Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 2, pp. 481 ~ 488

Initial Mode Decision Method for Clustering in Categorical Data

Soon Cheol Yang[1], Hyung Chang Kang[2], Chul Soo Kim[3]

Abstract

The k-means algorithm is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. The k-modes algorithm is to extend the k-means paradigm to categorical domains. The algorithm requires a pre-setting or random selection of initial points (modes) of the clusters.

This paper improved the problem of k-modes algorithm, using the Max-Min method that is a kind of methods to decide initial values in k-means algorithm. we introduce new similarity measures to deal with using the categorical data for clustering.

We show that the mushroom data sets and soybean data sets tested with the proposed algorithm has shown a good performance for the two aspects(accuracy, run time).

Keywords: Categorical Data, k-means Algorithm, k-modes Algorithm, Max-Min Method

1. 서론

k-means 알고리즘(MacQueen(1967))은 대용량 데이터 처리에 효율적(Anderberg (1973))이기 때문에 데이터 마이닝에 적합하다. 그러나 수치형 데이터(numerical data)만 적용 가능하며, 초기값과 군집의 수에 따라 군집 결과에 상당한 영향을 주게 된다.

대용량 데이터를 다루는 데이터 마이닝은 범주형 데이터(categorical data)를 포함하는 경우가 있다. Ralambondrainy(Ralambondrainy(1995))는 범주형 데이터를 포함하는

1) (690-756) 제주특별자치도 제주시 제주대학로 66, 제주대학교 전산통계학과 석사
E-mail: wwwmd-00@nate.com
2) (690-756) 제주특별자치도 제주시 제주대학로 66, 제주대학교 전산통계학과 박사수료
E-mail: hchkang@cheju.ac.kr
3) 교신저자 (690-756) 제주특별자치도 제주시 제주대학로 66, 제주대학교 전산통계학과 교수
E-mail: cskim@cheju.ac.kr

대용량 데이터를 군집화 하기위해 k-means 알고리즘을 이용하였다. 이 알고리즘은 데이터 집합(data sets)이 많은 범주와 속성을 포함할 때 2진 값으로 바꾸어야 하는 계산 비용 및 기억 장소를 증가시키게 되고, 범주의 속성에 대해 부여되는 0과 1을 이용한 수치의 평균은 의미가 없다.

k-modes 알고리즘(Huang(1998))은 범주형 데이터를 대상으로 k-means 알고리즘의 형식을 유지하면서 비유사도(dissimilarity measure)를 이용하여 범주형 데이터에 적합하도록 제안한 방법이고, ROCK(Guha 외 2인(1999))은 객체간의 유사도를 정의한 후 데이터의 모든 객체를 동시에 비교하여 유사도가 가장 큰 객체들을 순차적으로 병합해 가는 계층적 방법이다.

앞서 설명된 k-means, k-modes는 비계층적 군집화 방법이므로 초기값과 군집 수에 영향을 받게 되고, 부적절한 초기값 결정은 잘못된 군집의 생성과 군집 생성 과정에서 많은 반복이 발생하여 군집 생성에 많은 시간이 소요되므로 군집분석 성능에 상당한 영향을 주게 된다. 본 논문에서는 범주형 데이터를 군집화 하기 위한 k-modes 알고리즘의 초기값 결정 문제를 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 k-modes 알고리즘과 초기값 결정 방법을 살펴보고, 3장에서는 제안한 알고리즘을 기술하고, 4장에서는 small soybean 데이터와 mushroom 데이터를 이용하여 제안한 알고리즘에 대한 성능 평가를 하였으며, 5장에서는 결론에 대해 기술하였다.

2. k-modes 알고리즘과 초기값 결정 방법

2.1 k-modes 알고리즘

데이터 집합 $X=\{x_1,x_2,\cdots,x_n\}$는 n개의 객체로 구성되어 있고, 각 객체는 m개의 범주형 변수 값을 갖는다고 하자.

$$x_i=(x_{i1},x_{i2},\cdots,x_{im})',\ i=1,2,\cdots,n$$

j번째 범주형 변수 A_j는 l_j개의 수준을 가지며 각 수준을 $c_{j1},c_{j2},\cdots,c_{jl_j}$이라 하자.

