고차원 범주형 데이터를 위한 투영 군집화 기법의 핵심 요소 개발

김민호[O,1], R.S. Ramakrishna[2]
[1]한국전자통신연구원, 바이오인포매틱스팀
kimmh@etri.re.kr
[2]광주과학기술원 정보통신공학과
rsr@gist.ac.kr

Development of Core Components of Projected Clustering for High-Dimensional Categorical Data

Minho Kim[O,1] and R.S. Ramakrishna[2]
[1]Bioinformatics Team, Electronics and Telecommunications Research Institute (ETRI)
[2]Dept. of Information & Communications, Gwangju Institute of Science and Technology (GIST)

요 약

본 논문은 고차원의 범주형 데이터에 대한 군집화에 대해서 다룬다. 기존의 범주형 데이터 객체를 위한 유사성(상이성) 계측들의 기저에 깔려 있는 한계점은 수치형 데이터에서와 같은 순서화 (ordering)의 부재와 데이터의 고차원성과 희소성에 기인하는데, 이를 효과적으로 극복할 수 있는 기법이 투영 군집화이다. 본 논문에서는 고차원의 범주형 데이터를 효과적으로 처리할 수 있는 투영 군집화를 다루며 핵심 요소인 군집 차원의 정의와 군집 응집도를 제안한다.

1. 서 론

군집화에 있어서 범주형 데이터에 대한 가장 주요한 특징은 각 값들 사이의 순서가 정의되지 않기 때문에 수치형 데이터에서와 같은 두 객체의 값의 차에 기반한 자연스러운 거리 함수가 정의 될 수 없다는 것이다 [4]. 그 대신 Jaccard/Dice 계수(coefficient), Hamming 거리 함수, cosine 계측과 같이 두 객체의 교집합(과 합집합)을 이용하는 유사성(상이성) 계측이 사용될 수 있다. 하지만 이와 같은 계측들은 몇 가지 문제가 있다. 가장 먼저 들 수 있는 문제는 계측 레벨의 제한이다. 예를 들어, Jaccard/Dice 계수의 경우, 교집합에 기반하고 있기 때문에 기껏해야 ' $\min\{|T_1|, |T_2|\} + 1$' 만큼의 계측 레벨밖에 가질 수 없다 [5]. 이 때, T_1과 T_2는 항목 집합인 트랜잭션을 의미한다. 이것은 부가적인 다른 문제를 야기시킨다. 이른바, 한 쌍의 객체들에 대한 유사성이 다른 쌍의 객체들의 유사성에 비해 분명한 차이가 존재함에도 불구하고 동일한 크기의 교집합을 가질 때 그 차이를 구분할 수 없게 된다.

이러한 문제는 시장 바구니 (market basket)와 같은 트랜잭션 데이터집합을 다룰 때 훨씬 심각해진다 [7] [5] [8]. 이 트랜잭션 데이터집합의 특성은 항목(item) 집합의 크기는 매우 큰 반면에 실제 각 트랜잭션이 포함하는 항목의 수는 상대적으로 적다. 즉, 고차원성(high dimensionality)과 희소성(sparsity)의 특징을 가지고 있다. 이와 같은 문제점을 해결하기 위해 여러 가지 다양한 연구가 시도 되었다. ROCK [5]에서는 *link* 라는 객체 사이의 내부 연결도 (interconnectivity)를 두 객체 사이의 유사성 계측으로 이용하였다. 쌍 유사성(상이성) 계측과 같은 국지적인 (local) 계측과는 달리 전역적인 (global) 유사성 (또는 상이성) 계측 방법을 사용한 군집화 알고리즘들도 있다 [7] [8]. LargeItem [7]에서는 군집 내의 *large* 항목 (item)의 수를 최대화 시키는 반면, 군집들 사이의 중복된 large 항목의 수를 최소화시키는 방법으로 군집화를 시도하였고, CLOPE [8]에서는 높이(항목의 발생 빈도수)-넓이(항목의 총 수) 비를 증가시킴으로써 최적의 군집 구조를 찾고자 시도하였다.

