비대칭적 유사도 기반의 심볼릭 객체의 계층적 클러스터링

Hierarchical Clustering of Symbolic Objects based on Asymmetric Proximity

오승준* · 박찬웅**

Seung-Joon Oh* and Chan-Woong Park

*경기과학기술대학교 산업경영학과, **가천대학교 산업경영공학과

* GTEC(Gyeonggi College of Science and Technology)

요 약

패턴 인식, 데이터 분석, 침입 탐지, 이미지 처리, 바이오 인포매틱스 등과 같은 수많은 분야에서 클러스터링 분석이 사용되고 있다. 기존의 많은 연구들은 수치 데이터에만 기반을 두고 있다. 그러나 구간 데이터, 히스토그램, 심지어는 함수들을 값으로 갖는 변수들을 다루는 심볼릭 데이터 분석이 부상하고 있다. 본 논문에서는 이런 심볼릭 데이터들을 클러스터링하기 위하여 비대칭적 유사도를 제안한다. 또한 평균 유사도 값(ASV)에 기반한 클러스터링 방법도 개발한다. 제안하는 클러스터링의 결과는 기존 방법들과 다르며, 매우 고무적인 결과를 보여준다.

키워드 : 비대칭적 유사도, 심볼릭 객체, 클러스터링, 데이터 마이닝

Abstract

Clustering analysis has been widely used in numerous applications like pattern recognition, data analysis, intrusion detection, image processing, bioinformatics and so on. Much of previous work has been based on the numeric data only. However, symbolic data analysis has emerged to deal with variables that can have intervals, histograms, and even functions as values. In this paper, we propose a non symmetric proximity based clustering approach for symbolic objects. A method for clustering symbolic patterns based on the average similarity value(ASV) is explored. The results of the proposed clustering method differ from those of the existing methods and the results are very encouraging.

Key words : Asymmetric Proximity, Symbolic Objects, Clustering, Data Mining

1. 서 론

패턴 인식, 데이터 분석, 침입 탐지, 이미지 처리, 바이오 인포매틱스 등과 같은 수많은 분야에서 클러스터링 분석이 사용되고 있다. 기존의 많은 연구들은 수치 데이터에만 기반을 두고 있다. 그러나 구간 데이터, 히스토그램, 심지어는 함수들을 값으로 갖는 변수들을 다루는 심볼릭 데이터 분석이 부상하고 있다[1-3]. 심볼릭 객체들은 객체들을 개별적으로 표현하기 보다는 객체들 사이에 존재하는 개념들을 표현하기에 좋다. 즉 다양한 개념들을 저장하는데

심볼릭 표현의 강점이 있다.

기존의 데이터 셋들이나 심볼릭 객체들을 다루기 위한 유사도 측정방법들에 대한 많은 연구가 진행되어 왔다[4-10]. [5][6]에서 제안된 방법들은 스캔(속성값들의 상대적 크기), 위치(두 속성들의 상대적 위치), 내용(두 속성들간의 공통 부분)을 고려하여 유사도를 측정한다. [5][6]에서는 구간 데이터 속성들간의 비유사도를 측정하는데 있어 문제점이 발생하는 경우가 있다. 이러한 단점들을 극복하기 위해 [7][10]에서는 수정된 스캔, 위치, 내용을 이용한 유사도 측정방법을 제안하였다. 구간 데이터들간의 유사도를 측정하기 위해 [4]에서는 하이퍼 박스 모델 표현법이 제시되었다. 다중값 데이터들간의 유사도를 측정하기 위한 연구는 [8][9]에서 진행되었다. 그러나 기존 연구들은 구간 데이터나 다중값 데이터를 개별적으로 처리하였으며, 모두 대칭적 유사도 개념을 사용하였다.

