거리 히스토그램을 이용한 특성 추출 기법

최기석O 전성진 양명석
한국과학기술정보연구원
{choiO, sjchon, msyang}@kisti.re.kr

Feature Selection by Using Distance Histogram

Kiseok ChoiO Sungjin Chon Myungseok Yang
Korea Institute of Science and Technology Information

요 약

특성 추출은dimensionality reduction technique로서 잡음을 제거하기 위해 사용되는 중요한 전처리 방식이다. 이러한 과정을 통해 데이터의 크기를 줄일 수 있으며 학습의 정확성 및 이해도를 높일 수 있다. Classification에 사용되는 다양한 특성 추출방식들이 존재하는 반면에 클러스터링에 적용될 수 있는 방식들은 양적으로도 많이 부족하며 존재하는 방식들도 대부분 사용되는 클러스터링 알고리즘 자체에 의존적인 실세계 어플리케이션에는 적용하기 부적합한 wrapper 방식을 도입하고 있다. 본 논문에서는 클러스터링 알고리즘으로부터 독립적인 필터 솔루션(filter solution)을 제안하였다. 이 방식은 클러스터를 가진 데이터와 가지지 않고 있는 데이터 사이의 point-to-point 거리 히스토그램의 차이에 기반하고 있다

1. 서 론

현재 사용되고 있는 많은 어플리케이션들은 고차원 데이터를 다루고 있다. 특성 추출 방식은 주요한 특징들을 추출하여 이러한 데이터의 차원을 줄여주는 기법이다. 학습의 정확도 및 이해도를 저하시키는 불필요한 요소들을 제거하여 데이터의 크기를 줄여주는 효과 또한 가져온다. 학습에는 supervised와 unsupervised 학습의 두 가지가 있다. 클러스터링은unsupervised 학습방식을 도입하기 때문에 특성 추출 방식을 하는데 어려움이 많으며 실제로도 다양한 방식들이 존재하지 않는다.

클러스터링에서의 특성 추출은 클러스터로부터 주요한 특성들을 선택하는 것이다. 기존의 방식들은[8,9,11] 학습에 사용되는 알고리즘에 기반하여 후보가 되는 특성군들을 평가하는 wrapper 방식에 기반하고 있다. 그러나 이 방식은 클러스터링의 정확도를 측정할 수 있는 객관적인 기준의 부재와 다른 서브스페이스 상에 존재하는 클러스터링을 구분하는 방식을 필요로 하는 문제점 등을 안고 있다.

이러한 문제점들을 해결하기 위해 본 논문에서는 클러스터링에서 특성을 추출하기 위한 '필터' 방식을 도입한다. 이 방식은 클러스터링 알고리즘으로부터 독립적이며 데이터 상에 클러스터의 존재 유무에 따른 point-to-point 거리 히스토그램의 차이에 기반하고 있다. 이를 이용한 엔트로피 측정(entropy measure)을 사용하는데, 이는 구분이 되는 클러스터가 존재할 경우에 낮은 수치를 보인다. 엔트로피 수치는 데이터의 다차원성에 영향을 받지 않고 클러스터링 자체의 수준에 의해서만 영향을 받기 때문에 특성 추출에 사용하기 가장 적합하다. 본 방식의 효율성을 입증하기 위해 실제 데이터 및 벤치마크 데이터 등에 대한 실험을 수행한다.

2. 거리 히스토그램 (Distance Histogram)

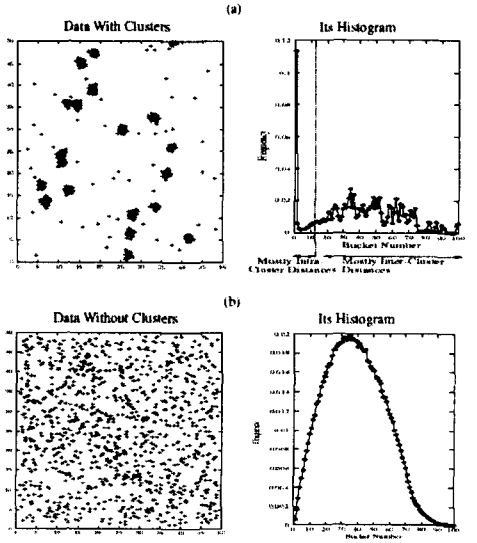

그림 1 클러스터 유무에 따른 히스토그램 변화

[그림 1]은 클러스터를 가진 데이터 집합과 클러스터가 없는 데이터 집합의 점-대-점 거리 히스토그램이다. 히스토그램 사이의 모든 거리는 [0..1] 사이의 값으로 정규화되며 각 히스토그램은 b개의 버켓을 가진다. 버켓값은 정규화된 거리값을 b와 곱한 값으로 정의된다. 각 버켓별로 카운터를 두어 하나의 거리가 해당 버켓으로 할당될때마다 증가된다. 히스토그램 상의 각 점은 하나의 버켓을 의미하며 해당 점의 x축 값은 버켓값을, y축은 버켓 x의 카운터 값을 나타낸다. 두 히스토그램을 살펴보면 클러스터를 가진 데이터 집합은 벨 형태로 구성되는 반면에 그렇지 못한 데이터