범주형 변수의 값들로 구성된 두 객체 x_{i1}, x_{i2}의 비유사도를 객체간의 일치하지 않는 변수의 수인 $d(x_{i1},x_{i2})$로 정의한다.

$$d(x_{i1},x_{i2})=\sum_{j=1}^{m}\delta(x_{i_1j},x_{i_2j}) \tag{1}$$

여기서, $\delta(a,b)$는 두 값이 일치하지 않을 때 1의 값을 갖고, 그렇지 않으면, 0의 값을 갖는 지시함수이다.

$$\delta(a,b)=\begin{cases}0, & a=b\\ 1 & a\neq b\end{cases} \tag{2}$$

데이터 집합 X에 대응되는 mode $q^{*}=(q_1^{*},q_2^{*},\cdots,q_m^{*})'$는 집합 X내의 객체들과 비유사도가 가장 적은 벡터로 정의한다. 즉 mode q^{*}는 벡터 $q=(q_1,q_2,\cdots,q_m)'$ 중에서 비유사도의 합 $D=(q,X)$을 최소로 하는 벡터이다.

$$D(q,X)=\sum_{i=1}^{n}d(x_i,q) \tag{3}$$

데이터 집합 X에서 범주형 변수 A_j가 수준 c_{jk}를 갖는 빈도수를 n_{jk}라 하면, A_j가 c_{jk}를 가질 상대빈도는 다음과 같다.

$$fre(A_j=c_{jk}|X)=\frac{n_{jk}}{n}\quad,\ k=1,2,\cdots,l_j \tag{4}$$

모든 $j(=1,2,\cdots,m)$에 대하여

$$fre(A_j=q_j^{*}|X)\geq fre(A_j=c_{jk}|X) \tag{5}$$

을 만족하는 $q^{*}=(q_1^{*},q_2^{*},\cdots,q_m^{*})'$는 비유사도의 합 $D(q,X)$를 최소로 하므로, 데이터 집합 X의 mode가 된다. 변수별로 빈도가 가장 큰 범주 값들의 조합이 그 집합의 mode가 된다.

k-modes 알고리즘의 단계는 다음과 같다.

1. k개 군집의 초기 mode $\{q_1^0,q_2^0,...,q_k^0\}$를 선택한다.
2. 모든 객체에 대해 초기 mode $\{q_1^0,q_2^0,...,q_k^0\}$와의 비유사도를 계산한다. 비유사도가 가장 작은 군집으로 객체를 할당한 후, k개 군집내의 mode를 갱신하여 갱신된 첫 번째 mode $\{q_1^1,q_2^1,...,q_k^1\}$를 얻는다.
3. 모든 객체와 갱신된 mode의 비유사도를 계산하다. 만일 다른 군집내의 객체와 mode와의 비유사도가 더 작으면 해당 객체를 그 군집으로 다시 할당하고 군집내의 mode를 갱신한다.
4. 단계 3을 변화가 없을 때까지 반복 실행한다.

2.2 초기값 결정 방법

k-means 알고리즘에서 초기값 결정은 군집 수와 군집생성에 걸리는 시간에 중요한 요인이다. MA(Macqueen Approach) 방법(MacQueen(1967))은 데이터에서 임의로 k개의 초기값을 선택하고 나머지 객체들은 초기값에 가까운 군집으로 포함시킨 후, 군집 중심을 다시 계산하여 군집 중심의 변화량을 임계값(threshold) 이하가 될 때까지 반복하여 군집을 생성한다. 이 방법은 초기값 선택이 쉽고, 편리하게 사용할 수 있으나 부적절한 초기값이 선택될 때는 잘못된 군집을 생성할 수 있다.