본 논문에서는 비대칭적 유사도 개념을 이용하여 유사도 측정을 하고자 한다. 비대칭적 유사도의 개념은 다음과 같이 설명할 수 있다. 예를 들면 일정 그룹의 사람들이 있는데, 이들 사이에 서로 친밀감이 있을 경우, 이들을 가까운 친구로 그룹핑 하는 경우를 생각해 보자. 만약 A와 B라는 두 사람이

접수일자: 2012년 10월 5일

심사(수정)일자: 2012년 12월 4일

게재확정일자: 2012년 12월 7일

* Corresponding author

This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

서로 상대방에 대해 친밀감을 가지고 있는 경우에만 가까운 친구로 생각할 수 있다. 만약 A는 B에게 친밀감을 느끼지만, B는 A에 대해 친밀감을 가지고 있지 않은 경우, A와 B가 가까운 친구가 될 가능성은 상대적으로 낮을 수밖에 없다. 만약 한 사람이 다른 사람을 가까운 친구로 생각하지 않는 경우에는 두 사람은 서로 친밀한 친구로 그룹핑 시키면 안된다. 즉 두 객체가 단지 가까이 근접해서라기 보다는 서로 가까운 이웃을 가지고 있을 때 같은 클러스터로 구성을 해야 한다. 기존의 대칭적 유사도 개념이 아니라 비대칭적 유사도 개념을 사용해서 클러스터링을 해야 하는 경우이다.

본 논문에서는 심볼릭 객체들을 클러스터링 하기 위해 비대칭적 유사도를 제안한다. 새로운 비대칭적 유사도 측정 방법은 다중값 데이터 및 구간 데이터에 적용된다.

기존의 계층적이거나 분할 클러스터링 알고리즘들은 대칭적 유사도 방법을 사용한다. 따라서 본 논문에서는 비대칭적 유사도 방법을 사용하기 위해 ASV에 기반한 클러스터링 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 심볼릭 데이터간의 유사도를 측정하기 위한 비대칭적 유사도를 제안한다. 3장에서는 심볼릭 객체들을 클러스터링하기 위한 ASV 개념을 제안하고 이를 이용한 계층적 클러스터링 방법을 제안한다. 4장에서는 다양한 데이터 셋에 대한 실험결과를 기술하고, 5장에서는 결론을 제시한다.

2. 비대칭적 유사도 측정

본 장에서는 심볼릭 객체들간의 유사도를 측정하기 위한 유사도 방법을 제안한다. 객체들을 클러스터링 하는데 있어서는 유사도 측정방법이 기본 개념이기 때문에, 어떤 유사도 측정방법을 사용하느냐에 따라 클러스터링 결과가 달라질 수 있다.

n 차원공간에 두 개의 심볼릭 객체 O_i 와 O_j 가 있다고 가정하자. 여기서 n 차원은 구간 데이터의 속성 u 와 다중값 데이터의 속성 v 의 합이다.

$$\text{즉, } O_i = I_{i1}, I_{i2}, \dots, I_{iu}, M_{i1}, M_{i2}, \dots, M_{iv},$$

여기서, $I_{ik} = [f_{ik}^-, f_{ik}^+]$, $\forall k=1, 2, \dots, u$ 은 구간 속성

$M_{il} = (m_{il}^1, m_{il}^2, \dots, m_{il}^v)$, $\forall l=1, 2, \dots, v$ 는 다중값 속성

$$O_j = I_{j1}, I_{j2}, \dots, I_{ju}, M_{j1}, M_{j2}, \dots, M_{jv},$$

여기서, $I_{jk} = [f_{jk}^-, f_{jk}^+]$, $\forall k=1, 2, \dots, u$ 은 구간 속성

$M_{jl} = (m_{jl}^1, m_{jl}^2, \dots, m_{jl}^v)$, $\forall l=1, 2, \dots, v$ 는 다중값 속성.

2.1. 구간 데이터의 심볼릭 유사도

k 번째 구간 데이터 속성에 대한 객체 O_i 로부터 O_j 의 유사도는 다음과 같이 정의한다.

$$s_{i \rightarrow j}^k = \frac{|I_{ik} \cap I_{jk}|}{|I_{jk}|} \quad (1)$$

여기서 $I_{ik} \cap I_{jk}$ 는 구간 데이터 속성 I_{ik} 와 I_{jk} 의 겹치는 부분이며, $|\cdot|$ 는 구간의 길이를 나타낸다.

