제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

범주형 데이터 집합에 대한 엔트로피 기반 군집 유효화 기술

박남현, 안창욱, R.S. Ramakrishna
광주과학기술원 정보통신공학과
e-mail : nhpark@gist.ac.kr

Entropy-based Clustering Validation Technique for Categorical Data Sets

Namhyun Park, Chang Wook Ahn, R.S. Ramakrishna
Dept. of Information Communication, Gwangju Institute of Science and Technology

요 약

본 논문에서는 고차원의 특성을 가진 범주형 데이터 집합의 군집 유효화 기술에 대하여 알아본다. 먼저, 범주형 데이터 집합에 대하여 한 군집의 센트로이드를 정의함에 따라 일반적인 군집화 방법에서 사용되는 쌍 유사성 측정을 가능하게 한다. 다음으로, 범주형 데이터 집합에 대한 증분 군집 알고리즘을 통하여 도출된 결과에 대해 최적 군집 수의 결정하기 위하여 엔트로피 기반 군집 유효화 지수를 사용한다. 이를 통하여 일반적인 군집 알고리즘에서 최적 결과를 얻기 위해 필요한 문턱값 결정 문제를 손쉽게 해결한다. 마지막으로, 위의 개념들을 여러 데이터 집합에 대해 실험한다.

1. 서론

군집화 (Clustering)는 전체 데이터를 유사한 객체들의 그룹들로 분할하는 작업으로 데이터 마이닝의 중요한 영역 중의 하나이다 [1]. 그리고 군집 유효화 기술 (Clustering validation technique) 은 특정 데이터 집합에 대한 군집 결과에 대한 최적 군집 수를 결정하는 기술을 말한다 [8] [9]. 그러나 대부분의 현존하는 군집 유효화 기술들은 수치 데이터 집합에만 적합하게 개발 되었고 범주형 데이터 집합에는 적합하지 않았다. 최근 들어 범주형 데이터 집합의 군집 유효화 기술에 대한 연구가 활발히 이루어지고 있다 [2] [3] [4] [5]. 그러나 종전의 범주형 데이터 집합에 대한 군집 유효화 기술에서 볼 수 있듯이, 센트로이드를 정의하고 이를 이용한 군집 유효화 기술의 연구는 미흡하다. 본 논문에서 제안 군집 유효화 기술에서는 범주형 데이터 집합에서 한 군집의 센트로이드를 정의하고, 이를 이용한 군집 유효화 지수 (Clustering validation index)를 설계한다.

우리는 한 군집에 포함되는 범주형 데이터 집합의 각 에트리뷰트의 값에 대한 빈도수에 대한 히스토그램을 각 군집에 속하는 객체의 총 수로 나누어 센트로이드를 정의를 한다. 이를 통하여 각 군집의 센트로이드를 통하여 변형된 Shannon entropy 를 이용하여 내부 유사성을 측정하고, 각 센트로이드 사이에서 변형된 Euclidean distance 를 이용하여 외부 비유사성을 측정한다 [12] [13]. 다음으로, 위의 개념들을 이용해 군집 유효화 지수를 설계하고 특정 데이터 집합에 대한 최적 군집 수를 결정하는 적절한 일반적인 군집 알고리즘에 필요한 문턱값 (Threshold)을 손 쉽게 얻을 수 있다. 마지막으로, 합성 데이터 집합과 실 세계 데이터 집합의 군집화된 결과에 대하여 제안 군집 유효화 지수로 실험한다.

본 논문의 구성은 다음과 같다. 2 절에서는 범주형 데이터 집합에 대한 기존의 일반적인 알고리즘 바탕으로 설계되진 증분 군집화 (Incremental clustering)에 대하여 알아 본다. 3 절에서는 제안된 군집 유효화 기술을 언급하고, 4 절에서는 실험과 이에 대한 분석한다. 마지막으로, 5 절에서 결론을 제시한다.