의 집합은 다른 형태를 취한다는 것을 알 수 있다. 클러스터를 포함하고 있는 데이터 집합의 경우 같은 클러스터 내의 거리(intra-cluster distance)가 클러스터들 사이의 거리(inter-cluster distance)보다 작다는 것을 알 수 있다. 클러스터들의 구분이 확연해질수록 이들 거리의 차이 또한 명확해진다. 본 논문에서는 실제 클러스터링을 적용하지 않고도 클러스터를 가진 데이터 집합과 그렇지 못한 데이터 집합을 구분할 수 있는 방식을 제안한다.

3. 거리-기반 엔트로피 값의 효율적인 계산

우리는 시스템 상의 엔트로피를 이용하여 시스템의 부조화를 측정할 수 있다는 것을 엔트로피 이론을 통하여 알 수 있다. 데이터 집합의 엔트로피는 다음의 수식에 의해 정의된다.

$$E = -\sum_{X_i} p(X_{i1},...,X_{iM}) \log p(X_{i1},...,X_{iM}) + (1-p(X_{i1},...,X_{iM})) \log(1-p(X_{i1},...,X_{iM}))$$

여기에서 $p(x_{i1}..x_{iM})$는 점 $(x_{i1}..x_{iM})$에서의 확률 또는 밀도를 나타낸다. 두 번째 항목은 수식을 대칭되게 해주는 역할을 한다. 모든 점들의 확률이 같을 경우 우리는 결과값을 전혀 예측할 수가 없으며 엔트로피 값은 최대치로 나타난다. 이러한 현상은 데이터 점들이 특성 공간 상에 일정한 형태로 분산되어 있을 경우에 나타난다. 반면에 데이터들이 클러스터를 가지고 있는 경우에는 예측 가능성이 높아지며 엔트로피 값 또한 최소화된다. 즉, 엔트로피 값을 이용하여 클러스터를 가진 데이터와 그렇지 못한 데이터를 구분할 수 있다.

그러나 각 점에서의 확률을 모르기 때문에 다음의 프록시(proxy) 방식을 이용하여 엔트로피 값을 예측하는 방식을 제안한다. 확률을 거리로 대체하는 방식으로 확률에 대한 사전 정보 없이 엔트로피 값을 측정할 수 있다.

$$E = -\sum_{X_i}\sum_{X_j} Dij \log Dij + (1-Dij)\log(1-Dij)$$

여기에서 D_{ij}는 인스턴스 X_i와 X_j 사이의 거리를 [0.0-1.0] 사이의 범위에서 정규화한 값이다. 그림 4(a)는 E를 [0,0-1.0] 사이의 값으로 정규화한 후에 엔트로피 값과 거리 사이의 관계를 나타낸다. 거리가 가장 작거나 가장 큰 거리에 가장 낮은 엔트로피 값을 할당하고 거리의 평균값에 가장 높은 엔트로피 값을 할당한다. 이 방식은 클러스터에 기반하여 데이터 집합들을 구분해주는 장점을 가지는 반면 다음의 두 가지 단점 또한 안고 있다.

첫 째, 평균 거리 (0.5)에 의해 구분되는 클러스터 사이의 거리는 양 쪽 점의 교차점이 0.5로 정해져 있기 때문에 최대 엔트로피 값을 할당받게 된다. 또한 점들의 집합이 가지는 모양으로 인하여 한 클러스터 내의 데이터들의 거리가 약간만 증가해도 엔트로피 값이 따라서 증가하게 된다.

