범주형 값들이 순서를 가지고 있는 데이터들의 클러스터링 기법
Clustering Algorithm for Sequences of Categorical Values

오승준*
Oh, Seung Joon
김재련**
Kim, Jae Yearn

We study clustering algorithm for sequences of categorical values. Clustering is a data mining problem that has received significant attention by the database community. Traditional clustering algorithms deal with numerical or categorical data points. However, there exist many important databases that store categorical data sequences. In this paper we introduce new similarity measure and develope a hierarchical clustering algorithm. An experimental section shows performance of the proposed approach.

1. 서론

클러스터링(Clustering)이란 물리적 혹은 추상적 객체들을 서로 비슷한 객체들의 집합으로 그룹화 하는 과정으로, 하나의 클러스터에 속하는 데이터 객체들 간에는 서로 다른 클러스터 내의 객체들과는 구분되는 유사성을 갖게 된다 [2].

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 (agglomeratvie) 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.

* 한양대학교 산업공학과 박사과정
**한양대학교 산업공학과 교수

클러스터링 기법들은 통계학(statistics), 패턴인식(pattern recognition) 등의 분야에서 연구되어 왔으며, 현재는 데이터 마이닝 분야에서 이 기법을 응용하려는 연구가 활발히 진행되고 있다. 기존의 클러스터링 기법들은 주로 수치형 데이터 [5][7][8][11]와 범주형 데이터[9][12]들을 문제영역으로 다루어 왔다. 그러나, 실제로는 범주형 값들이 순서를 가지고 있는 데이터들이 존재하며, 기존의 기법들은 이러한 데이터들을 고려하지 않았다.

범주형 값들이 순서를 가지고 있는 데이터들에 대한 클러스터링은 행동에 의한 세분화(behavioral segmentation) 분야에 많은 응용 문제를 가지고 있다[6]. 예를 들면 웹 사용자를 세분화하는 문제에서, 웹 로그 파일을 이용하여 웹 사용자들을 클러스터링 하는 문제이다.

본 논문에서는 범주형 값들이 순서를 가지고 있는 데이터들을 클러스터링 하기 위해서 새로운 유사도 척도를 제안하고, 이 척도를 이용하여 계층적 방법으로 클러스터링을 수행한다.

2. 기존 연구

다양한 클러스터링 기법들에 대한 연구는 [2][3]에 있고, 이중에서 범주형 속성들에 대한 연구는 [9][12]에 있다. [9][12]는 단지 데이터들이 범주형 속성들로 이루어진 경우의 클러스터링 문제만을 다루고 있다.

범주형 값의 시퀀스에 대한 연구는 주로 빈발하는 순차 패턴을 찾는데 집중되어 왔다. 이 문제는 [1]에서 처음으로 제안되었는데, 이 분야의 순차패턴을 탐사하는 문제는 시퀀스의 지지도가 사용자가 정의한 최소지지도보다 같거나 큰 시퀀스를 발견하는 것이다.

복잡한 객체들의 시퀀스를 클러스터링 하는 문제는 [4]에서 제안되었다. 이 논문은 클래스 계층을 사용하여 클러스터링을 수행하는데, 수치형 데이터로 표현되는 객체들(예를 들면 움직이는 객체들의 궤적)의 시퀀스를 클러스터링 하는데 적합하다.

[10]에서 범주형 값들이 순서를 가지고 있는 데이터들의 클러스터링을 제안하였다. 이 논문은 빈발 패턴이 주어져 있다고 가정을 하고, 이 빈발 패턴을 하나 이상 포함한 데이터만을 대상으로 클러스터링을 수행한다. 그러나, 본 연구에서는 빈발 패턴에 상관없이 본 논문에서 제안하는 새로운 유사도 척도에 따라 모든 데이터를 대상으로 클러스터링을 진행한다.

3. 범주형 값들이 순서를 가지고 있는 데이터들의 클러스터링

이 장에서는 본 연구에서 사용된 새로운 용어를 정의하고 예를 들어 설명한다.

정의 1. $I = \{ i_1 , i_2 , \cdots, i_j , \cdots, i_m \}$은 항목 i_j들의 집합이다.

정의 2. 시퀀스(sequence) S는 n 개의 항목들의 집합이고 $< x_1\ x_2 \cdots x_j \cdots\ x_n >$ 으로 표시하고 여기서 x_j 는 항목이다. 데이터베이스 D는 시퀀스들의 집합이다.

정의 3. 시퀀스의 크기는 그 시퀀스에 존재하는 항목들의 개수이며, 크기가 k인 시퀀스를 k-시퀀스라고 한다.