2. 증분 군집화 (Incremental Clustering)

본 절에서는 범주형 데이터 집합에 대한 센트로이드를 이용한 증분 군집화에 대하여 알아본다. 종전의 센트로이드를 이용한 군집 알고리즘은 *Ordonez 알고리*

즘 그리고 *Huang 알고리즘* 등이 있다 [10] [11]. 이들 모두 국부적인 접근 방법으로 정의된 평가 함수를 이용하고, 일반적으로 수치 데이터 집합에서 사용되는 유사성 측정법으로 두 점 간의 거리를 중심으로 하여 군집화하는 방법을 채택하고 있다. *Ordonez 알고리즘*과 *Huang 알고리즘*은 데이터의 각 에트리뷰트 값에 대하여 히스토그램을 바탕으로 한다. 먼저, *Ordonez 알고리즘*은 주어진 히스토그램을 각 군집 객체의 총 수로 나누어 센트로이드를 정의하고, 이에 대해 Euclidean distance 를 이용하여 쌍 유사성을 측정하면서 군집화한다. 두 번째로, *Huang 알고리즘*은 주어진 히스토그램을 일정한 문턱값을 주어 한 군집에 대한 센트로이드를 정의하고, 이를 Jaccard coefficient 를 이용하여 쌍 유사성을 측정하면서 군집화한다 [2].

본 논문에서 제안된 군집 유효화 지수를 실험하기 위하여 *Ordonez 알고리즘*을 바탕으로 하는 증분 군집 알고리즘을 사용한다 [10]. 2.1 에서는 범주형 데이터 집합에 대한 센트로이드 결정 문제에 대하여 언급한다. 그리고 2.2 에서는 제안된 군집 유효화 지수를 실험에 필요한 군집 결과를 생성시킬 증분 군집 알고리즘을 소개한다.

2.1 센트로이드 결정

직관적으로 범주형 트랜잭션 데이터 집합에서의 센트로이드 결정은 일반적인 수치 데이터 집합과 다르다. 다음 예를 통하여 센트로이드가 어떻게 설정되는지 알아 보자. 그림 1 의 (a)에서 볼 수 있듯이, 한 군집의 히스토그램이 있다고 가정하면 (b)에서 각 군집의 히스토그램을 그 군집이 가지는 객체의 총 수로 나누어 센트로이드로 설정할 수 있다.

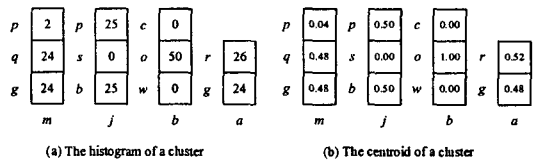

그림 1. 한 군집 C_i 의 히스토그램과 센트로이드

이로써 위에서 언급한 (a)의 한 군집의 히스토그램에서 (b)와 같은 센트로이드를 얻을 수 있다. 왜냐하면 대형 할인형 매장에서 다수의 구매자들이 물건을 구입할 때, 구매자들이 물건을 항상 같은 패턴으로 구입하지 않기 때문이다. 다시 말하면, 그림 1 의 (a)와 같이 대부분이 {*m*q, j*p, b*o, a*r*}과 {*m*g, j*b, b*o, a*g*} 등과 같은 패턴으로 구매자가 물건을 사지만, 어떤 구매자들은 {*m*p, j*p, b*o, a*r*}과 {*m*p, j*b, b*o, a*g*}등과 같이 물건을 살 때도 있다.

2.2 증분 군집 알고리즘

범주형 데이터 집합에 대한 군집 과정 중에 필요한 쌍 유사성은 객체와 센트로이드 사이에서 얻는다. 그림 2 는 증분 군집 알고리즘에 필요한 쌍 유사성의 예에 대한 간단한 도식이다.

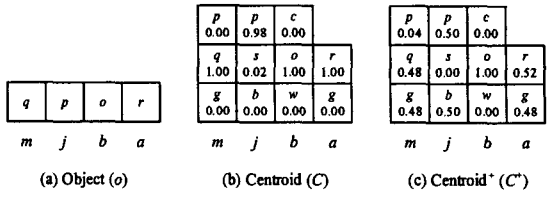

그림 2. 군집 결정에 필요한 쌍 유사성의 예

위에서 만들어진 각 군집의 센트로이드를 이용해 객체와 센트로이드 사이에서 일반적인 유사성 측정법인 Euclidean distance 를 그림 2 와 같이 적용시킬 수 있다. 여기에서 (a)는 특정 군집에 포함되는 한 객체 (o)라고 하고 (b)와 (c)를 각 군집이 가지는 센트로이드로 C 와 C^+ 라고 하자. 그리고 A 는 에트리뷰트 개수를 의미한다. 그래서 (a)와 (b) 간의 Euclidean distance 인 $d(o,C)=\sqrt{\sum_{i=0}^{A-1}(o_i-C_i)^2}=\sqrt{(1-1)^2+(1-0.98)^2+(1-1)^2+(1-1)^2}=\sqrt{0.0004}$ 를 구할 수 있다. 또한, (a)와 (c)에서는 $\sqrt{0.7508}$ 를 얻을 수 있다. 이처럼 비슷한 패턴을 가진 객체들이 모여 한 군집을 만들어진 경우 객체와 센트로이드 간의 관계에서 Euclidean distance 정의에 따라 값이 작을 때 쌍 유사성이 크다는 것을 알 수 있다.