두 개의 구간 속성들 간의 유사도는 속성들 간의

공통부분이 많을수록 크다고 볼 수 있으므로, 이점을 기본개념으로 식(1)을 제안하였다.

마찬가지로 k 번째 구간 데이터 속성에 대한 객체 O_j 로부터 O_i 의 유사도는 다음과 같이 정의한다.

$$s_{j \rightarrow i}^k = \frac{|I_{ik} \cap I_{jk}|}{|I_{ik}|} \quad (2)$$

모든 n개의 구간 데이터 속성에 대한 객체 O_i 로부터 O_j 의 유사도는 다음과 같이 정의한다.

$$S_{i \rightarrow j} = [s_{i \rightarrow j}^1, s_{i \rightarrow j}^2, \dots, s_{i \rightarrow j}^n] = \sqrt{\sum_{k=0}^n (s_{i \rightarrow j}^k)^2} \quad (3)$$

마찬가지로 모든 n개의 구간 데이터 속성에 대한 객체 O_j 로부터 O_i 의 유사도는 다음과 같이 정의한다.

$$S_{j \rightarrow i} = [s_{j \rightarrow i}^1, s_{j \rightarrow i}^2, \dots, s_{j \rightarrow i}^n] = \sqrt{\sum_{k=0}^n (s_{j \rightarrow i}^k)^2} \quad (4)$$

식(3)과 (4)로부터 얻어진 유사도는 비대칭적 유사도임을 알 수 있다.

2.2. 다중값 데이터의 심볼릭 유사도

다중값 데이터에 대한 두 객체간의 유사도는 두 객체 O_i 와 O_j 의 공통 부분으로 측정할 수 있다.

l 번째 다중값 데이터 속성에 대한 객체 O_i 로부터 O_j 의 유사도는 다음과 같이 정의한다.

$$s_{i \rightarrow j}^l = \frac{|M_{il} \cap M_{jl}|}{|M_{jl}|} \quad (5)$$

여기서 $M_{il} \cap M_{jl}$ 는 집합 M_{il} 와 M_{jl} 의 교집합을 의미하며 $|\cdot|$ 구간의 카디널리티를 나타낸다.

두 개의 다중값 속성들 간의 유사도는 속성들 간의 공통집합에 비례하므로, 이점을 이용하여 식(5)를 제안하였다.

마찬가지로 l 번째 다중값 데이터 속성에 대한 객체 O_j 로부터 O_i 의 유사도는 다음과 같이 정의한다.

$$s_{j \rightarrow i}^l = \frac{|M_{il} \cap M_{jl}|}{|M_{il}|} \quad (6)$$

모든 n개의 다중값 데이터 속성에 대한 객체 O_i 로부터 O_j 의 유사도는 다음과 같이 정의한다.

$$S_{i \rightarrow j} = [s_{i \rightarrow j}^1, s_{i \rightarrow j}^2, \dots, s_{i \rightarrow j}^n] = \sqrt{\sum_{l=0}^n (s_{i \rightarrow j}^l)^2} \quad (7)$$

마찬가지로 모든 n개의 다중값 데이터 속성에 대한 객체 O_j 로부터 O_i 의 유사도는 다음과 같이 정의한다.

$$S_{j \rightarrow i} = [s_{j \rightarrow i}^1, s_{j \rightarrow i}^2, \dots, s_{j \rightarrow i}^n] = \sqrt{\sum_{l=0}^n (s_{j \rightarrow i}^l)^2} \quad (8)$$

식(7)과 (8)로부터 얻어진 유사도는 비대칭적 유사도임을 알 수 있다.

3. 계층적 클러스터링 방법

이 장에서는 평균 유사도 값(ASV; average similarity value)을 제안하며, 이것을 이용한 계층적 클러스터링 방법도 제안한다.

3.1. 평균 유사도 값 (ASV)

기존의 계층적이거나 분할 클러스터링 방법들은 대칭적 유사도 방법을 사용해야 한다. 따라서 본 논문에서 제안하는 비대칭적 유사도 측정 방법에 평균 유사도 값이라는 개념을 도입하여 대칭적 유사도를 구할 수 있다. 평균 유사도 값(ASV; Average similarity value)은 다음과 같이 정의한다.

$$ASV(O_i, O_j) = \frac{(a S_{i \rightarrow j} + \beta S_{j \rightarrow i})}{2} \quad (9)$$

여기서 a 와 β 는 스칼라 값이다.