정의 4. 시퀀스 $S_1 = < a_1\ a_2 \cdots\ a_n >$과 시퀀스 $S_2 = < b_1\ b_2 \cdots\ b_m >$가 존재할 때, $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, …, $a_n \subseteq b_{in}$ 을 만족하는 정수 $i_1 < i_2 \cdots < i_n$ 이 존재하면 시퀀스 S_2이 시퀀스 S_1를 지지한다고 정의한다. 또는 S_1가 S_2의 서브시퀀스라고 하며, $S_1 \subset S_2$로 표현한다.

정의 5. 시퀀스 $S = < x_1\ x_2 \cdots\ x_i\ x_j \cdots\ x_n >$에서 2 개의 항목집합으로 구성된 $x_i\ x_j$ $(i < j)$를 시퀀스 요소 e_i 라고 하며, 시퀀스 S에는 $\sum_{k=1}^{n-1} k$ 개의 시퀀스 요소가 존재한다. $E = \{ e_1, e_2, \cdots, e_i, \cdots, e_{n-1} \}$는 시퀀스 요소 e_i 들의 집합이다.

정의 6. 두 개의 시퀀스 $S_1 = < a_1\ a_2 \cdots\ a_n >$과 시퀀스 $S_2 = < b_1\ b_2 \cdots\ b_m >$ 의 시퀀스 요소 집합을 각각 $E_1 = \{ ea_1, ea_2, \cdots, ea_i, \cdots, ea_{n-1} \}$, $E_2 = \{ eb_1, eb_2, \cdots, eb_j, \cdots, eb_{m\,1} \}$라고 하면, S_1, S_2 의 유사도(similarity) $sim(S_1, S_2)$는 다음과 같이 정의한다.

$$sim(S_1, S_2) = \sum_{ea_i \in E_1, eb_j \in E_2} \delta(ea_i, eb_j) \qquad (1)$$

$$\text{여기서, } \delta(ea_i, eb_j) \begin{cases} = 1 & \text{if } ea_i = eb_j \\ = 0 & \text{otherwise} \end{cases}$$

예) 두 개의 시퀀스 $S_1 = < ABDE >$, $S_2 = < ACDEG >$가 있다. S_1 과 S_2 는 각각 4-시퀀스와 5-시퀀스이며, 각각의 시퀀스 요소 집합은 각각 $E_1 = \{ AB, AD, AE, BD, BE, DE \}$ 과 $E_2 = \{ AC, AD, AE, AG, CD, CE, CG, DE, DG, EG \}$ 이다. 두 시퀀스의 유사도 $sim(S_1, S_2)$는 3이다.

정의 7. 클러스터 $C_1 = \{ S_1, S_2, \cdots, S_i, \cdots, S_n \}$과 클러스터 $C_2 = \{ S_1, S_2, \cdots, S_j, \cdots, S_m \}$의 유사도 $sim(C_1, C_2)$는 다음과 같이 정의한다.

$$sim(C_1, C_2) = \frac{1}{|C_1|\ |C_2|} \sum_{S_i \in C_1, S_j \in C_2} sim(S_i, S_j) \qquad (2)$$

$|C_1|$, $|C_2|$ 는 각각 클러스터 C_1, C_2 에 있는 시퀀스의 총 개수

예) 네 개의 시퀀스 $S_1 = < ABDE >$, $S_2 = < ACDEG >$, $S_3 = < ABG >$, $S_4 = < AEG >$ 로 이루어진 두 개의 클러스터 $C_1 = \{ S_1, S_2 \}$, $C_2 = \{ S_3, S_4 \}$ 가 있다. 두 클

러스터 C_1, C_2 의 유사도 $sim(C_1, C_2)$는 다음과 같다.

$$sim(C_1, C_2) = \frac{1}{2 \cdot 2}\{sim(S1, S3) + sim(S1, S4) + sim(S2, S3) + sim(S2, S4)\}$$

$$= \frac{1}{2 \cdot 2}\{1+1+1+3\} = 1.5$$

문제정의

시퀀스들의 집합인 데이터베이스 $D = \{ S_1, S_2, \cdots, S_n \}$에서, 지정된 클러스터 개수나 최소 유사도 이하의 클러스터가 없을 때까지 유사도가 높은 시퀀스들부터 병합하여 클러스터링 수행하는 것이다.

4. 알고리듬

본 장에서는 범주형 값들이 순서를 가지고 있는 데이터들의 클러스터링에 대한 알고리듬을 제안한다. 본 연구에서 제안하는 알고리듬의 개요는 그림 1과 같다.

Step 0. 초기화
　for all i
　　$C_i \leftarrow S_i \in D$

Step 1. 유사도 계산
　for each C_i, $C_j \in D$
　　Compute $sim(C_i, C_j)$

Step 2. 병합
　$C_{new} \leftarrow$ Merge (C_i, C_j) for $max\_sim(C_i, C_j)$

Step 3. 유사도 갱신
　Update $sim(C_{new}, C_i)$

Step 4. 조건 검사
　If ($|C_i| > n$) And
　　(exist C_i, $C_j \in D$ such that $sim(C_i, C_j) > min\_sim$)
　Then Goto Step 2.
　Else Goto Step 5.