3. 제안 군집 유효화 지수

본 절에서는 센트로이드를 통하여 특정 데이터 집합의 각 군집의 센트로이드에서 내부 유사성과 각 센트로이드 사이에서 외부 비유사성을 알아 낸다. 이를 통하여 범주형 데이터 집합의 군집 유효화 지수를 알 수 있다. 또한, 군집 유효화 지수를 통하여 군집 알고리즘에서 최적 군집 수 결정에 필요한 문턱값을 구할 수 있다.

3.1 유사성 측정

그림 3 에서는 데이터 집합의 각 센트로이드에서 내부 유사성 측정과 각 센트로이드 사이에서 외부 비유사성 측정에 알아 본다.

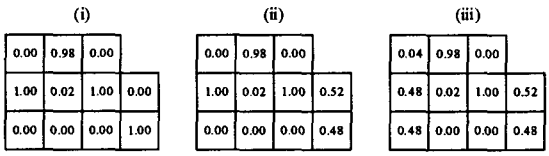

그림 3. 내부 유사성과 외부 비유사성

우리는 위에서 만들어진 각 군집의 센트로이드를 이용해 내부 유사성을 변형된 Shannon entropy, 즉, $H_{intra}=\sum_{j=0}^{A-1}\left(-\sum_{u=0}^{AV_j-1}p_u(x)\log_2 p_u(x)\right)$ 를 (a)~(c)까지 얻을

수 있다 [12]. 여기에서 A 는 에트리뷰트 개수, AV 는 각 에트리뷰트에 포함되는 에트리뷰트 값에 대한 개수, 그리고 $p(x)$는 한 군집의 에트리뷰트 값에 대한 확률을 의미한다. 각 센트로이드 사이에서 외부 비유사성을 변형된 Euclidean distance, $H_{inter} = \sum_{i,j=0}^{A-1} \sum_{v,u=0}^{AV_i,_{j-1}} \|p_v(x) - p_u(x)\|^2$ 를 이용해 (c), (d)를 구할 수 있다.

이를 일반화하면, 군집화가 잘 이루어졌다면 H_{intra} 가 작고, 군집화가 잘 이루어지지 않았다면 H_{intra} 가 크다. 또한, 각 센트로이드 사이의 거리가 가깝다면 H_{inter} 가 작고, 멀다면 H_{intra} 가 커지는 것을 알 수 있다. 이는 [7]에서 언급한 Compactness 와 Separation 의 개념으로 전자와 후자를 대치하여 설명할 수 있다. 더 자세히 말하면, H_{intra} 나 H_{inter} 가 작다면 내부 유사성과 외부 비유사성이 크고, H_{intra} 나 H_{inter} 가 크다면 내부 유사성과 외부 비유사성이 작다는 것을 알 수 있다. 다음 절에서는 새로운 엔트로피 기반 군집 유효화 지수를 제시하고 최적 군집 수를 얻는데 필요한 문턱값에 대해 언급한다.

3.2 군집 유효화 지수

군집화의 과정에서 최적화된 군집 결과를 얻기 위하여 적절한 문턱값의 결정이 매우 중요하다. 그러나 실세계 사용자들에게는 문턱값 결정 자체가 매우 어려운 일이다. 여기에서는 유사성 측정법인 거리 함수로 Shannon entropy 와 Euclidean distance 를 사용한다. 각 군집의 센트로이드에서 내부 유사성과 각 센트로이드 사이에서의 외부 비유사성을 통하여 최적 군집 수를 찾아 내는데 필요한 적절한 문턱값을 결정할 수 있다. 여기에서 K 는 특정 데이터 집합을 군집화할 때 예상되는 대략적인 범위의 최대값을 의미한다.

