step 5에서 히스토그램 변수의 경우 분할점 $v_r$은 분할점에 인접한 두 히스토그램의 평균이 되는 히스토그램의 중앙값이다. 히스토그램 변수에 의해 분할되는 두 개의 군집을 찾기 위해 제안된 알고리즘은 중앙값을 이용하여 심볼릭 관찰값을 정렬하고 그 순서를 이용하여 최적의 분할을 찾는다. 따라서, 기준이 중앙값이므로 분할점 역시 경계에 있는 두 히스토그램으로 부터의 중앙값에 의해 결정되는 것이 자연스럽다. 물론, 심볼릭 관찰값의 정렬의 기준으로 중앙값 대신 평균을 이용할 수 있으나 일반적으로 히스토그램 자체가 원자료에 대한 요약이므로 만일 원자료에 이상값(outlier)이 있다면 평균이 크게 변동될 수 있으므로 중앙값을 이용하는 것이 이상값에 로버스트(robust)하다는 이점을 가진다. 마찬가지로 멀티모달 변수의 경우에는 대표항목의 가중값에 의해 관찰값들이 정렬되므로, 분할점은 경계에 있는 대표항목의 두 가중값의 평균이 된다.

이 분할점을 이용하면 군집분석의 각 단계에서 이진 질문(binary question)을 제공할 수 있다. 즉, 히스토그램 변수의 경우 $\mathrm{Med}(Y_j) \le v_r$ ?와 같은 질문이 주어지고 이것이 의미하는 바는 변수 $Y_j$에 대해 분할점 $v_r$보다 작은 중앙값을 갖는 심볼릭 관찰값들과 그 중앙값보다 큰 관찰값의 그룹으로 두 개의 군집이 분할되었음을 의미한다. 마찬가지로 멀티모달 변수에 대해 $Y_j : b_{jk} \le v_r$ ?이라는 질문을 가질 수 있고, 이는 변수 $Y_j$의 항목 $b_{jk}$의 가중값이 $v_r$보다 작은 그룹과 그 가중값이 큰 그룹으로 나뉘어졌음을 의미한다. 이는 심볼릭 자료의 구조를 이해하고 해석하는데 도움이 된다.

군집분석의 결과에 대해 최적의 군집의 개수를 결정하는 것은 군집분석에서 매우 중요한 일 중 하나이다. 이를 위해 Kim과 Billard (2011)은 Dunn (1974)과 Davis와 Bouldin (1979)이 제안한 군집 유효성 지수(cluster validity index)를 식 (3.3)과 식 (3.4)를 이용하여 확장하였다. 이 두 확장된 지수를 소개하면 다음과 같다:

**정의 3.3** $r$번째 분할 $P_r = (C_1^r, \ldots, C_r^r)$에 대한 확장된 Dunn 지수(extended Dunn index)는 다음과 같이 주어진다:

$$\mathrm{DI}_r = \min_{u=1,\ldots,r}\left[\min_{t=1,\ldots,r,t\neq u}\left\{\frac{I(C_t^r \cup C_u^r) - I(C_t^r) - I(C_u^r)}{\max_{l=1,\ldots,r}\{I(C_l^r)\}}\right\}\right], \quad r = 2, \ldots, n-1, \tag{3.11}$$

여기서 $I(\cdot)$은 식 (3.3)에서 정의된 군집내 변동이다.

식 (3.11)의 분자는 군집들 사이가 얼마나 잘 분리되었는지를 측정하고 분모는 군집 내부의 변동이 얼마나 작은가를 측정한다. 일반적으로 군집간의 변동이 크고 군집내의 변동이 작을 경우 잘된 군집분석결과로 고려되므로, Dunn 지수가 가장 큰 값을 갖는 분할이 최적의 분할로 식별될 수 있다.