객체들 간의 거리를 정의하는데 사용되는 척도가 유용한 척도가 되기 위해서는 메트릭 조건이나 세미메트릭 조건을 만족해야 한다. 두 객체 O_i, O_j 간의 제안하는 유사도 척도는 다음과 같은 간단한 변환식을 이용하면 비유사도로 변환할 수 있다.

$$\text{Dissimilarity}(O_i, O_j) = 1 - ASV(O_i, O_j) \quad (10)$$

이제, 두 객체 O_i, O_j 간의 비유사도가 세미메트릭 조건을 만족하지 알아보자. 객체들간의 거리 척도 d 가 다음과 같은 세가지 조건을 만족하면, 척도 d 가 세미메트릭이라고 한다.

- ① $d(Y_i, Y_j) \geq 0$
- ② $d(Y_i, Y_i) = 0$, or $d(Y_i, Y_j) = 0$ if and only if $Y_i = Y_j$
- ③ $d(Y_i, Y_j) = d(Y_j, Y_i)$

여기서, Y_i, Y_j, Y_k 는 심볼릭 객체들이다.

제안하는 유사도 척도는 $0 \leq ASV(O_i, O_j) \leq 1$ 이므로, $\text{dissimilarity}(O_i, O_j) \geq 0$ 이 되어 첫 번째 조건을 만족한다. 또한, $ASV(O_i, O_i) = 1$ 이므로, 두 번째 조건을 만족한다. 마지막으로, $ASV(O_i, O_j) = ASV(O_j, O_i)$ 이므로, 세 번째 조건도 만족을 한다. 그러므로, 제안하는 척도는 세미메트릭 조건을 만족하는 척도이다.

3.2. ASV를 이용한 계층적 클러스터링 알고리즘

클러스터링 방법은 크게 분할 (partitioning) 방법과 계층적 (hierarchical) 방법으로 나눌 수 있다. 분할 방법은 범주 함수를 최적화 시키는 K개의 분할영역을 결정해 나가는 방법으로, Euclidean distance 측정법에 기반을 둔다. 계층적 방법은 통합 (agglomerative) 방법과 분리 (divisive) 방법으로 나눌 수 있다. 통합 방법은 처음에 각각의 데이터 객체를 하나의 클러스터로 설정 한 후 이들 쌍간의 거리를 기반으로 가장 가까운 클러스터들끼리 합병을 수행한다. 최종적으로 한 클러스터에 모든 데이터 객체들이 포함될 때까지 과정을 수행한다. 분리 방법은 통합 방법과 반대로 과정을 진행한다.

본 연구에서는 클러스터링 방법으로 널리 활용되고 있고,

2장에서 제안한 유사도 측정방법을 효과적으로 이용하기 위하여 통합방식의 계층적 클러스터링 알고리즘을 사용하고자 한다. n 개의 심볼릭 객체들을 클러스터링 하는 문제를 생각해 보자. 처음에는 $n \times (n-1)/2$ 개의 클러스터간 합병을 고려할 수 있는데, 이중에서 합병을 했을 경우 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 1번째 합병 후에는 $(n-1) \times (n-1)/2$ 개의 클러스터간 합병을 고려하며, 이중에서 가장 높은 평가함수 값을 주는 두 개의 클러스터를 합병한다. 최종적으로는 주어진 개수의 클러스터가 남을 때까지 위의 과정을 반복한다.

본 연구에서는 평가함수로 식(11)을 사용한다.

Maximize Cf =

$$\sum_{r=1}^k \frac{1}{n_r} \sum_{i,j \in C_r} \text{sim}(i, j) \quad (11)$$

여기서, n_r 은 C_r 내의 심볼릭 객체들 개수,
 k 는 클러스터 개수

본 연구에서 제안하는 클러스터링 알고리즘의 단계는 그림 1과 같다. 여기서 C_i 와 S_i 는 각각 i 번째 클러스터와 심볼릭 객체이며, $|C_i|$ 는 현재 단계의 클러스터 개수이다.