앞서 언급했던 연구들은 범주형 데이터 객체 사이의 유사성 계측이 수치형 데이터의 유사성 계측에서 기대했던 것과는 다르기 때문에 발생한 문제를 해결하기 위한 시도였다. 그런데, 수치형 데이터에 대해서도 앞에서 언급했던 것과 유사한 문제가 있었다. 수치형 데이터집합에서, 차원이 높아짐에 따라 모든 쌍의 데이터 객체들에 대한 거리가 거의 같은 값을 가지게 되는 차원의 저주 (the curse of dimensionality)가 바로 그것이다 [2].

수치형 데이터집합에서 차원의 저주 문제에 대한 해결책 중에서 주목을 받았던 방법 중의 하나가 투영 군집화 (projected clustering)이다. 투영 군집화에 이용된 현상은 고차원의 수치형 데이터집합에서 각 군집은 각기 다른 차원의 부분 집합에 대해 연관되어 있으며, 자신의 연관 차원 이외의 차원들은 군집화에서 방해가 될 뿐이라는 점이다. 주목할 만한 점은 이러한 현상을 범주형 데이터집합에 대한 군집화에서도 쉽게 확인할 수 있다는 것이다 [5] [8].

이와 같은 유사성에 대한 직관의 발견이 본 논문의 연구 동기가 되었으며, 본 논문에서는 고차원의 수치형 데이터 집합에서 사용되었던 투영 군집화 기법에 기반하여 고차원의 범주형 데이터집합을 다룰 수 있는 군집화 기법을 위한 주요 요소들을 제안한다. 첫 번째 핵심 요소는 군집 차원으로써 범주형 데이터에 적합한 군집 차원을 재정의한다. 다음으로, 최종 타겟 군집화 알고리즘은 계층적 군집화를 기반할 예정인데, 계층적 군집화로 인해 발생하는 오류군집을 효과적으로 판별하기 위해 정보 이론에 기반을 둔 응집도 계측인 InCoh'(C_k)에 의해 정의한다.

본 논문의 구성은 다음과 같다. 2 절에서는 투영 군집화를 간단히 소개한다. 3 절과 4 절에서는 범주형 데이터를 위한 군집 차원과 군집 응집도를 정의하고 설명한다. 마지막으로 5 절에서 결론을 제시한다.

2. 투영 군집화

고차원의 데이터 집합에서 모든 차원이 임의의 한 군집과 연관되어 있을 가능성은 매우 적다. 이것은 연관된 차원 이외의 차원들은 군집화에 방해될 수도 있음을 의미한다.

이러한 문제점에 대한 해결책은 서로 다른 그룹들은 차원에 대해 서로 다른 상관성을 가진다는 사실에서 찾을 수 있다 [1]. 이를 이용하면 다음과 같은 투영된 군집(Projected Cluster)을 정의할 수 있으며, 고차원의 데이터 집합에 대한 군집화 문제는 이 정의를 바탕으로 해결할 수 있다. *투영된 군집*은 전체 차원의 한 부분 집합에 대해 전체 데이터 집합의 일부 데이터 객체들이 밀접하게 연관되어 있을 때, 차원의 부분 집합 D와 함께 정의된 데이터 객체의 부분 집합 C를 의미한다. 여기에서, D의 각 차원을 군집차원이라 한다.

이것을 실제 군집화에 응용한 알고리즘이 PROCLUS(PROjected CLUStering) [1] 이다. 여기서는 PROCLUS 알고리즘의 기저 아이디어만을 제시하였으며, 좀 더 자세한 설명은 [1]에서 구할 수 있다.

