Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

K-모드 알고리즘과 ROCK 알고리즘의 비교 및 개선방안

김 보 화[1] · 김 규 성[2]

요 약

데이터 마이닝에서 분석의 대상으로 하는 대용량 자료에는 연속형 자료와 범주형 자료가 모두 포함된다. 전통적인 군집분석은 연속형 자료를 대상으로 하는 방법들이다. 본 연구에서는 범주형 자료를 대상으로 하는 군집분석방법인 K-모드 알고리즘과 락(ROCK) 알고리즘을 비교·분석하였다. 그리고 두 알고리즘이 갖는 방법론적인 단점을 보안하여 군집의 효과를 높일 수 있는 개선 방안을 제안하였다.

주요용어 : 군집분석, 대용량 자료, 범주형 자료, 사후 할당.

1. 서론

군집분석은 객체들을 몇 개의 군집으로 나누는데, 군집내의 객체들은 비슷한 성질을 갖도록 하고 다른 군집의 객체들과는 상이한 속성을 갖도록 하는 기법으로, 크게 계층적 군집방법(hierarchical clustering)과 최적분리 군집방법(partitional clustering)으로 나눌 수 있다. 계층적 군집방법은 가까운 객체들끼리 묶어감으로써 군집을 만들어 가는 병합적(agglomerative) 방법과 반대로 먼 객체들을 나누어 가는 분할적(divisive) 방법으로 나눌 수 있으며, 계층적 군집 방법은 어떤 객체가 일단 다른 군집에 속하면 다시는 같은 군집에 속하지 못하는 성질을 가지고 있다. 최적분리 군집방법은 계층적으로 군집을 형성하는 것이 아니라 객체들을 몇 개의 군집으로 구분시키는 형태로, 어떤 기준 함수(criterion function)를 최적화하는 군집을 찾는 기법이다. 계층적 군집방법이 초기에 부적절한 병합(또는 분리)이 일어났을 때 회복할 수 없는 반면, 최적분리 군집방법은 군집을 형성하는 과정에서 객체들을 재 할당할 수 있다.

대용량 자료를 다루는 데이터 마이닝에서 군집분석은 객체들을 유사한 속성을 갖는 군집으로 분할하여 그 속에서 의미 있는 정보를 발견하는 것이다. 전통적으로 널리 이용 되어온 방법인 K-평균(means) 알고리즘은 수렴한다는 사실이 알려져 있다(MacQueen 1967, Bezdek 1980, Selim and Ismail 1984). 또한 대용량 자료를 다루는데 효과적이기 때문에 데이터 마이닝에서 군집분석 방법으로 활용하기에 알맞다. 그러나 K-평균 알고리즘은 객체간의 비유사성을 유클리디안 거리로 정의하고 객체의 평균을 계산하여 비용함수를 최소화시키기 때문에 연속형 자료에 대해서만 적용 가능하다는 한계가 있다. 통상적으로 데이터 마이닝에서 다루는 대용량 자료에는 연속형 자료는 물론 범주형 자료도 포함되어 있기 때문에 K-평균 알고리즘만으로는 대용량 자료의 군집분석을 하기에는 어려움이 있다.

대용량 범주형 자료에 대한 군집분석에 대한 연구로는 K-모드(modes) 알고리즘(Huang,

1) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 컴퓨터·통계학과 대학원.
2) (130-743) 서울특별시 동대문구 전농동 90, 서울시립대학교 컴퓨터·통계학과, 부교수
E-mail : kskim@uoscc.uos.ac.kr

1997)과 락 알고리즘(RObust Clustering using linKs, ROCK, Sudipto 등, 1999) 등이 대표적으로 알려져 있다. 본 연구의 제 2장에서는 K-모드 알고리즘과 락 알고리즘을 비교·분석하고, 제 3장에서는 두 알고리즘의 단점을 보완하는 개선방안을 제안한다. 마지막으로 제 4장에서는 간단한 요약과 향후 과제를 언급한다.