Step 0. 초기화

For all $i, C_i \leftarrow O_i (\in D,$
where D is the database of
symbolic objects)

Step 1. 평가함수 값 계산

For each $C_i, C_j \in D,$
compute the ASV (Eq. (9))

Step 2. 합병

$C_{\text{new}} \leftarrow \text{merge}(C_i, C_j)$ for the max value
of ASV

Step 3. 조건 검사

If $(|C_i| > k)$ then go to Step 1.
Else go to Step 4.

Step 4. 종료

Exit the algorithm

그림 1. 제안하는 계층적 클러스터링 알고리즘
Fig. 1. Proposed hierarchical clustering algorithm

Step 0는 초기화 단계로서 데이터베이스 D 를 액세스하여 각각의 심볼릭 객체를 하나의 클러스터로 설정한다. Step 1은 두 클러스터가 합병이 될 경우의 평가함수의 값을 구하는 단계로, 현재 n 개의 클러스터가 있다고 하면, $n \times (n-1)/2$ 개의 평가함수 값을 계산한다. Step 2는 합병 단계로서, Step

1에서 계산한 평가함수 값들 중 가장 큰 값을 주는 두 개의 클러스터를 합병한다. Step 3은 조건 검사 단계로서 클러스터의 개수가 지정된 클러스터 개수보다 크면 Step 1으로 간다. 그렇지 않으면 Step 4로 간다. 마지막으로, Step 4는 종료단계로서 알고리즘을 끝낸다.

계층적 클러스터링 방법에서는 계산량이 많이 필요한 2 단계가 있다[11]. 첫 번째 단계는 데이터 셋 안에 있는 모든 객체들 간의 상호 유사도를 구하는 단계이다. n 개의 심볼릭 객체들에 대한 2장의 유사도 계산방법에 대한 시간 계산량은 $O(n^2)$ 이다. 두 번째 단계는 평가함수를 만족하는 두 개의 클러스터를 반복적으로 선택하는 단계이다. 본 연구에서 제안하는 계층적 클러스터링 방법의 시간 계산량은 $O(n^2 \log n)$ 이다. 따라서 본 연구에서 제안하는 방법론에 대한 전체적인 시간 계산량은 $O(n^2 \log n)$ 이다.

4. 실험 결과

본 장에서는 Fat oil 데이터 셋, 마이크로 컴퓨터 데이터 셋, 마이크로 프로세스 데이터 셋 등 다양한 데이터 셋[12]들에 대한 실험결과를 보여준다. 이들 데이터 셋들은 기존 연구들에서 많이 사용되었던 표준 데이터 셋들로서, 제안하는 알고리즘을 기존방법들과 비교하기 위하여 사용하였다.

데이터 셋의 패턴들 간의 유사도는 C++언어로 코딩을 하여 측정된 후, 미네소타 대학의 Karypis 연구실에 만든 CLUTO 프로그램[11]으로 클러스터링을 수행하였다.

첫 번째 실험은 Fat Oil 데이터 셋에 대해 수행하였으며, 이들 데이터 셋은 구간 데이터들을 가지고 있는 표준적인 데이터 셋으로서, 4개의 구간 데이터 속성과 8개의 패턴으로 이루어져 있다.

표 1은 식(9)를 이용하여 계산한 각 패턴들 간의 유사도 행렬이며, 2번 패턴과 5번 패턴이 가장 높은 유사도 값을 가지기 때문에 제일 먼저 같은 클러스터로 할당이 된다. 예를 들어, 2번 패턴과 5번 패턴의 ASV를 구해보자. 식(3)과 식(7)을 이용하면, $S_{2 \rightarrow 5} = 0.7186$ 이고, $S_{5 \rightarrow 2} = 0.7167$ 이다. 즉, $ASV(O_2, O_5) = (\alpha S_{2 \rightarrow 5} + \beta S_{5 \rightarrow 2})/2 = (0.7186 + 0.7167)/2 \approx 0.72$ 이다.