Step 5. 종료
끝낸다.

그림 1. 클러스터링 알고리듬

Step 0은 초기화 단계로서 데이터베이스 D를 액세스하여 각각의 시퀀스를 하나의 클러스터로 설정한다. Step 1은 유사도 계산 단계로서 각 클러스터간의 유사도를 계산한다. Step 2는 클러스터 합병 단계로서 유사도가 가장 높은 두 개의 클러스터를 합병한다. Step 3은 유사도 갱신 단계로서 Step 2에서 합병된 클러스터와 나머지 클러스터간의 유사도를 갱신한다. Step 4는 조건 검사 단계로서 클러스터의 개수가 지정된 클러스터 개수보다 크고 두 클러스터간의 유사도중 최소 유사도 이상의 것이 존재하면 Step 2.로 간다 그렇지 않으면 Step 5로 간다. 마지막으로, Step 5는 종료단계로서 알고리듬을 끝낸다.

본 알고리듬에서 가장 많은 계산량을 필요로 하는 부분은 유사도 계산 및 갱신 단계이다. 따라서, 이들 유사도 계산을 효율적으로 수행하기 위하여 그림 2.와 같은 유사도 계산 알고리듬을 제안한다. 그림 2.는 두 개의 시퀀스 $S_1 = < a_1\ a_2 \cdots\ a_n >$과 $S_2 = < b_1\ b_2 \cdots\ b_m >$의 유사도를 계산하는 알고리듬이다.

```
Step 0. 초기화
    S1, S2, Sa = { }, Sb = { }, count=0
Step 1. Sa 생성
    for i=1 to n {
      for j=1 to m {
        if ai = bj
          insert(Sa, ai)
      }
    }

Step 2. Sb 생성
    for j=1 to m {
      for i=1 to n {
        if bj = ai
          insert(Sb, bj)
      }
    }
```

Step 3. S_a, S_b 간의 유사도 계산
```
for i=1 to n'
  for j=1 to m' {
    if ea_i = eb_j
      count=count+1
  }
}
```
Step 4. 종료

$sim(S_1, S_2) = count$

그림 2. 유사도를 계산하는 알고리듬 : $sim(S_1, S_2)$

웹 사용자 세분화 문제를 예로 들어 설명하겠다. 그림 3. 과 같은 웹 로그 파일이 주어져 있다. (예: 사용자 u1은 사이트 A를 방문한 후, 차례대로 B, C, F, Z를 방문함) 웹 사용자들을 웹 로그 파일을 기초로 몇 개의 그룹으로 클러스터링 하는 문제이다. 즉, 웹 사용자의 사이트 방문 기록을 다른 사용자의 방문 기록과 비교하여 유사한대로 클러스터링을 하는 문제이다.

그림 1.의 클러스터링 알고리듬에서 보면, Step 0.에서 각 시퀀스들을 하나의 클러스터로 할당하여 모두 6개의 클러스터가 생성된다. Step 1.에서 각 클러스터들간의 유사도를 계산하고, Step 2.에서 유사도가 가장 높은 u1과 u2를 합병한다. 그러면 그림 4. 와 같이 모두 5개의 클러스터가 생성된다. 이후, Step 3.에서 유사도를 갱신한 후, 현재의 클러스터 개수가 지정된 클러스터 개수 보다 크고 최소 유사도 이하의 클러스터가 없을 때까지 앞의 과정을 반복 수행한다.

u1: A→B→C→F→Z　　u4: B→G→I

u2: A→C→F→H　　u5: C→B→A→F

u3: B→E→G→I　　u6: Z→C→B→A

그림 3. 웹 사용자들의 사이트 방문기록

c1	c2	c3	c4	c5
u1: A→B→C→F→Z u2: A→C→F→H	u3: B→E→G→I	u4: B→G→I	u5: C→B→A→F	u6: Z→C→B→A

그림 4. 첫 번째 합병후의 클러스터

그림 1.의 클러스터링 알고리듬 단계 중 Step 2.의 유사도 계산 부분을 좀 더 자세히 살펴보자. 예를 들어, u_1 = <ABCFZ>, u_2 = <ACFH>의 유사도 계산 과정을 주어

진 시퀀스로부터 직접 계산하면 u_1의 시퀀스 요소 {AB, AC, AF, AZ, BC, BF, BZ, CF, CZ, FZ}를 S_2 의 시퀀스 요소 {AC, AF, AH, CF, CH, FH}와 비교하여 유사도를 계산한다. (정의 6 참조) 그러나, 그림 2의 알고리듬을 사용하면, Step 1.에서 S_a = <ACF>를 구하고, Step 2.에서 S_b = <ACF> 구하여, 이들 S_a, S_b 로부터 유사도를 계산하기 때문에 훨씬 효율적으로 유사도를 계산하게 된다.