KA(Kaufman Approach) 방법(Kaufman and Rousseeuw(1990))은 자료의 가장 중앙에 위치한 값을 첫 번째 초기값으로 선택하고, 나머지 모든 관측값에 대해서 초기값과 관측값과의 거리를 계산한다. 나머지 초기값은 첫 번째 초기값과 일정한 거리 이상 떨어져 있도록 군집을 생성한다. 이 방법은 MA 방법보다는 정교한 반면, 초기값을 구한 후 다음 단계의 초기값을 구하는 과정에서 주변의 모든 값들을 고려하기 때문에 대용량 데이터에는 적합하지 않다.

Max-Min 방법(Bae and Roh(2005))은 단계적으로 초기값을 선택하되 선택된 초기값들이 다음 초기값을 결정하는데 도움을 주며 많은 계산을 하지 않도록 고안됐다.

Max-Min 방법은 자료에서 랜덤하게 하나의 관측값을 선택하여 첫 번째 초기값으로 선택하고, 첫 번째 초기값에서 나머지 관측값과의 거리를 구하여 그 거리를 최대로 하는 관측값을 두 번째 초기값으로 선택한다. 다음 단계의 초기값을 구하기 위해서 초기값에 선택되지 않은 나머지 관측값들에 대하여 첫 번째 초기값과의 거리와 두 번째 초기값과의 거리를 구하면 각 관측값에 대해 두 종류의 계산된 거리가 얻어진다. 이 두 거리 중, 최소값을 선택하여 각 관측값에 대해 두 초기값과의 거리로 대응하게 한다. 각 관측값에 대응되어 있는 관측값과 두 초기값과의 거리를 비교하여 이 값을 최대로 하는 관측값을 구하여 세 번째 초기값으로 선택함으로써 초기값들이 적절하게 떨어져 선택되도록 하였다. 다음 단계의 초기값을 선택하기 위해서는 이전 단계까지 초기값으로 선택되지 않은 관측값에 대하여 이 과정을 반복적으로 시행하여 k개의 초기값이 모두 선택될 때까지 계속 실시한다.

3. 제안 알고리즘

본 논문에서는 2장에서 언급한 Max-Min 방법을 k-modes 알고리즘에 사용할 수 있도록 변형하였고, 유사도를 이용하여 k-modes 알고리즘의 효율을 높일 수 있는 방안을 제시한다.

3.1 유사도

Max-Min 방법을 k-modes 알고리즘에 적용하기 위해 두 객체 (x_i, x_j)의 유사도를 다음과 같이 정의한다.

데이터 집합 $X=\{x_1, x_2, \cdots, x_n\}$는 n개의 객체로 구성되어 있고, 각 객체는 m개의 범주형 변수 값을 갖는다고 하자.

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{im})', \ i = 1, 2, ..., n$$

j번째 범주형 변수 A_j는 l_j개의 수준을 가지며 각 수준을 $c_{j1}, c_{j2}, \cdots, c_{jl_j}$이라고 하자.

유사도를 다음과 같이 정의한다.

$$d(x_{i_1}, x_{i_2}) = \frac{\sum_{j=1}^{m} \delta(x_{i_1 j}, x_{i_2 j})}{\sum_{j=1}^{m} \delta(x_{i_1 j}, x_{i_2 j}) + w} \tag{6}$$

여기서, $\delta(x_i, y_i)$는 (2)와 반대로 두 값이 일치하지 않을 때 0의 값을 갖고, 그렇지 않으면, 1의 값을 갖는 지시함수이다.

가중치 w는 $\delta = 0$일 때, 즉 동일한 변수에 대한 각 객체의 값이 일치하지 않을 때 그 변수의 (1/수준 수)들의 합이다. 즉 2개의 수준을 갖는 변수보다 3개의 수준을 갖는 변수가 서로 다를 가능성이 높기 때문에 이를 유사도에 포함하였다.

$$w = 2 \times \sum_{j=1}^{m} \frac{1}{|l_j|} \tag{7}$$

기존 k-modes 알고리즘에서는 수준 수가 다른 변수를 수준 수가 동일한 변수로 취급하여 비유사도를 계산하였기 때문에 이 부분을 개선하였다.