표 1. Fat oil 데이터 셋의 유사도
Table 1. Similarity matrix of Fat oil data

	0	1	2	3	4	5	6	7
0	1.00	0.70	0.51	0.44	0.53	0.48	0.39	0.44
1	0.70	1.00	0.63	0.66	0.53	0.67	0.49	0.54
2	0.51	0.63	1.00	0.69	0.61	0.72	0.62	0.65
3	0.44	0.66	0.69	1.00	0.56	0.66	0.40	0.46
4	0.53	0.53	0.61	0.56	1.00	0.65	0.31	0.46
5	0.48	0.67	0.72	0.66	0.65	1.00	0.50	0.56
6	0.39	0.49	0.62	0.40	0.31	0.50	1.00	0.59
7	0.44	0.54	0.65	0.46	0.46	0.56	0.59	1.00

표 2는 본 연구에서 제안하는 방법으로 클러스터링 한 결과이다. [13][14]는 패턴들을 3개의 클러스터로 그룹핑 하였고, [7]은 2개의 클러스터로 그룹핑 하였다. 제안하는 방법은 [8]의 결과와 동일한 결과를 제시한다.

표 2. Fat oil 데이터 셋의 실험결과
Table 2. Experiments of Fat oil data set

	Two cluster level	A stage just before two cluster level
Proposed method	{0,1,2,3,4,5}, {6,7}	{0,1},{2,3,4,5}, {6,7}
Gowda and Diday[13]	NA	{0,1},{2,3,4,5}, {6,7}
Gowda and Ravi[14]	NA	{0,1},{2,3,4,5}, {6,7}
Gowda and Ravi[7]	{0,1,2,3,4,5}, {6,7}	NA
Guru et al.[8]	{0,1,2,3,4,5}, {6,7}	{0,1},{2,3,4,5}, {6,7}
Kiranagi et. al.[15]	{0,1,2,3,4,5}, {6,7}	{0},{1,3}, {2,4,5},{6,7},

다음으로 마이크로 컴퓨터 데이터 셋에서 실험을 수행하였으며, 이들 데이터 셋은 3 그룹의 마이크로 컴퓨터들을 설명하는 12개의 패턴들로 구성되어 있다. 각각의 패턴들은 5개의 속성을 가지고 있다. 표 3은 각 패턴들 간의 유사도 행렬이며, 표 4는 본 연구에서 제안하는 방법으로 클러스터링 한 결과이다. 제안하는 방법은 비대칭적인 유사도 개념을 사용하였기 때문에 기존의 방법들과는 서로 다른 실험결과를 보여준다.

표 3. 마이크로 컴퓨터 데이터 셋의 유사도
Table 3. Similarity matrix of Micro computer data set

	0	1	2	3	4	5	6	7	8	9	10	11
0	1.00	0.50	0.30	0.13	0.13	0.00	0.00	0.00	0.35	0.35	0.25	0.25
1	0.50	1.00	0.25	0.75	0.00	0.00	0.00	0.50	0.35	0.35	0.25	0.25
2	0.30	0.25	1.00	0.00	0.50	0.13	0.00	0.25	0.00	0.00	0.00	0.50
3	0.13	0.75	0.00	1.00	0.00	0.25	0.00	0.75	0.25	0.35	0.43	0.35
4	0.13	0.00	0.50	0.00	1.00	0.43	0.25	0.25	0.00	0.00	0.00	0.75
5	0.00	0.00	0.50	0.25	0.43	1.00	0.25	0.39	0.00	0.00	0.25	0.85
6	0.00	0.00	0.13	0.00	0.25	0.25	1.00	0.00	0.25	0.00	0.00	0.25
7	0.00	0.50	0.00	0.75	0.25	0.39	0.00	1.00	0.00	0.85	0.25	0.25
8	0.35	0.35	0.25	0.25	0.00	0.00	0.25	0.00	1.00	0.85	0.75	0.00
9	0.35	0.35	0.00	0.35	0.00	0.00	0.00	0.85	0.85	1.00	0.85	0.25
10	0.25	0.25	0.00	0.43	0.00	0.25	0.00	0.25	0.75	0.85	1.00	0.35
11	0.25	0.25	0.50	0.35	0.75	0.85	0.25	0.25	0.00	0.25	0.25	1.00