2. K-모드 알고리즘과 ROCK 알고리즘 비교

2.1 K-모드 알고리즘

K-모드 알고리즘은 K-평균 알고리즘의 기본 구조를 유지하면서 대용량 범주형 자료를 처리할 수 있도록 제안된 방법이다(Huang 1997). K-평균 알고리즘은 크게 네 단계로 나타낼 수 있다. 첫째, k개 군집의 평균에 대한 초기값을 정하고 둘째, 각 객체와 k개 군집의 초기 평균과의 비유사성을 계산하며 셋째, 각 객체를 가장 비유사성이 적은 군집으로 할당한다. 마지막으로 k개 군집에 대해서 군집내의 비유사성이 최소가 되도록 군집의 평균을 갱신한다. 두 번째부터 네 번째까지의 절차를 알고리즘이 수렴할 때까지 반복 수행한다. K-모드 알고리즘은 K-평균 알고리즘에서 세 가지를 변경하는데, 군집의 평균 대신 모드의 개념을 사용하고, 범주형 속성을 갖는 객체간의 비유사성을 일치하지 않는 속성의 개수로 정의하며, 또한 군집에서 속성별로 빈도가 가장 큰 범주 값을 사용하여 모드를 갱신하는 것이다.

K-모드 알고리즘의 각 단계를 정리하면 아래와 같다.

- 단계 1. k개 군집의 초기 모드를 선택한다.
- 단계 2. 각 객체를 비유사성이 가장 적은 군집으로 할당한다. 여기서 비유사성은 두 객체간에 일치하지 않는 속성의 개수로 정의한다.
- 단계 3. k개 군집에 대해 속성별로 빈도가 가장 큰 범주 값으로 모드를 갱신한다.
- 단계 4. 모든 객체에 대해 갱신된 k개 모드와 비유사성을 구해서 비유사성이 가장 적은 군집으로 객체를 재 할당한다.

단계 3과 4를 반복하여 알고리즘이 수렴할 때까지 실행한다.

K-모드 알고리즘은 절차가 간단하고, 범주형 자료를 포함하는 대용량 자료에 적용하기에 적합하다는 장점이 있다. 반면 알고리즘의 첫 단계에서 초기 모드를 어떻게 정하는가에 따라 군집의 결과가 상당히 달라지기 때문에 초기 모드 설정에 대단히 민감하다는 단점이 있다.

Huang(1997)은 초기 모드 선택 방법으로 두 가지를 제안하였다. 첫 번째는 자료로부터 임의로 서로 다른 k개의 레코드를 뽑아 초기 모드로 사용하는 것이고, 두 번째는 자료의 속성별로 범주의 빈도를 구해서 k개의 초기 모드에 빈도가 큰 범주가 골고루 반영되도록 하는 것이다.

2.2 ROCK 알고리즘

Sudipto 등(1999)에 의해 제안된 락 알고리즘은 링크(link)라는 개념을 이용하여 객체들을 병합해 나가는 방법이다. 락 알고리즘의 절차는 다음과 같다.

우선 범주형 자료를 부울(boolean)형 자료로 변환하는데, 모든 범주형 속성의 각 범주를 하나의 변수로 생각해서 속성의 값이 그 범주에 해당하면 1을, 그렇지 않으면 0을 값으로 갖도록 한다. 변환된 자료에서 두 객체 T_1과 T_2의 유사성은 자카드 계수(Jaccard coefficient)를 사용하여 (2.1)식과 같이 정의한다.

$$\mathrm{sim}(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \tag{2.1}$$

여기서 $|T_i|$는 T_i에 속하는 것의 개수이다. 부울형으로 변환된 각 속성에 대해서 두 객체가 공통으로 1을 값으로 갖는 변수가 많을수록 즉, $|T_1 \cap T_2|$ 값이 클수록 두 객체는 유사한 것이다. $|T_1 \cup T_2|$ 값으로 나누어주는 것은 두 객체간의 유사성이 0과 1사이의 값을 가지도록 하기 위해서이다.

두 객체간의 유사성이 어떤 기준값(threshold) θ 보다 크면 두 객체는 이웃(neighbor)이 되는데, 이 때 기준값 θ는 자료의 특성에 따라 정해 주어야 한다. 모든 두 객체들에 대해서 유사성을 구하고 유사성이 기준값 θ 보다 크면 이웃군(neighborhood)을 형성한다. 링크(link)는 두 객체간의 공통 이웃의 개수로 정의하는데, 링크의 값이 클수록 두 객체는 유사한 것이다. 이와 같은 링크의 개념을 사용하는 것은 객체간의 유사성을 판단할 때 자료의 모든 객체를 고려하는 것이기 때문에 폭넓은(global) 접근방법이라 할 수 있다.