3. 군집 차원

투영 군집화에서 가장 중요한 요소중의 하나는 당연히 군집 C_i 에 대한 차원 집합 D_i 를 어떻게 정의하느냐 하는 것이다. 범주형 데이터집합에서 차원 집합을 어떻게 정의할지 알아보기 이전에 먼저 그림 1의 예제를 살펴보자.

그림 1에서 알 수 있듯이, 모든 데이터 객체는 동일한 수의 애트리뷰트를 가진 트랜잭션이다. 그림 1 (a) 의 각 셀에 있는 숫자는 각 군집에 대한 각 애트리뷰트값(a_{xy}: 애트리뷰트 a_x 의 y 번째 값) 의 발생 빈도수를 의미한다. 그림 1 (b) 는 각 애트리뷰트의 지배 패턴을 보여준다. 즉, 애트리뷰트 a_1의 a_{13}가 다른 애트리뷰트값들에 비해 가장 현저한 발생 빈도를 가지기 때문에 ' 1' 의 값을 가지는 지배 패턴으로 설정되고 다른 애트리뷰트값들(a_{11}, a_{12}, a_{14})은 ' 0' 의 값을 가지는 비지배패턴으로 설정된다. 이렇게 정의한 중심을 지배패턴 기반 중심 (dominant pattern-based center, DPBC)라 하자.

각 애트리뷰트에 속한 지배적 패턴의 수에 따라, 애트리뷰트를 3가지로 분류할 수 있다: SDA (single dominant attribute), PDA(partial dominant attribute), CDA(complete dominant attribute). 여기에서, SDA는 애트리뷰트가 단 하나의 지배 애트리뷰트값을 가지며, PDA는 (애트리뷰트 a_i 의 카디널리티(cardinality)보다 작은) 몇 개의 지배패턴을 가진다. 그리고 CDA 는 모든 애트리뷰트값 이 지배패턴인 경우이다 (이것은 구현상으로 정의한 것에 불과하며 의미적으로는 지배패턴이 전혀 없는 경우와 동일한 경우라 할 수도 있다). 참고로, 각 애트리뷰트의 지배패턴을 찾기 위해 [6] 에서 제안된 군집 차원 자동결정 알고리즘을 이용할 수 있다.

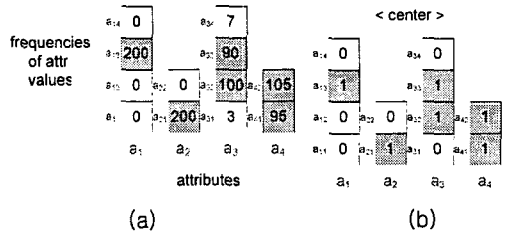

그림 1. 지배패턴 기반 중심 (Dominant pattern-based center, DPBC)

본론으로 다시 돌아와서, 임의의 군집에 속한 데이터 객체들이 임의의 차원에서 다른 차원들에 비해 훨씬 더 연관되어 있는 경우 그 차원을 *군집 차원*이라 정의한다. 그리고, 이 군집 차원들이 모여서, 군집 차원 집합을 이룬다 [1]. 범주형 데이터에 대해서도 이 개념을 그대로 적용시켜 보자. 그러면, 자연스럽게 SDA가 이에 부합함을 알 수 있을 것이다. 따라서, 본 논문에서는 SDA를 군집의 범주형 데이터에 대한 군집 차원으로 정의한다. 그림 1의 예제에서는 a_1과 a_2가 SDA로써 군집 차원으로 정의된다. 그리고, a_3과 a_4는 각각 PDA와 CDA로써 비군집 차원으로 정의된다.