표 4. 마이크로 컴퓨터 데이터 셋의 실험결과
Table 4. Experiments of Micro computer data set

	Two cluster level	A stage just before two cluster level
Proposed	{0,1,3,7,8,9},	{0,1,3,7,8,9},

method	10}, {2,4,5,6,11}	10}, {2,4,5,11}, {6}
Gowda and Diday[13]	NA	{0,1,3,9,10}, {6},{2,8}, {4,5,11},{7}
Gowda and Ravi[14]	NA	{0,1,3,5,7,8,9, 10,11},{2},{6}, {4}
Gowda and Ravi[7]	{0,1,2,3,4,5, 7,8,9,10,11}, {6}	NA
Guru et al.[8]	{0,1,2,3,4,5, 7,8,9,10,11}, {6}	{0,1,2,3,8,9, 10}, {4,5,7}, {6}, {11}
Kiranagi et. al.[15]	{0,1,2,3,4,5, 7,8,9,10,11}, {6}	{4,5,7,11}, {0,1,2,3,8,9, 10},{6}

마지막 실험은 마이크로 프로세서 데이터 셋에서 수행되었으며, 이 데이터 셋은 6개의 속성과 9개의 패턴들로 구성되어 있다. 표 5는 각 패턴들 간의 유사도 행렬이며, 표 6은 본 연구에서 제안하는 방법으로 클러스터링 한 결과이다. 마이크로 컴퓨터 데이터 셋의 실험결과와 마찬가지로 제안하는 방법의 결과가 기존 방법들과는 다른 결과를 제시한다.

표 5. 마이크로 프로세서 데이터 셋의 유사도
Table 5. Similarity matrix of Micro processor data set

	0	1	2	3	4	5	6	7	8
0	1.00	0.54	0.45	0.51	0.70	0.63	0.51	0.32	0.35
1	0.54	1.00	0.25	0.25	0.61	0.40	0.25	0.00	0.00
2	0.45	0.25	1.00	0.44	0.44	0.68	0.60	0.00	0.19
3	0.51	0.25	0.44	1.00	0.51	0.51	0.35	0.30	0.26
4	0.70	0.61	0.44	0.51	1.00	0.58	0.51	0.31	0.32
5	0.63	0.40	0.68	0.51	0.58	1.00	0.61	0.29	0.68
6	0.51	0.25	0.60	0.35	0.51	0.61	1.00	0.30	0.51
7	0.32	0.00	0.00	0.30	0.31	0.29	0.30	1.00	0.31
8	0.35	0.00	0.19	0.26	0.32	0.68	0.51	0.31	1.00

표 6. 마이크로 프로세서 데이터 셋의 실험결과
Table 6. Experiments of Micro processor data set

	Two cluster level	A stage just before two cluster level
Proposed method	{0,1,2,3,4,5,8}, {7}	{0,1,2,3,4,5,6} {7}, {8}
Gowda and Diday[13]	NA	NA
Gowda and Ravi[14]	NA	{0,1,4,5,7}, {2,3,6}, {7,8}
Gowda and Ravi[7]	{0,1,2,3,4,5,6, 8}, {7}	NA

Guru et al.[8]	{0,1,4,7}, {2,3,5,6,8}	{0,1,4,7},{2,3, 5,6},{8}
Kiranagi et. al.[15]	{0,1,2,3,4,5,6, 7},{8}	{0,1,4,7},{2,3, 5,6}, {8}

표 2, 4, 6에서 NA로 표시된 것은 기존 연구들에서 해당하는 결과가 없는 경우이다. 본 연구에서 제시하는 계층적 클러스터링 알고리즘은 기존의 방법들과 다르게 사전 지식이 필요 없다. 또한 제안하는 방법의 클러스터링 결과는 기존 방법들과 다르며, 기존 방법들과의 비교 실험을 통하여 타당성을 입증하였다.