첫 번째 문제점은 클러스터 사이의 거리 및 클러스터 내부의 거리를 구분할 수 있는 교차 지점 (u)를 지정하여 해결할 수 있다. 위의 모든 사항들을 고려하여 다음과 같은 방식을 제안한다.

$$E = \sum_{X_i}\sum_{X_j} Eij$$

$$Eij = \begin{cases} \dfrac{exp(b^* Dij) - exp(0)}{exp(b^* u) - exp(0)} & , 0 <= Dij <- u \\ \dfrac{exp(b^*(1-Dij)) - exp(0)}{exp(b^*(1-u)) - exp(0)} & , u <= Dij <= 1 \end{cases}$$

그림 4(b)는 b의 값을 증가시키면 개별 엔트로피 값 및 결과적으로는 전체적인 엔트로피 값이 감소된다는 것을 보여준다. 또한 각기 다른 u 값을 설정하게 되면 양쪽 엔트로피의 교차점을 이동시키는 효과를 가져올 수 있다는 것을 그림 4(c)를 통해 알 수 있다.

4. 특성 추출 알고리즘

이번 장에서는 엔트로피 값에 기반한 특성 추출 알고리즘에 대하여 설명한다. 특성 추출은 크게 다음과 같은 두 과정을 통해서 수행된다. 첫 번째 단계는 생성 또는 검색과정이며 다음은 특성 집합들의 평가 단계이다. 엔트로피 값은 이러한 특성 집합들을 평가하기 위해 사용될 수 있다. 잘 구성되어 있는 클러스터의 경우 엔트로피 출력 값이 낮으며 그렇지 않은 경우 엔트로피 출력 값이 높게 나타난다. 검색 방법의 경우 정확도는 최적성에 의해 결정된다. 최적성은 엔트로피 값이 최소인 데이터의 서브 집합으로 정의된다. 지금까지 많은 검색기법이 제안되어 왔다.[7] 이러한 검색기법에는 랜덤 검색, 경험적 방법 및 몇 가지 방식을 혼합한 혼합방식 등 다양한 방식들이 존재한다. 검색은 가장 낮은 엔트로피 값을 추출해내는 방식을 찾기 위한 과정으로 이들 검색 방식에 대한 자세한 내용은 생략하도록 하겠다. [표 1]은 forward selection 방식의 간단한 알고리즘을 보여준다.

알고리즘: ForwardSelect

```
INPUT: DataD, Original Feature Set S
OUTPUT: Selected Features
1. selected = ∅
2. overallLowestEntropy = A-veryHigh-value
3. for size = 1 to M
4.     lowestEntropy = A-veryHigh-value
5.     for I = 1 to M
6.         if S[i] ∉ selected
7.             tempSubset = append(selected, s[i])
8.             tempEntropy = E(tempSubset)
9.             if tempEntropy < lowestEntropy
10.                lowestEntropy = tempEntropy
11.                selectedFeature = S[i]
12.        selected = append(selected, selectedFeature)
13.        if E(selected) < overallLowestEntropy
14.            overallLowestEntropy = E(selected)
15.            overallSelected = selected
16. return overallSelected
```

표 1 특성추출 알고리즘

5. 실험 평가

클러스터링과 같은 unsupervised 학습 방식에 대한 평가는 공통적으로 적용 가능한 평가 방법이 없기 때문에 classification과 같은 supervised 학습 방법에 비해 많은 어려움들이 있다. 또한 클

러스터링으로부터 특성을 추출하는 것은 데이터의 다차원성에 영향을 받고 각 데이터 서브 집합별로 별도의 클러스터가 형성될 수 있기 때문에 더더욱 어려움이 가중된다. 이러한 여러 가지 어려움 때문에 클러스터링에 사용되는 특성 추출 방식에 대한 평가는 추출되어진 특성들의 적합성 여부를 판단하는 방식이 가장 적합하다고 할 수 있다. 즉, 해당 방식에 의해 추출된 특성들을 실제로 중요한 특성들과 비교하는 방식을 사용한다. 이렇게 하기 위하여 인의적으로 구성된 주요 특성들이 알려진 데이터 집합에 대한 평가를 우선 수행하게 된다. 그 이후에 여기에서 추출된 평가 항목에 따라서 사람이 직접 눈으로 주요한 특성들을 추출할 수 있는 실제 데이터 집합 및 벤치마크 데이터에 이들 방식을 적용하여 보게 된다. 또한 forward selection 방식을 적용한 후 이 결과를 exhaustive 검색 방식을 이용하여 결과가 최소 엔트로피 값을 가진 서브 집합들을 제대로 추출했는지 여부를 검증하게 된다. 그 이후에 제안된 필터 방식과 기존의 wrapper 방식의 결과를 비교 분석하게 된다. 본 논문의 실험은 3장에서 나온 가이드라인을 따라서 진행되었으며 그 결과는 다음의 그림들에 나타나 있다.