5. 실험결과

본 연구에서 제안한 클러스터링 알고리듬의 수행도를 평가하기 위해, 본 연구의 알고리듬 중 가장 많은 계산량을 필요로 하는 유사도 계산 알고리듬의 성능을 실험하였다. 그림 5에서 방법 1은 주어진 데이터로부터 직접 유사도를 계산하는 방법이고, 방법 2.는 본 연구에서 제안하는 유사도 계산 알고리듬이다.

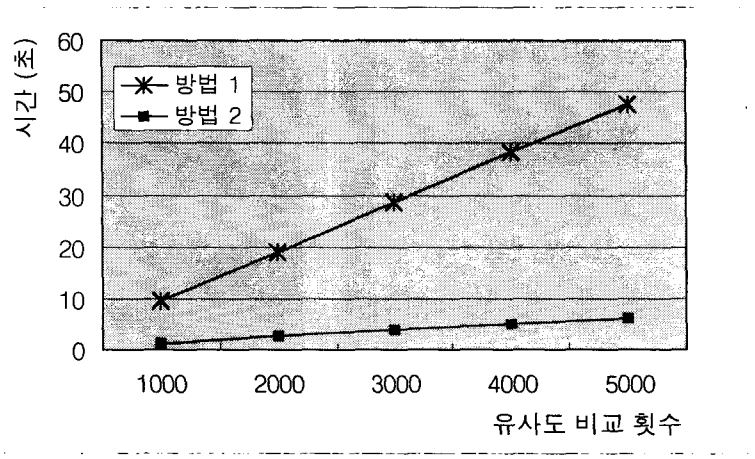

그림 5. 실험결과

본 실험은 인텔 550MHz 사양의 586 컴퓨터에서 C++ 언어로 코딩하여 수행하였고, 시퀀스의 크기가 12~15인 데이터를 대상으로 하였다. 그림 5.에서 보는 바와 같이 유사도 비교 횟수가 늘어나도 본 연구에서 제안하는 유사도 계산 알고리듬의 수행시간이 방법 1보다 훨씬 우수함을 알 수 있다.

6. 결론

본 논문에서는 범주형 값들이 순서를 가지고 있는 데이터들의 클러스터링 문제를 연구하였다. 본 문제를 풀기 위하여 새로운 유사도 척도를 제안하였고, 이 척도를 이용하여 계층적 방법으로 클러스터링을 수행하는 알고리듬을 제안하였다. 또한 효율적으로 유사도를 계산하는 알고리듬을 제안하였고, 실험을 통하여 성능의 우수함을 보여주었다.

본 논문의 클러스터링 알고리듬은 기존의 방법과 달리 빈발 패턴이 주어져 있지 않아도 되며, 모든 입력 데이터들을 대상으로 지정된 클러스터 개수나 최소 유사도 이하의 클러스터가 없을 때까지 클러스터링을 수행한다.

참고문헌

[1] Agrawal R., Srikant R.; "Mining Sequential Patterns", Proceedings of the 11th International Conference on Data Engineering, 1995

[2] J. Han and M. Kamber; Data Mining: Concepts and Techniques, Morgan kaufmann Publishers, pp335-393, 2001

[3] J. Han, M. Kamber, and A.K.H. Tung; "Spatial Clustering Methods in Data Mining: A Survey", H. J. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, NY: Taylor and Francis, 2001.

[4] Ketterlin A.; "Clustering Sequences of Complex Objects", Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 1997

[5] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu; "A Dennsity-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", KDD, 1996.

[6] Perkowitz M., Etzioni O.; "Towards Adaptive Web Sites: Conceptual Framework and Case Study", Computer Networks 31, Proceedings of the 8th International WWW Conference, 1999

[7] Raymond T. Ng, Jiawei Han; "Efficient and Effective Clustering Method for Spatial Data Mining", VLDB 1994.

[8] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim; "CURE : An Efficient Clustering Algorithm for Large Databases", SIGMOD98, 1998

[9] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim; "ROCK: A Robust Clustering Algorithm for Categorical Attributes", IEEE99, 1999

[10] Tadeusz Morzy, Marek Wojciechowski, Maciej Zakrzewicz: "Scalable Hierarchical Clustering Method for Sequences of Categorical Values", Proc. of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'01), Kowloon, Hong Kong, 2001

[11] Tian Zhang, Raghu Ramakrishnan, Miron Livny; "BIRCH : An Efficient Data Clustering Method for Very Large Databases", ACM SIGMOD96, 1996

[12] Wang K., Xu C., Liu B.; "Clustering Transactions Using Large Items", Proceedings of the '99 ACM, 1999