4. 군집 응집도

본 논문에서 최종 목표로써 설계하고 있는 군집화 알고리즘은 근본적으로 계층적 군집화를 취할 예정이다. 그런데 계층적 군집화 (최종 결과물)의 문제점은 병합에서 발생한 오류가 다음 단계로 계속해서 증대되면서 전파된다는 것이다. 이러한 문제점을 줄이기 위해서는 오류를 포함하고 있는 군집을 결정해야 한다. 즉, 오류군집을 판단할 수 있는 계측이 요구되며, 아래와 같이 정의한다. 이를 위해서, 우선 몇가지 개념 및 정의들을 살펴 보자. a_i는 트랜잭션의 i 번째 애트리뷰트이고, $a_{i,j}$는 a_i의 j번째 애트리뷰트 값이다. 즉, $a_i \in \{a_{i,1}, a_{i,2}, \Lambda, a_{i,j}, \Lambda, a_{i,s_i}\}$ 이다. 여기에서 s_i는 a_i의 카디널리티(cardinality)이다. 즉, $s_i = |\{a_{i,j}\}|$이다. $f_k(a_{i,j})$는 군집 C_k에서 애트리뷰트 값 $a_{i,j}$의 발생 빈도수

이다. $\mathbf{f}_k(i)$는 군집 C_k에서 i 번째 애트리뷰트에 대한 발생 빈도 벡터로써, $\mathbf{f}_k(i) = (f_k(a_{i,1}), f_k(a_{i,2}), \Lambda, f_k(a_{i,s_i}))$ 이다. $\mathbf{p}_k(i)$는 $\mathbf{f}_k(i)$의 확률 벡터로써

$$\begin{aligned}\mathbf{p}_k(i) &= (f_k(a_{i,1}), f_k(a_{i,2}), \Lambda, f_k(a_{i,j}), \Lambda, f_k(a_{i,s_i}))/|C_k| \\ &= (p_k(a_{i,1}), p_k(a_{i,2}), \Lambda, p_k(a_{i,j}), \Lambda, p_k(a_{i,s_i}))\end{aligned} \quad (1)$$

로 정의 된다.

다음으로, $\text{InCoh}_k(i)$는 애트리뷰트 a_i의 응집성을 측정하는 계측으로써 $\{p_k(a_{i,j})\}$의 엔트로피를 이용하여 아래와 같이 정의된다.

$$InCoh_k(i) = -\sum_{j=1}^{s_i}[p_k(a_{i,j}) \cdot \log_{s_i} p_k(a_{i,j})] \quad (2)$$

잘 알려져 있듯이 엔트로피는 불확실성이 높아질수록 더 큰 값을 가지게 된다. 애트리뷰트 값 $a_{i,j}$의 발생 빈도 패턴의 입장에서 본다면, 모든 애트리뷰트 값에 대한 발생 빈도가 균등할수록 더 높은 값을 가지게 되며, 어느 한 값에 편중될수록 작은 값을 가진다. 예를 들어, $p_k(a_{i,1}) = p_k(a_{i,2}) = \cdots = p_k(a_{i,s_i}) = 1/s_i$ 일경우 $\text{InCoh}_k(i) = 1$ 이 되고, $p_k(a_{i,j}) = 1$, $p_k(a_{i,l}) = 0$ $(l \neq j)$ 일 경우, $\text{InCoh}_k(i) = 0$ 의 값을 가지게 된다. 수식 (1)에서 응집도를 ' Coh' (erence) 대신에 ' InCoh' (erence) 를 쓰고 있는데, 이것은 높은 $\text{InCoh}_k(i)$ 값이 응집 구성이 아닌 비응집 구성을 나타내기 때문이다.

위의 수식에서 한가지 주목할 만한 사항은 log의 베이스로 s_i, 즉, $\log_{s_i}$ 를 취하고 있다는 사실이다. 이것은 카디널리티가 각기 다른 애트리뷰트에 대한 $\text{InCoh}_k(i)$가 [0, 1] 의 범위를 가지도록 정규화시킨 것이다.