5. 결 론

본 논문에서는 심볼 객체들을 클러스터링 하는 문제에 대해 다루었다. 심볼 객체들을 클러스터링 하기 위해서 심볼 객체들간의 유사도를 측정하기 위한 새로운 방법을 제안하였다. 또한 ASV라는 값과 이를 이용한 계층적 클러스터링 방법을 제안하였다. 표준적인 데이터 셋을 이용하여 우리가 제안하는 클러스터링 방법이 심볼 객체들을 클러스터링 하는데 효율적이라는 것을 보여주었다.

향후에는 본 논문에서 사용된 심볼 객체들을 표현 하는 방법을 좀 더 확장하고, 사람들간의 친밀도나 혈액 공급 등 비대칭적 유사도가 실제로 적용될 수 있는 사회과학적 분야의 데이터 셋들에 본 연구에서 제안한 방법들을 적용하고 싶다.

References

- [1] J. Han and M. Kamber, *Data mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.
- [2] S. Asharaf, M. N. Murty, S. K. Shevade, "Rough set based incremental clustering of interval data," *Pattern Recognition Letters*, vol. 27, pp. 515-519, 2006.
- [3] T. V. Ravi and K. C. Gowda, "A new non-hierarchical clustering procedure for symbolic objects," *LNCS 1983*, pp. 35-41, 2000.
- [4] T. Denoeux and M. Masson, "Multidimensional scaling of interval valued dissimilarity data," *Pattern Recognition Letters*, vol. 21, pp. 83-92, 2000.
- [5] K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, vol. 24, no. 6, pp. 567-578, 1991.
- [6] K. C. Gowda and E. Diday, "Symbolic clustering using a new similarity measure," *IEEE Trans. SMC*, vol. 22, no. 2, pp. 368-378, 1992.
- [7] K. C. Gowda and T. V. Ravi, "Disjunctive clustering of symbolic objects using the concepts of both similarity and dissimilarity," *Pattern Recognition*, vol. 28, pp. 1277-1282, 1995(b).

- [8] D. S. Guru, B. B. Kiranagi and P. Nagabhushan, "Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns," *Pattern Recognition Letters*, vol. 25, pp. 1203-1213, 2004.
- [9] D. S. Guru and B. B. Kiranagi, "Multivalued type dissimilarity measure and the concept of mutual similarity value useful for clustering symbolic patterns," *Pattern Recognition Letters*, vol. 25, no. 10, pp. 1203-1213, 2005.
- [10] S. H. N. Prakash, *Classification of remotely sensed data: some new approaches*, Ph. D. Thesis, University of Mysore, Mysore, India, 1998.
- [11] Y. Zho, G. Karypis, *Comparison of agglomerative and partitional document clustering algorithms*, Proc. 2nd SIAM Int. Conf. Data Mining, VA, 2002.
- [12] M. Ichino, H. Yaguchi, "Generalized minkowski metrics for mixed feature-type data analysis," *IEEE Trans. On systems, man, and cybernetics*, vol. 24, 1994.
- [13] K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, vol. 24, pp. 567-578, 1991.
- [14] K. C. Gowda and T. V. Ravi, "Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity," *Pattern Recognition Letters*, vol. 16, pp. 647-652, 1995(a).
- [15] B. B. Kiranagi, D. S. Guru, S. Manjunath and B. S. Harish, "MNV for clustering based on non symmetric symbolic proximity," *Int. J. of Recent Trends in Engineering*, vol. 2, pp. 121-124, 2009.

저 자 소 개

오승준(Seung-Joon Oh)

1992년 : 한양대학교 산업공학과 공학사
1994년 : 한양대학교 산업공학과 공학석사
2004년 : 한양대학교 산업공학과 공학박사
2005년~현재 : 경기과학기술대학교
산업경영과 교수

관심분야 : 데이터마이닝, 클러스터링
Phone : +82-31-496-4719
E-mail : osj@gtec.ac.kr

박찬웅(Chan-Woong Park)

1988년 : 한양대학교 산업공학과 공학사
1991년 : 한양대학교 산업공학과 공학석사
1997년 : 한양대학교 산업공학과 공학박사
2007년~현재 : 가천대학교
산업경영공학과 교수

관심분야 : CIM, 데이터베이스 응용
Phone : +82-31-750-5805
E-mail : cwpark@gachon.ac.kr