**정의 3.4** $r$번째 분할 $P_r = (C_1^r, \ldots, C_r^r)$에 대한 확장된 Davis-Bouldin 지수(extended Davis-Bouldin index)는 다음과 같이 주어진다:

$$\mathrm{DB}_r^s = \frac{1}{r}\sum_{u=1}^{r}\left[\frac{\max_{t=1,\ldots,r,t\neq u}\{I(C_t^r)+I(C_u^r)\}}{\min_{l=1,\ldots,r,l\neq u}\left\{I(C_l^r\cup C_u^r)-I(C_l^r)-I(C_u^r)\right\}}\right], \quad r=2,\ldots,n-1, \tag{3.12}$$

식 (3.12)의 Davis-Bouldin 지수는 Dunn 지수와는 달리 분자가 군집내 변동을 측정하고 분모가 군집간 변동을 측정한다. 또한, Davis-Bouldin 지수는 그 두 변동의 비의 평균을 이용하고, 지수의 가장 작은 값에 대응하는 분할이 주어진 군집분석결과에 대한 최적의 분할로 간주된다.

## 4. 산업재해자료분석

고용노동부에서는 매년 산업재해예방 정책 수립의 기초자료로 활용하기 위해 업종별, 규모별, 발생시기별, 원인별 재해의 분포와 같은 산업재해현황에 대한 통계들을 발표한다. 이러한 통계들은 근로복지공단의 산재 요양승인된 자료를 기반으로 작성된다. 앞서 언급했듯이 산재보험과 산재예방은 업종 단위로 이루어지므로 개개의 사업장 보다는 업종별 분석이 주된 관심이다. 또한, 더 심도 깊은 분석을 위해 통계의 원자료인 개인에 대한 요양승인자료를 이용하면 좋겠지만 개인정보보호의 이유로 그러한 자료는 현실적으로 쉽게 이용할 수 없다. 이러한 상황하에서 업종들에 대한 군집분석을 할 수 있는 방법으로 심볼릭 데이터 기법을 고려할 수 있다. 본 연구에서는 앞서 제안한 분할적 군집분석방법을 이용하여 한국의 업종별 산업재해자료를 분석하고자 한다. 자료는 고용노동부에서 발간한 2013년 산업재해현황분석에서 수집되었으며 고용노동부 홈페이지 `http://www.moel.go.kr`에서 이용가능하다.

제 1절에서 언급하였듯이 산재보험의 보험요율은 같은 업종 내의 사업장들에 대해 동일하게 적용된다. 따라서, 산업재해의 발생현황과 재해의 심도를 업종별로 분석해 보는 것은 의미있는 일이다. 따라서, 본 연구에서는 57개의 업종별로 작성된 산업재해현황에 대한 4개의 통계표로 부터 구한 심볼릭 데이터를 이용하여 업종들의 군집분석을 실시한다. 산업재해현황분석에 있는 4개의 통계표는 사업장 규모별, 근속기간별, 요양기간별, 발생형태별 산업재해가 발생한 건수를 보여준다. 각 업종은 심볼릭 관찰값의 대상(object)이 될 수 있고 그 각각의 통계표는 심볼릭 변수로써 고려될 수 있다. 따라서, 4개의 심볼릭 변수가 자료에 존재하고, 각 변수는 $Y_1$ = '규모별 산업재해현황', $Y_2$ = '근속기간별 산업재해현황', $Y_3$ = '요양기간별 산업재해현황', $Y_4$ = '발생형태별 산업재해현황'으로 정의될 수 있다. 규모별 산업재해현황 변수에서는 업종들의 대기업과 중소기업의 재해형태가 비슷한 업종들을 묶을 수 있을 것으로 기대한다. 근속기간별 변수에서는 작업에 대한 숙련도에 따른 재해의 패턴(pattern)이 유사한 업종들이 무엇인지 볼 수 있을 것이고, 요양기간별 변수에서는 산업재해가 발생했을 때 그 재해의 심도가 비슷한 업종들을 조사할 수 있을 것으로 기대한다. 마지막으로 발생형태별 변수에서는 재해발생유형이 비슷한 업종들을 볼 수 있을 것이다. 변수 $Y_1$과 $Y_2$는 히스토그램 변수이고, $Y_3$과 $Y_4$는 멀티모달 변수이다. $Y_3$의 경우 이는 요양기간에 대한 변수이지만 '사망'이라는 질적인 항목이 있기 때문에 멀티모달 변수로 취급한다. 멀티모달 변수 $Y_3$과 $Y_4$의 항목들은 Table A.2에서 볼 수 있다. 또한, Table A.1은 57개의 업종들을 보여준다.