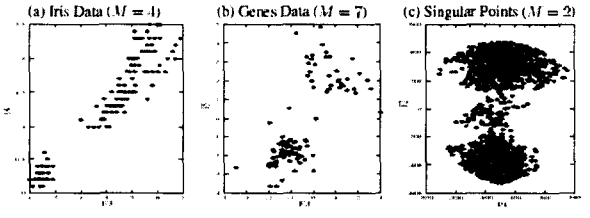

그림 2 클러스터가 있는 데이터

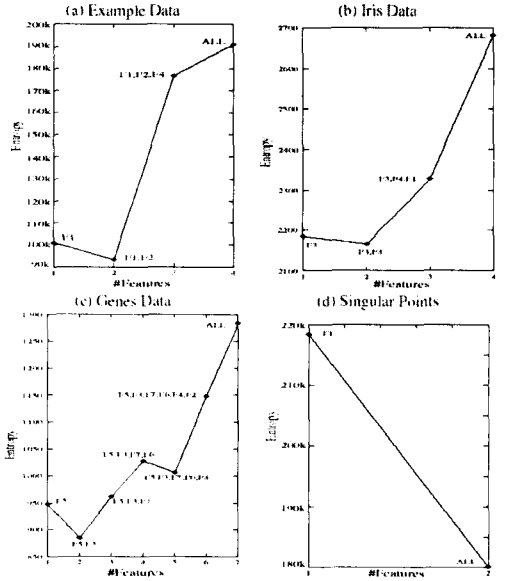

그림 3 특징추출 결과

6. 결론

본 논문에서는 클러스터링에서 특성 추출을 위한 필터 방식을 제안하고 이들의 성능을 평가하여 분석했다. 5장에서 얘기한 인위적으로 구성된 데이터 집합, 실제 데이터 및 벤치마크 데이터에 본 논문에서 제안한 방식을 적용해본 결과 본 논문에서 제안한 방식의 효율성을 입증할 수 있었다. 기존의 방식들의 경우 특성 집합을 평가하기 위해 클러스터링 알고리즘에 의존적인 wrapper 방식을 사용하였다. 클러스터링 알고리즘은 클러스터의 숫자와 같은 다양한 파라미터 값에 민감하다. 또한 이러한 정보는 실세계 데이터로부터 추출하기 어렵기 때문에 적용이 어렵다. 반면에 본 논문에서 제안한 방식은 클러스터 내부의 거리와 같이 구하기 쉬운 파라미터 값을 사용하고 있다. 또한 본 논문에서 제안한 방식은 사전 정보 없이도 데이터 집합들로부터 클러스터를 추출해내는데도 용이하다.

이후에는 이 방식을 서브스페이스 클러스터링[2] 및 불규칙적인 형태의 클러스터에 적용하는 방식에 대한 연구를 수행해볼 수 있을 것이다.

[참고 문헌]

[1] C. C. Aggarwal, C. Procopiuc, J. L. Wolf, P. S. Yu, and J. S. Park. Fast algorithms for
projected clustering. In *Proceedings of ACM SIGMOD Conference on Management of Data*,
pages 61?2, 1999.
[2] R Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of
high dimensional data for data mining applications. In *Proceedings of ACM SIGMOD Conference
on Management of Data*, 1998.
[3] R Agrawal and R. Srikant. Fast algorithm for mining association rules. In *Proceedings of
20th International Conference on Very Large Databases (VLDB)*, Santiago, Chile, 1994.
[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is ?earest neighbor?meaningful?
In *Proceedings of the International Conference on Database Theory (ICDT)*, pages
217?35, 1999.
[5] C.L. Blake and C.J. Merz. UCI repository of machine learning databases.
http://www.ics.uci.edu/?mlearn/MLRepository.html, 1998.
[6] C. Cheng, A. W. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical
data. In *Proceedings of International Conference on Knowledge Discovery and Data Mining
(KDD)*, 1999.
[7] M. Dash and H. Liu. Feature selection for classication. *International Journal of Intelligent
Data Analysis*, 1(3), 1997.
[8] M. Dash and H. Liu. Feature selection for clustering. In *Proceedings of Fourth Pacic-Asia
Conference on Knowledge Discovery and Data Mining, (PAKDD)*, 2000.
[9] M. Devaney and A. Ram. Efcient feature selection in conceptual clustering. In *Proceedings
of the International Conference on Machine Learning (ICML)*, pages 92?7, 1997.
[10] R. O. Duda and P. E. Hart. *Pattern Classication and Scene Analysis*, chapter Unsupervised
Learning and Clustering. John Wiley & Sons, 1973.
[11] J. G. Dy and C. E. Brodley. Visualization and interactive feature selection for unsupervised
data. In *Proceedings of the International Conference on Knowledge Discovery and Data
Mining (KDD)*, pages 360?64, 2000.