자료의 모든 두 객체간의 링크를 계산한 다음, 어떤 군집들을 먼저 병합할 것인가를 판단해야 한다. 락 알고리즘은 처음에 모든 객체를 하나의 군집으로 생각하고 시작한다. 두 개의 군집 C_i와 C_j에 대해서 $\mathrm{link}[C_i, C_j]$는 C_i와 C_j사이의 교차 링크(cross link)의 개수로 식 (2.2)와 같다.

$$\mathrm{link}[C_i, C_j] = \sum_{p_q \in C_i, p_r \in C_j} \mathrm{link}(p_q, p_r) \tag{2.2}$$

군집 C_i와 C_j를 병합할 것인가에 대한 판단기준(goodness measure)은 아래와 같이 정의된다.

$$g(C_i, C_j) = \frac{\mathrm{link}[C_i, C_j]}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \tag{2.3}$$

여기서 n_i, n_j는 각각 군집 C_i와 C_j에 속하는 객체들의 개수이다. (2.3)식에서 $f(\theta)$를 결정하는 것은 선택사항인데, Sudipto 등(1999)은 모의실험에서 $f(\theta) = (1-\theta)/(1+\theta)$사용하였다. 군집을 형성하는 각 단계에서 (2.3)식이 최대가 되도록 하는 두 개의 군집이 병합된다.

락 알고리즘에서는 군집을 형성할 때 두 객체간 유사성을 링크라는 개념을 사용하여 정의하는데, 이것은 자료의 모든 객체를 고려하는 방법이므로 폭넓은 접근 방법이라 할 수 있다. 락 알고리즘은 처음에 모든 객체를 하나의 군집으로 보고 유사성이 가장 큰 군집들을 하나의 군집으로 병합해 나가므로 유사성이 큰 객체들을 순차적으로 군집화할 수 있는 장점이 있는 반면, 객체내의 모든 변수들을 동시에 고려하므로 알고리즘이 수렴하였는데도 군집을 형성하지 못하고 남아있는 객체가 있을 수 있다는 단점이 있다.

2.3 K-모드 알고리즘과 ROCK 알고리즘의 비교

대용량 범주형 자료에 대한 군집분석 방법으로 제안된 K-모드 알고리즘은 군집의 개수를 미리 정해서 군집 분석하는 최적분리 군집방법이고, 락 알고리즘은 가장 유사성이 큰 군집들을

단계적으로 병합하는 계층적 군집방법으로 군집을 형성해 나가는 과정이 서로 상이하다.

객체간의 비유사성을 정의하는 방법으로 K-모드 알고리즘에서는 범주형 속성에서 일치하지 않는 것의 개수로 비유사성을 정의하고 있는 반면, 락 알고리즘은 범주형 자료를 각 범주값이 하나의 변수가 되게 부울형 자료로 변환한 후 두 객체간의 교집합의 개수를 합집합의 개수로 나눈 값이 기준값 보다 크면 서로 이웃이 되고, 두 객체간의 공통 이웃의 개수를 링크라고 하여 이를 두 객체간의 유사성으로 정의하고 있다.

다음으로 두 알고리즘의 장·단점을 비교해보자. K-모드 알고리즘은 절차가 간단하고, 대용량 자료에서 계층적 군집방법에 비해 수렴 속도가 빠르다는 장점이 있다(Murtagh 1992, Huang 1997). 반면 군집분석 전에 미리 군집의 개수를 정해야 하고, 알고리즘의 첫 번째 단계에서 초기 모드를 어떻게 정하는가에 따라 군집의 결과가 굉장히 민감하다는 단점이 있다. 이에 비해 락 알고리즘은 두 객체간의 유사성을 정의할 때 링크라는 개념을 이용하여 전체 객체를 고려하는 폭넓게 접근방법을 사용하고 있고, 어떤 군집을 병합할 것인가를 판단할 때 군집의 크기까지 고려하기 때문에 상대적으로 크기가 작은 군집도 제대로 유지할 수 있다는 장점이 있다. 반면 락 알고리즘은 범주형 자료를 부울형 자료로 변환해서 사용하는 과정상의 번거로움이 있고, 알고리즘이 수렴한 후에도 군집을 형성하지 못하고 남아 있는 객체가 있을 수 있다는 단점이 있다.