Figure 4.1은 3절에서 소개된 군집방법과 식 (3.1)의 거리척도를 사용하여 57개 업종을 산업재해의 형태와 특징에 의해 분류한 댄드로그램(dendrogram)을 보여준다. Figure 4.1에서 댄드로그램의 세로축

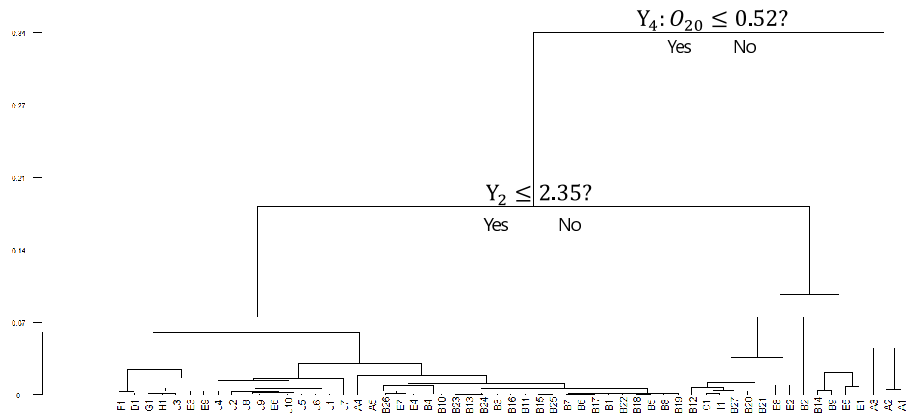


**Figure 4.1.** Dendrogram for industrial accident data.

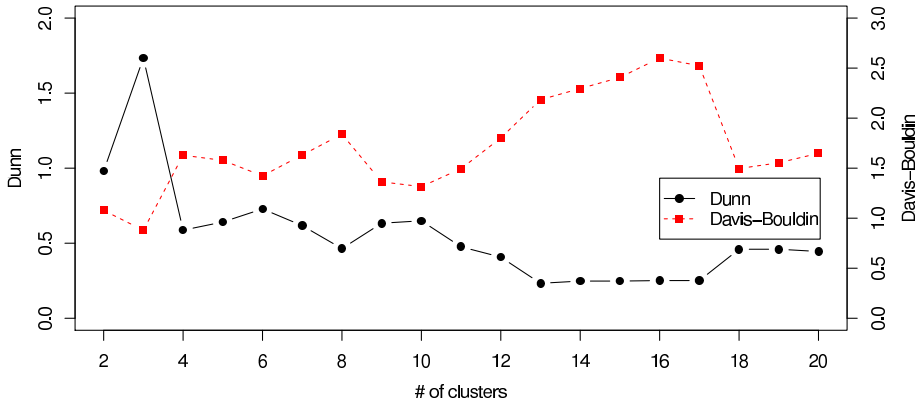


**Figure 4.2.** Cluster validity indexes for industrial accident data.

**Table 4.1.** Three clusters

| 군집 | 업종 |
|---|---|
| 군집1 | A1, A2, A3 |
| 군집2 | A4, A5, B1, B10, B11, B13, B15, B16, B17, B18, B19, B20, B22, B23, B24, B25, B26, B3, B4, B5, B6, B7, B8, D1, E3, E4, E6, E7, E9, F1, G1, H1, J1, J10, J2, J3, J4, J5, J6, J7, J8, J9 |
| 군집3 | B2, B9, B12, B14, B20, B21, B27, C1, E1, E2, E5, E8, I1 |

은 식 (3.2)의 총 군집내 변동값을 나타낸다. Figure 4.2는 최적의 군집의 개수를 결정하기위해 식 (3.11)과 식 (3.12)를 이용하여 구한 각 분할에 대한 두 개의 지수 값들을 나타낸다. Figure 4.2로 부터 Dunn 지수는 군집이 세 개인 분할에 대해 가장 큰 값을 가졌으며 Davis-Bouldin 지수 역시 군집이 세 개 일 때 가장 작은 값을 나타냈다. 따라서, Figure 4.1의 군집결과에 대해 세 개의 군집을 갖는 분할이 가장 분류가 잘 된 결과로 식별되었고, 그 세 개 군집들 내의 업종들은 Table 4.1에서 보여준다.