$$
\begin{aligned}
H(K) &= \frac{H(K)_{intra}}{H(K)_{inter}} \\
H(K)_{intra} &= \sum_{i=0}^{K-1} \sum_{j=0}^{A_i-1} \left(- \sum_{u=0}^{AV_j-1} p_u(x) \log_2 p_u(x) \right) \\
H(K)_{inter} &= \underset{g,h=0}{\overset{K-1}{MIN}} \left(\sum_{i,j=0}^{A_g,_{h-1}} \sum_{v,u=0}^{AV_i,_{j-1}} \|p_v(x) - p_u(x)\|^2 \right) \\
& \qquad , where\ g \neq h
\end{aligned}
\tag{1}
$$

이처럼 식 1 의 군집 유효화 지수에서 3.1 에서 언급한 내부 유사성을 통하여 $H(K)_{intra}$ 를 얻을 수 있고, 외부 비유사성을 통하여 $H(K)_{inter}$ 를 얻을 수 있다. 이는 내부 유사성과 외부 비유사성을 [7]의 군집의 Compactness 와 군집 사이의 Separation 의 관계로 연관을 지을 수 있기 때문이다. 이 문제를 해결하기 위해서 Compactness 나 Separation 을 $H(K)_{intra}$ 나 $H(K)_{inter}$ 를 이용해 최적 군집 수를 얻을 수 있는 적절한 문턱값을 얻을 수 있다 [7] [8]. 여기에서 $H(K)$가 최소값일 때 최적 군집 결과를 도출한다.

4. 실험 결과

이번 절에서는 여러 합성 데이터 집합과 실 세계 데이터 집합에 대한 증분 군집 알고리즘을 실험하고, 이 결과를 군집 유효화 기술에 적용하여 최적의 문턱값을 찾아 낸다. 일반적으로 군집화에 대한 성능은 주로 군집 수와 군집 순도에 의해 평가된다. 즉, 군집 결과가 데이터 집합과 일치하는 군집 수를 가지며 순도가 높을수록 좋은 품질을 가졌다라고 할 수 있다. 이는 군집 유효화 지수에 의해 예측된 최적 군집 수가 실제 데이터의 최적 군집 수 보다 작거나 클 경우 순도가 급격하게 악화되는 현상에서 이를 확인할 수 있다.

4.1 합성 데이타 집합

위에서 언급한 것처럼 군집 수와 군집 순도와의 관계에서 군집 결과가 실제 군집과 일치하는 군집 수를 가진다면 직관적으로 군집 순도가 높다는 것을 알 수 있다. 이를 증명하기 위하여 다음의 범주형 데이터 집합인 t0~t2 에서 군집 수와 군집 순도와의 관계를 규명한다. 다음 표 1 은 합성 데이터 집합의 요약이다.

표 1. 합성 데이터 집합의 요약

Data sets	# objects	# attributes	# cluster	Perturbation
t0	9000	20	20	0%
t1	9000	20	20	10%
t2	9000	20	20	20%

표 2 에서는 t2 데이터 집합에 대해 $H(K)_{intra}$ 와 $H(K)_{inter}$ 를 통하여 제안된 군집 유효화 지수, $H(K)$를 이용하여 최소값을 얻었고, 이로 인해 이미 알려진 20 개의 군집을 얻었다. 이때 입력한 문턱값이 t2 데이터 집합의 최적 군집 수를 결정 짓는 적절한 문턱값임을 알 수 있다.

표 2. t2 데이터 집합에 대한 군집 유효화 지수

	$H(K)_{intra}$	$H(K)_{inter}$	$H(K)$
11	0.576288	0.986543	0.584149
20	0.370603	0.989774	0.374432
25	0.370209	0.61388	0.603065
48	0.368783	0.558387	0.660443
91	0.365499	0.466389	0.783678
217	0.355417	0.421945	0.842330
439	0.337917	0.421945	0.800856
835	0.308955	0.375294	0.823234

다음의 그림 4 는 일반적인 특정 데이터 집합의 군집화에서 얻어지는 내부 유사성과 외부 비유사성과 관련된 $H(K)_{intra}$, $H(K)_{inter}$, $H(K)$의 관계를 보여준다. 이에 대한 자세한 내용은 [7]에서 확인할 수 있다.