InCoh (C_k)는 군집에 대한 응집도을 계측하기 위해 제안 된 것으로써 아래와 같이 정의 된다. 이 값은 소위 말하는 군집의 조밀도(compactness)을 나타내게 된다.

$$InCoh(C_k) = \frac{1}{d}\sum_{i=1}^{d} InCoh_k(i) = -\frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{s_i}[p_k(a_{i,j}) \cdot \log_{s_i} p_k(a_{i,j})] \quad (3)$$

참고로, 사용 방법과 용도가 틀리지만, 엔트로피를 군집의 응집성을 측정하는 데 사용한 예를 [3]에서도 찾아 볼 수 있다.

본 논문에서는 투영 군집화를 다루고 있다. 따라서, 투영 군집화 알고리즘에서 실질적으로 사용되는 응집도는 수식 (4)의 투영 군집에 대한 대응인 투영 응집도이다. 그 정의는 아래와 같다.

$$InCoh'(C_{k_x}) = -\frac{1}{d'_k}\sum_{a'_i \in A'_k}\sum_{j=1}^{s'_i}[p_k(a'_{i,j}) \cdot \log_{s'_i} p_k(a'_{i,j})] \quad (4)$$

수식에서 a'_i은 군집 C_k의 군집 차원이다. 그리고, A'_k은 군집 차원의 집합이며, $d'_k = |A'_k|$이다.

5. Conclusions

본 논문에서는 고차원의 범주형 데이터 집합을 군집화할 때 발생하는 문제점을 고찰해 보고 그에 대한 해결책을 제시하였다. 고차원의 범주형 데이터 집합은 수치형 데이터와 달리 값들 사이의 순서가 존재하지 않으며 그 표현 능력의 한계로 인해 더욱 더 심각한 차원의 저주라는 문제점을 보여준다. 본 논문에서는 이러한 문제를 해결할 수 있는 투영 군집화를 위한 핵심 요소를 제안하였다. 그 첫번째 요소가 범주형 데이터 타입을 위한 군집차원으로써 군집의 지배패턴 기반 중심의 SDA 를 이용해 제안하였다. SDA 는 해당 군집의 데이터 객체들이 군집차원에서 고도의 연관성을 가진다는 특성을 효과적으로 표현할 수 있다. 다음으로 군집 응집도로써 엔트로피의 변형으로 정의하였다. 잘 알려져 있듯이 엔트로피는 분산도를 효과적으로 표현하는 것으로 알려져 있다. 본 논문에서 제안한 군집 응집도는 카디널리티가 각기 다른 애트리뷰트들로 구성된 데이터 객체 집합의 분산도를 표현할 때 아주 유용할 것으로 기대된다.

References

[1] C.C. Aggarwal, C. Procopiuc, J.L. Wolf, P.S. Yu, and J.S. Park, "Fast Algorithms for Projected Clustering," Proc. ACM SIGMOD Conf. Management of Data, pp. 61-72, 1999.

[2] K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is `Nearest Neighbor' Meaningful?," Proc. Int'l Conf. Database Theory (ICDT), pp.217-235, Jerusalem, Israel, 1999.

[3] M. Dash, H. Liu, and J. Yao, "Dimensionality Reduction of Unsupervised Data," Proc. ICTAI'97 (the 9th Int'l Conf. Tools with Aritficial Intelligence), pp. 532-539, 1997.

[4] V. Ganti, J. Gehrke, and Raghu Ramakrishnan, "CACTUS-Clustering Categorical Data Using Summaries," Proc. KDD'99, pp. 73-83, 1999.

[5] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," J. Information Systems, vol. 25, no. 5, pp. 345-366, 2000.

[6] M. Kim, H. Yoo, and R.S. Ramakrishna, "Cluster Validation for High-Dimensional Datasets", Lecture Notes in Artificial Intelligence (AIMSA 2004), Vol. 3192, pp. 178-187, Sept. 2004.

[7] K. Wang, C. Yu, and B. Liu, "Clustering Transactions Using Large Items," Proc. CIKM'99, pp. 483 ~ 490, 1999.

[8] Y. Yang, X. Guan, and J. You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data," Proc. KDD'02, pp. 682-687, 2002.