군집분석의 첫 번째 단계에서 전체 57개 업종은 두 개의 군집으로 분할되었고 그 분할은 Table 4.1의 (군집1)과 (군집2, 3)으로 구성된다. 이 분할에 대응하는 이진 질문은 $Y_4 : O_{20} \leq 0.52$?이었으며, 이는 산업재해의 발생형태에서 군집1에 해당하는 업종은 업무상 질병의 비율이 0.52보다 높았다는 것을 의

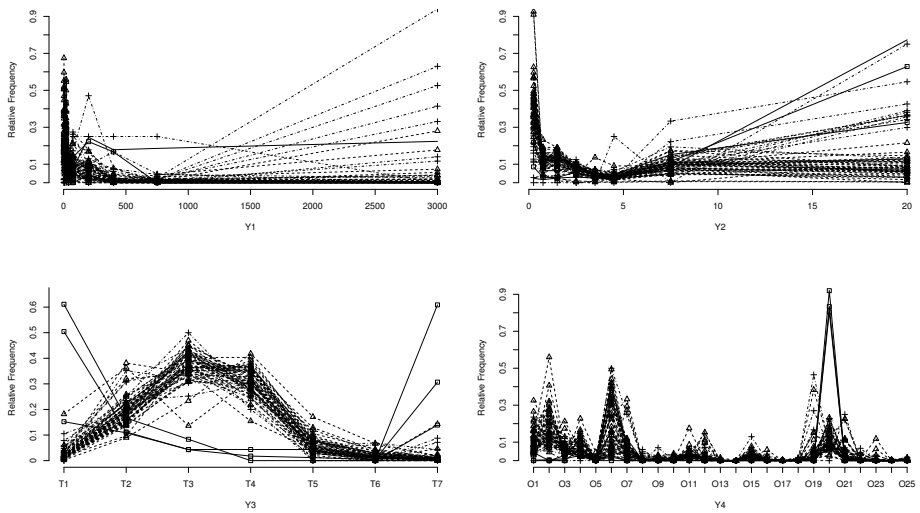


**Figure 4.3.** Distributions of each cluster for each symbolic variable; cluster1 (□), cluster2 (△) cluster3 (+).

미한다. 이에 반해 군집2와 3은 업무상 질병의 비율이 0.52보다 낮았다는 뜻이다. 군집1에 속한 업종은 석탄광업(A1), 금속 및 비금속 광업(A2), 채석업(A3)으로 광업에 속하는 업종들이었고 업무상 질병이 다른 업종들 보다 현저하게 높았다. 광산업 중에서 석탄광업과 금속 및 비금속 광업은 작업장이 탄광이고 탄광에서의 지속적인 작업은 호흡기와 폐질환에 질병을 유발할 가능성이 높다. 이로 인해 업무상 질병의 비율이 타업종에 비해 현저히 높은 것으로 보인다. 두 번째 분할에서는 (군집2, 3)이 (군집2)와 (군집3)으로 심볼릭 변수 $Y_2$인 근속기간에 의해 나뉘어졌다. 즉, 근속기간에 대한 중앙값이 2.35보다 작은 업종들은 군집2로 분류되었고, 그 값보다 큰 중앙값을 갖는 업종들은 군집3으로 분류되었다. 군집3은 7개의 제조업과 전기, 가스, 증기 및 수도사업(C1), 운수, 창고 및 통신업의 4개 업종, 금융 및 보험업(I1)의 13개 업종을 포함한다. 이들 업종에서는 산업재해를 당한 사람들 중 상대적으로 군집2에 속한 업종들에 비해 근속년수가 높은 직원들의 비율이 높았다는 것을 알 수 있다.