유사도 (6)은 초기 mode 결정뿐만 아니라 군집에 할당하기 위하여 mode(또는 초기 mode)와 비교할 때도 사용된다.

3.2 유사도를 이용한 초기 mode 결정

범주형 데이터를 군집화 하기 위하여 초기 mode를 결정한다. 초기 mode 결정을 위해 Max-Min 방법에 유사도를 적용한 알고리즘은 다음과 같다.

1. 대용량 데이터일 경우 데이터의 10%를 표본으로 추출하여 객체간의 유사도를 구한다.
   a. 소용량 데이터일 경우 모든 객체간 유사도를 구한다.
2. 유사도의 분산이 가장 큰 객체를 첫 번째 초기 mode로 결정하고 q_1^0와 유사도가 가장 낮은 객체를 두 번째 초기 mode q_2^0로 결정한다.
3. 초기 mode q_m^0, $m=3,4,\cdots,k$를 구하기 위해서 이미 구해진 초기 mode를 추가하면서 초기 mode를 제외한 나머지 객체 $X_i(X_i \neq q_l^0, l=1,2,\cdots,m-1)$에 대하여 다음을 계산한다.
   $$X_i \leftarrow D_i = max\{d(X_i,q_1^0), d(X_i,q_2^0),\cdots,d(X_i,q_{m-1}^0)\},\ i=1,2,...,n,$$
4 각 객체 X_i에 대해서 얻어진 D_i를 비교하여 이들의 값을 최소로 하는 객체를 다음 초기값으로 선택한다.
   $$q_m^0 = X_q \leftarrow min_{1 \le j \le n}(D_j) = D_q$$
5. 각 객체별로 mode와 유사도를 계산하여 가장 유사한 군집에 객체를 할당한다.
6. 모든 객체들에 대해서 군집으로 할당이 끝나면 mode를 갱신한다.
7. 단계 5를 변화가 없을 때까지 반복 실행한다.

4. 성능평가

본 논문에서 성능평가를 위해 군집화 실험에 빈번하게 사용되는 UCI Machine Learning Repository의 small soybean 데이터와 mushroom 데이터를 사용하였다.

small soybean 데이터는 47개의 객체를 가지고 있으며 각각의 객체는 35개의 범주 속성으로 이루어져 있다. 각각의 속성들은 월(date), 잎 모양, 줄기 상태, 크기 등으로 구성되어 있다.

제안한 알고리즘이 대용량 데이터에 적합한지를 알아보기 위하여 mushroom 데이터를 사용하였다. mushroom 데이터는 총 8,124개로 구성되어 있으며, 그중 4,208개는 식용 버섯을 나타내고, 3,916개는 독성 버섯을 나타낸다. 데이터들은 버섯의 물리적 특성인 색, 크기, 냄새, 모양 등 22개의 속성을 가지고 있다.

실험에 좀 더 적합하게 하기 위하여 실험 데이터 속성의 성분을 수치화하였고, 결측은 0으로 표시하여 실험에 포함하였다.

4.1 성능평가

제안한 알고리즘의 성능을 평가하기 위하여 기존의 k-modes 알고리즘과 정밀도 및 수행 시간을 비교하였다.

1) small soybean 데이터

small soybean 데이터는 총 객체 수가 47개로 소용량 데이터이므로 표본을 추출하지 않고 모든 객체간의 유사도를 계산하여 초기 mode를 선택하였다.