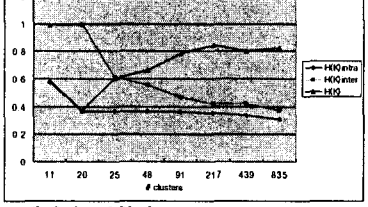

그림 4. t2 데이터 집합의 $H(K)_{intra}$, $H(K)_{inter}$, $H(K)$에 대한 그래프

4.2 실 세계 데이터 집합

위의 4.1 에서 군집 수와 군집 순도와의 관계에 대해서 알아 보았다. 이번 절에서는 실 세계 데이터 집합인 Congressional votes 와 Mushroom 의 실험 결과를 나타낸다 [6]. 표 3 은 실 세계 데이터 집합의 요약이다. 다음의 표 4 과 5 는 $H(K)$가 최소값을 가질 때에 문턱값에 따라 얻어지는 각 군집 결과이다. 결과는 종전의 범주형 데이터 집합의 알고리즘인 *LargeItem*, *CLOPE*, 그리고 *ROCK* 등과 비교하고 있다 [2] [4] [5].

표 3. 실 세계 데이터 집합의 요약

Data sets	# objects	# attributes	# cluster	Remark
Congressional votes	435	16 (few missing values)	2	Rep. (168) Dem. (267)
Mushroom	8124	22 (few missing values)	2	Edi. (4208) Poi. (3916)

표 4. Congressional votes 데이터 집합의 군집 결과

No. of cluster	No. of republicans	No. of democrats
1	153	62
2	6	214

표 5. Mushroom 데이터 집합의 군집 결과

# cluster	# edi.	# poi.	# cluster	# edi.	# poi.
1	0	256	14	0	864
2	512	0	15	48	0
3	424	0	16	48	0
4	96	0	17	0	32
5	192	0	18	0	8
6	96	0	19	0	648
7	864	0	20	32	72
8	344	0	21	192	0
9	0	648	22	0	864
10	0	192	23	288	0
11	864	0	24	0	36
12	192	0	25	0	8
13	0	288	26	16	0

5. 결론

본 논문에서는 고차원의 특성을 가진 범주형 데이터 집합의 군집 유효화 기술에 대하여 알아 보았다. 먼저, 범주형 데이터 집합을 일반적인 수치 데이터 집합처럼 한 군집의 센트로이드로 정의하였다. 그리고 증분 군집 알고리즘으로 여러 합성 데이터 집합과 실 세계 데이터 집합을 군집화하였다. 또한, 센트로이드를 통해 변형된 Shannon entropy 와 변형된 Euclidean distance 를 바탕으로, 데이터 집합의 각 센트로이드에서 내부 유사성과 외부 비유사성 측정을 가능하게 하였다. 그리고 군집 유효화 지수를 설계하였다. 마지막으로, 증분 군집 알고리즘에서 도출된 군집 결과를 제안된 군집 유효화 지수에 적용하여 최적 군집 수 결정에 필요한 문턱값을 결정하였다.

참고문헌

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001.

[2] S. Guha, R. Rastogi, and K. Shim, *ROCK: A Robust Clustering Categorical Data using Attributes*, Information systems Vol. 25, No. 5, pp. 345-366, 2000

[3] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS: Clustering data using summaries," *Proc. Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, San Diego, CA, 1999.

[4] K. Wang, C. Xu, and Liu, B., "LargeItem: Clustering Transactions using Large Items," *Proc. Int'l. Conf. Information and Knowledge Management (CIKM)*, Sydney, Australia, 1999.

[5] Y. Yang, X. Guan, and J. You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data," *Proc. Int'l Conf. ACM Special Interest Group on Knowledge Discovery in Data and Data Mining (SIGKDD)*, Edmonton, Alberta, Canada, 1002.

[6] C.L. Blake and C.J. Merz, *UCI Repository of machine learning databases*, http://www.ics.uci.edu/~mlearn/MLRepository.html, Univ. of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[7] M.J.A. Berry, G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.

[8] X.L. Xie and G.A. Beni, *A Validity Measure for Fuzzy Clustering*, IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI), 1991.

[9] D.-J. Kim, Y.-W. Park, and D.-J. *Park, A Novel Validity Index for Determination of the Optimal Number of Clusters*, IEICE Trans. Inf. & Syst., 2001.

[10] Z. Huang, *Extensions to the k-means Algorithm for Clustering Large Data Sets Categorical Values*, Data Mining and Knowledge Discovery, Vol. 2, No. 3, 1998.

[11] Carlos Ordonez, "Clustering Binary Data Streams with K-means," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, San Diego, California, USA, 2003.

[12] Jianhua Lin, *Divergence Measures Based on the Shannon Entropy*, IEEE Transactions On Information Theory, Vol. 37, No. 1, January 1991.

[13] Erhan Gokcay and Jose G., *Information Theoretic Clustering*, IEEE Transactions On Pattern Analysis and Machine Intelligence Vol. 24, No. 2, February 2002.