Figure 4.3은 각 업종들의 각 심볼릭 변수에 대한 상대도수(가중값)를 꺾은선 그래프로 보여주고 있다. 첫 번째 분할이 일어난 변수 $Y_4$를 보면 군집1에 해당하는 □선이 $O_{20}$ 항목에서 타군집들에 비해 훨씬 높이 솟아있음을 볼 수 있다. 이는 첫 번째 분할에 대한 이진 질문의 결과와 일치한다. 따라서, 군집 알고리즘이 분할되는 지점을 잘 찾았다고 볼 수 있다. 두 번째 분할은 $Y_2$에 의해 일어났으므로, Figure 4.3의 변수 $Y_2$를 보면 군집2에 해당하는 △선들은 짧은 근속년수에서 높은 빈도를 보인 반면 근속년수가 긴 경우에 낮은 재해율을 보였다. 그에 반해 군집3에 해당하는 +선은 높은 근속년수에서 상대적으로 높은 상대도수를 보였다. 요양기간별($Y_3$)의 결과를 보면 군집1은 재해가 사망($T_1$) 또는 4–7일의 요양기간($T_7$)을 요하는 가벼운 부상의 비율이 높은 반면 군집2와 3은 1–6개월 사이의 요양($T_3$, $T_4$)을 요하는 재해의 비율이 높았다. 따라서, Figure 4.3의 결과는 군집분석이 업종들의 각 변수에 대한 분포에 맞게 잘 이루어졌음을 방증한다.

## 5. 결론

심볼릭 데이터는 기존의 전통적인 데이터들과는 달리 관찰값 자체가 내부적으로 변동을 갖는 형태이다. 이는 전통적인 데이터를 우리가 관심이 있는 범주나 그룹으로 요약한 형태로써 고려될 수 있는데 대용량

의 자료들이 많은 현실에서 유용한 방법이 될 수 있다. 또한, 정부기관에서 발행하는 통계표들을 분석할 때도 사용될 수 있다.

본 연구에서는 심볼릭 데이터의 종류들인 히스토그램과 멀티모달 데이터가 혼합되어있는 자료에 대한 분할적 군집분석방법을 제안하였다. 이 방법은 히스토그램 데이터에 대한 방법과 멀티모달 데이터에 대한 방법을 확장하고 결합한 형태로 각 단계에서 한 번에 하나의 변수를 이용하여 최적의 분할을 찾기를 시도한다. 이러한 방법의 장점으로는 분할적 군집분석방법의 문제점 중 하나인 각 단계에서 검토해야 할 $2^{n-1}-1$개의 많은 분할들을 $p(n-1)$개의 분할들로 줄일 수 있다는 점이다. 그럼으로써 군집분석의 속도를 높일 수 있다. 또한, 한 번에 하나의 변수를 이용하므로 각 단계에서 어떠한 변수에 의해 그리고 어떤 지점에서 군집이 나누어졌는지에 대한 정보를 제공함으로써 군집분석결과의 해석을 도울 수 있다. 제안된 방법에서 히스토그램 데이터에 대한 군집분석은 각 히스토그램의 중앙값을 이용하여 최적의 분할을 찾는다. 이는 히스토그램의 평균을 이용하는 방법보다 이상값에 로버스트(robust)하다. 히스토그램 데이터는 원자료를 요약한 형태이고 원자료에 이상값들이 존재한다면 히스토그램으로 부터 구한 평균은 실제 그룹의 평균과 다를 수 있다.

그러나, 이렇게 한 번에 하나의 변수를 사용하는 계층 분할적 군집분석방법은 군집의 구조가 여러 개의 변수들의 조합에 의존하는 경우에는 만족스러운 군집을 구하지 못할 수 있고 (Chavent, 1998), 자료의 일부만 변해도 군집의 결과가 매우 달라질 수 있다는 점에 주의해야만 한다.