<표 1> soybean 데이터에 대한 제안 알고리즘 수행결과

	q_1	q_2	q_3	q_4
1		10		
2			10	
3	10			
4				17

<표 1>은 small soybean 데이터에 대한 제안 알고리즘 수행결과로 정밀도 100%로 나타났다. small soybean 데이터는 객체 수가 적기 때문에 모든 객체에 대한 유사도를 계산한 다음 초기 mode를 한 번에 선택할 수 있었고, 각 군집마다 1개의 초기 mode가 결정되었다. 초기 mode는 11번, 21번, 5번, 46번이 선택되었고, 1번부터 10번까지 1번 군집, 11번부터 20번까지 2번 군집, 21번부터 30번까지 3번 군집, 31번부터 46번까지 4번 군집으로 생성되었다.

<표 2>는 small soybean 데이터에 대해 k-modes 알고리즘과 제안 알고리즘의 수행시간을 비교한 것이다(100회 반복).

<표 2> k-modes 알고리즘과 제안 알고리즘의 수행시간

	Mean	Worst	Best
k-modes	7.74sec	8.60sec	4.97sec
제안	7.55sec		

2) mushroom 데이터

mushroom 데이터에서 1,000개를 랜덤하게 추출하여 실험 데이터를 만들었다. 1,000개 중 521개는 식용 버섯이고, 나머지 479개는 독성 버섯이다. 이 데이터를 이용하여 10%(100개) 표본을 다시 추출하여 실험하였다.

<표 3> k-modes 알고리즘과 제안 알고리즘의 정밀도

정밀도	90% 초과	90-80%	80-70%	70-60%	60% 미만	평균	표준편차
k-modes	0회	25회	11회	33회	31회	69.47%	11.78%
제안	4회	18회	23회	31회	24회	71.02%	11.37%

<표 3>에서 k-modes 알고리즘은 80%이상의 정밀도를 보인 횟수가 25회로 비교적 높게 나타났으나, 70%미만의 정밀도를 보인 횟수도 64회 나타났다. 제안한 알고리즘은 70%이상의 정밀도를 나타낸 횟수가 45회이고 또한 정밀도가 90%가 넘는 경우도 4회가 나타났다. k-modes 평균 정밀도는 평균 69.47%이고, 제안한 알고리즘은 평균 71.02%의 정밀도를 보였다.

<표 4>는 mushroom 데이터에 대해 k-modes 알고리즘과 제안 알고리즘의 수행시간을 비교한 것이다(100회 반복).

<표 4> k-modes 알고리즘과 제안 알고리즘의 수행시간

	Mean	Worst	Best
k-modes	284.20sec	735.69sec	132.11sec
제안	254.52sec	616.22sec	142.50sec

5. 결론

본 논문에서 제안한 알고리즘은 범주형 데이터 군집화에 필요한 초기값 결정을 위해 Max-Min 방법을 수정하였다. 데이터 크기가 커질수록 많은 객체간의 유사도를 계산해야 하기 때문에 수행속도 측면에서 성능이 떨어질 수 있지만, k-modes 알고리즘에 비해 평균 수행시간 및 정밀도가 다소 향상되었음을 실험결과 알 수 있었다.

참고 문헌

1. Anderberg, M. R.(1973). *Cluster Analysis for Applications. Academic Press.*
2. Bae W, S and Roh S, W.(2005). *A Study on K-Means Clustering, The Korean Communications in Statistics Vol. 12 No. 2, pp. 497-508.*
3. Guha, S. Rastogi, R. and Shim, K. (1999). *ROCK: A robust clustering algorithm for categorical attributes. Proceedings of the IEEE International Conference on Data Engineering, Sydney.*
4. Huang, Z.(1998). *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283-304.*
5. Kaufman L and Rousseeuw P.J.(1990). *Finding Groups in Data. An Introduction to Cluster Analysis.*
6. MacQueen, J.B.(1967). *Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium Mathematical Statistics and Probability, pp. 281-297.*
7. Ralambondrainy, H.(1995). *A Conceptual Version of the k-Means Algorithm. Pattern Recognition Letters. 16, pp. 1147-1157.*

[2007년 4월 접수, 2007년 5월 채택]