3. K-모드 알고리즘과 ROCK 알고리즘의 개선방안

3.1 K-모드 알고리즘의 개선방안

K-모드 알고리즘의 가장 큰 단점은 초기 모드 설정에 따라 군집의 결과가 민감하게 반응한다는 것이다. 따라서 효과적인 초기 모드 설정방법을 이용하면 K-모드 알고리즘의 효율을 높일 수 있는 여지가 있다. 본 소절에서는 K-모드 알고리즘의 효율을 높일 수 있는 개선 방안을 제안한다.

군집 분석에서 중요하게 다루어지는 변수를 종속변수로 하고 나머지 다른 변수를 설명변수로 하여 군집형성에 영향을 주는 정도에 따라 설명변수에 우선순위를 준다. 그리고 자료를 우선순위가 높은 변수 순으로 다중 정렬을 한다. 여기에서 정렬된 자료의 수를 n이라 하고, 군집의 개수를 k라 하자. 그러면 1과 n/k 사이에서 하나의 난수를 발생하여, 이를 m이라 하자. 자료의 m번째 레코드를 첫 번째 초기 모드로 한다. 이 후 순차적으로 $(m+(j-1)\times n/k$ $j=2,\ldots,)k$번째의 레코드를 초기 모드로 한다. 만일 $(j-1)\times n/k$값이 정수가 아니면 반올림하여 사용한다.

3.2 ROCK 알고리즘의 개선방안

락 알고리즘은 알고리즘이 수렴하였는데도 군집을 형성하지 못하고 남아 있는 객체가 있을 수 있는 점이 가장 큰 단점이다. 군집을 형성하지 못한 객체들은 다른 객체들에 비해 유사성이 적은 것들이며 락 알고리즘은 유사성이 큰 객체들을 우선적으로 묶어가기 때문에 알고리즘을 실행해 나가다 보면 유사성이 적은 객체들은 뒤에 남겨지게 된다. 그리고 앞에서 정의한 링크

라는 개념을 이용하면 이미 묶여진 군집에 포함되기 어려운 상황이 발생할 수 있다. 따라서 이렇게 남겨진 객체들을 이미 만들어진 군집에 사후 할당하는 방법을 생각할 수 있다.

사후 할당을 위해 각 군집의 모든 범주형 속성에 대해 빈도가 가장 큰 값으로 그 군집의 대표치를 만든다. 그런 다음 각 군집의 대표치와 군집을 형성하지 못하고 남아 있는 객체와의 유사성을 구해서 유사성이 가장 큰 군집으로 객체를 할당한다. 이러한 사후 할당 방법을 락 알고리즘에 추가하면 군집분석이 완료되었을 때 군집을 형성하지 못하고 남아있는 객체는 없어질 것이다.

4. 요약 및 향후 과제

전통적인 군집분석은 연속형 자료에 대해서 연구되어왔다. 최근의 대용량 자료에는 범주형 자료도 포함되어 있으므로 본 연구에서는 대용량 범주형 자료에 대한 군집분석 방법인 K-모드 알고리즘과 락 알고리즘에 대해 비교·분석해 보았다. 또한 이 두 알고리즘의 단점을 보안하기 위한 개선방안을 제안하였다.

향후 실제 자료에 K-모드 알고리즘과 락 알고리즘을 적용하여 두 알고리즘의 효율을 비교하고, 제안된 개선방법이 군집의 결과를 얼마나 향상시키는지를 연구해보고자 한다.

참 고 문 헌

[1] Bezdek, J.C. (1980). A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2(8), 1-8.

[2] Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values. Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining. World Scientific.

[3] Huang, Z. (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. Workshop on Research Issues on Data Mining and Knowledge Discovery.

[4] MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 281-297.

[5] Murtagh, F. (1992). Comments on "Parallel Algorithms for Hierarchical Clustering and Cluster Validity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(10), 1056-1057.

[6] Selim, S. Z. and Ismail, M. A. (1984). K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(1), 81-87.

[7] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim (1997). A Clustering algorithm for categorical attributes. Technical report, Bell Laboratories, Murray Hill.

[8] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim (1999). Rock: A robust clustering algorithm for categorical attributes. Proceedings of the IEEE International Conference on Data Engineering, Sydney.