## 부록: 산업재해자료

**Table A.1.** 57 types of business

| | 코드 업종 | | 코드 업종 |
|---|---|---|---|
| 광업 | A1 석탄광업 | 제조업 | B25 섬유 및 섬유제품 제조업(을) |
| | A2 금속 및 비금속광업 | | B26 자동차 및 모터사이클 수리업 |
| | A3 채석업 | | B27 코크스, 연탄 및 석유정제품 제조업 |
| | A4 석회석광업 | | C1 전기, 가스, 증기 및 수도사업 |
| | A5 기타광업 | | D1 건설업 |
| 제조업 | B1 식료품 제조업 | 운수창고 및 통신업 | E1 철도, 궤도 및 삭도 운수업 |
| | B2 담배 제조업 | | E2 여객자동차 운수업 |
| | B3 섬유 또는 섬유제품 제조업(갑) | | E3 화물자동차 운수업 |
| | B4 목재 및 나무제품 제조업 | | E4 수상 운수업, 항만하역 및 화물취급사업 |
| | B5 펄프, 지류 제조업 및 제본 또는 인쇄물 가공업 | | E5 항공 운수업 |
| | B6 신문, 화폐발행, 출판업 및 인쇄업 | | E6 운수관련 서비스업 |
| | B7 화학제품 제조업 | | E7 창고업 |
| | B8 의약품 및 화장품 향료 제조업 | | E8 통신업 |
| | B9 고무제품 제조업 | | E9 소형화물 운수업 및 택배업 퀵서비스업 |
| | B10 유리 제조업 | | F1 임업 |
| | B11 도자기 및 기타요업제품 제조업 | | G1 어업 |
| | B12 시멘트 제조업 | | H1 농업 |
| | B13 비금속광물제품 및 금속제품 제조업 | | I1 금융 및 보험업 |
| | B14 금속제련업 | 기타의 사업 | J1 검물등의 종합관리사업 |
| | B15 금속재료품 제조업 | | J2 위생 및 유사 서비스업 |
| | B16 도금업 | | J3 기타의 각종사업 |
| | B17 기계기구 제조업 | | J4 해외 파견자 |
| | B18 전기기계기구 제조업 | | J5 전문기술 서비스업 |
| | B19 전자제품 제조업 | | J6 보건 및 사회복지사업 |
| | B20 선박건조 및 수리업 | | J7 교육 서비스업 |
| | B21 수송용 기계기구 제조업 | | J8 도소매 및 소비자용품 수리업 |
| | B22 계량기, 광학기계, 기타정밀기구 제조업 | | J9 부동산업 및 임대업 |
| | B23 수제품 제조업 | | J10 오락, 문화 및 운동관련사업 |
| | B24 기타 제조업 | | |

**Table A.2.** Items of multimodal variables $Y_3$ and $Y_4$

| $Y_3$ | | $Y_4$ | | | |
|---|---|---|---|---|---|
| 코드 | 항목 | 코드 | 항목 | 코드 | 항목 |
| $T_1$ | 사망자 | $O_1$ | 떨어짐 | $O_{14}$ | 광산사고 |
| $T_2$ | 6개월 이상 | $O_2$ | 넘어짐 | $O_{15}$ | 불균형 및 무리한 동작 |
| $T_3$ | 91–180일 | $O_3$ | 부딪힘 | $O_{16}$ | 화학물질 누출접촉 |
| $T_4$ | 29–90일 | $O_4$ | 물체에 맞음 | $O_{17}$ | 산소결핍 |
| $T_5$ | 15–28일 | $O_5$ | 무너짐 | $O_{18}$ | 사업장내 교통사고 |
| $T_6$ | 8–14일 | $O_6$ | 끼임 | $O_{19}$ | 사업장외 교통사고 |
| $T_7$ | 4–7일 | $O_7$ | 절단, 베임, 찔림 | $O_{20}$ | 업무상 질병 |
| | | $O_8$ | 감전 | $O_{21}$ | 체육행사 |
| | | $O_9$ | 폭발, 파열 | $O_{22}$ | 폭력행위 |
| | | $O_{10}$ | 화재 | $O_{23}$ | 동물상해 |
| | | $O_{11}$ | 깔림, 뒤집힘 | $O_{24}$ | 기타 |
| | | $O_{12}$ | 이상온도접촉 | $O_{25}$ | 분류불능 |
| | | $O_{13}$ | 빠짐, 익사 | | |

## References


Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley and Sons, New Jersey.

Billard, L. and Kim, J. (2013). Clustering in contemporary mixed-valued data, In *Proceedings of the 2013 World Statistics Congress*, International Statistical Institute.

Bock, H. H. and Diday, E. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, New York.

Cha, S. H. and Srihari, S. H. (2002). On measuring the distance between histograms, *Pattern Recognition Letter*, **35**, 1355–1370.

Chavent, M. (1998). A monothetic clustering method, *Pattern Recognition Letters*, **19**, 989–996.

Chavent, M. (2000). Criterion-based divisive clustering for symbolic data. In: Bock, H.H., Diday, E. (Eds.), *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, New York, 299–311.

Davis, D. L. and Bouldin, D. W. (1979). A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1**, 224–227.

De Carvalho, F. A. T. (1994). Proximity coefficients between boolean symbolic objects. In: Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., (Eds.), *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin, 387–394.

De Carvalho, F. A. T. (1998). Extension based proximity coefficients between constrained boolean symbolic objects. In: Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., Baba, Y., (Eds.), In *Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96)*, Springer-Verlag, Berlin, 370–378.

De Carvalho, F. A. T., Brito, P. and Bock, H. H. (2006). Dynamic clustering for interval data based on $L_2$ distance, *Computational Statistics*, **2**, 231–245.

De Carvalho, F. A. T. and Lechevallier, Y. (2009). Partitional clustering algorithms for symbolic interval data based on single adaptive distances, *Pattern Recognition*, **42**, 1223–1236.

De Carvalho, F. A. T. and De Souza, R. M. C. R. (2010). Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letters*, **31**, 430–443.

De Souza, R. M. C. R. and De Carvalho, F. A. T. (2007). A clustering methods for mixed feature-type symbolic data using adaptive squared Euclidean distances, *The 7th International Conference on Hybrid Intelligent Systems*, 168–173.

Diday, E. (1987). Introduction à l'approche symbolique en analyse des données, *Première Journées Symbolique-Numérique*, CEREMADE, Université Paris IX, 21–56.

Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions, *Journal of Cybernetica*, **4**, 95–104.

Gowda, K. C. and Diday, E. (1991). Symbolic clustering using a new dissimilarity measure, *Pattern Recognition*, **24**, 567–578.

Gowda, K. C. and Ravi, T. V. (1995a). Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity, *Pattern Recognition Letters*, **16**, 647–652.

Gowda, K. C. and Ravi, T. V. (1995b). Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity, *Pattern Recognition*, **28**, 1277–1282.

Ichino, M. and Yaguchi, H. (1994). Generalized minkowski metrics for mixed feature type data analysis, *IEEE Transactions on Systems, Man, and Cybernetics*, **24**, 698–709.

Irpino, A. and Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data, *IFCS 2006*, 185–192.

Kim, J. and Billard, L. (2011). A polythetic clustering process and cluster validity indexes for histogram-valued objects, *Computational Statistics & Data Analysis*, **55**, 2250–2262.

Kim, J. and Billard, L. (2012). Dissimilarity measures and divisive clustering for symbolic multimodal-valued data, *Computational Statistics & Data Analysis*, **56**, 2795–2808.

Kim, J. and Billard, L. (2013). Dissimilarity measures for histogram-valued observations, *Communications in Statistics - Theory and Methods*, **42**, 283–303.

# 혼합형태 심볼릭 데이터의 군집분석방법

김재직[a,1]

[a]성균관대학교 통계학과

(2015년 9월 14일 접수, 2015년 11월 2일 수정, 2015년 11월 3일 채택)

**요 약**

오늘날 데이터는 $p$-차원의 공간에서 점들로써 표현되는 전통적인 형태를 벗어나 시그널(signal), 함수, 이미지(image), 모양(shape) 등과 같은 다양한 형태의 자료들이 데이터로써 고려되고 분석되고있다. 그러한 종류의 새로운 종류의 데이터 중 하나로 심볼릭 데이터(symbolic data)를 고려할 수 있다. 심볼릭 데이터는 구간(interval), 히스토그램(histogram), 목록(list), 통계표, 분포, 또는 모형 등과 같은 다양한 형태들을 가질 수 있다. 지금까지의 연구가 주로 심볼릭 데이터의 각각의 형태별 자료를 고려했다면, 본 연구에서는 이를 확장하여 수집된 히스토그램과 멀티모달의 혼합된 형태로 이루어진 자료에 대한 계층 분할적 군집분석방법을 소개하고 이를 업종별 산업재해자료의 분석을 위해 이용한다.

주요용어: 혼합형태 심볼릭 데이터, 군집분석, 산업재해

[1](03063) 서울시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: jaejik@skku.edu