Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 3, pp. 549~556

More Efficient k-Modes Clustering Algorithm

Dae-Won Kim[1)] · Yigeun Chae[2)]

Abstract

A hard-type centroids in the conventional clustering algorithm such as k-modes algorithm cannot keep the uncertainty inherently in data sets as long as possible before actual clustering(decision) are made. Therefore, we propose the k-populations algorithm to extend clustering ability and to keep the data characteristics. This k-population algorithm as found to give markedly better clustering results through various experiments.

Keywords : Categorical data analysis, Clustering, Fuzzy k-means algorithm

1. 머리말

군집화 알고리즘이 점차 대량의 데이터를 분류하는 기법으로 많이 이용되고 있다. 특히 데이터마이닝이나 생물정보학 분야처럼 수치 혹은 비수치의 속성을 가진 대량의 데이터를 다루기 위하여 다양한 군집화 알고리즘이 사용되고 있다. 꾸준히 군집화 알고리즘이 제안되고 있는데 예를 들자면 Gowda와 Diday(1991)의 유사도계수를 이용한 계층적 군집화 등이 있다. Huang과 Ng(1999)는 이러한 알고리즘이 범주형 데이터를 포함하는 대량의 데이터에 적용할 경우 비효율성이 발생 할 수 있음을 보였다. 그리하여 Huang(1998)은 최근 표준 k-means 알고리즘을 확장한 k-modes 알고리즘을 제안하고 간결한 비유사도 측정을 통하여 군집화 데이터를 분류하고 발현 도수를 기반하여 군집화의 센트로이드를 결정하였다. Huang의 확장된 방법은 실제 데이터의 처리에 상당한 효율성을 가지고 있다. 이와 더불어 Huang은 k-modes 알고리즘을 일반화한 퍼지 k-modes 알고리즘을 개발하여 서로 다른 군집에 대하여 소속도(membership degree)를 적용하였다.

k-modes 알고리즘이 비록 실제 데이터를 효율적으로 군집화 하고, 결정적 센트로이

1) 대전광역시 유성구 구성동 한국과학기술원 바이오시스템학과 박사후연구원
E-mail : dwkim@if.kaist.ac.kr
2) 제1저자, 충남 공주시 옥룡동 공주대학교 교수
E-mail : ygchae@kongju.ac.kr

드와 간결한 거리도(distance measure)를 적용하여 보다 정교한 군집화가 가능하도록 하였지만 최종 군집화된 데이터는 오류를 포함하고 있다. 본 연구에서는 이러한 문제점을 해결하기 위하여 k-population 알고리즘을 제안하였으며 데이터의 특성이 센트로이드를 계산할 때 반영되도록 하여 오류를 줄였다. 여기서 population이란 센트로이드의 불확실성을 최소화 하고 군집화 과정에서 센트로이드를 보다 정교하게 표시하는 개념이다. 본 알고리즘에서는 데이터가 가지는 특성을 순차적인 갱신에서 센트로이드를 정할 때 반영하고, 최종 군집이 완성될 때까지 그 특성을 유지하도록 하여 다른 군집화 알고리즘에 비하여 국소적 극소화(local minima)의 오류를 회피하도록 하였다. 제안된 방법은 다양한 데이터를 대상으로 실험을 통하여 그 성능을 입증하였다.

2. 기존 k-modes 군집화 알고리즘

$X=\{x_1, x_2, \cdots, x_n\}$을 n개의 범주형 데이터이고 $x_j(1\le j\le n)$를 군집 속성 A=$\{A_1, A_2, \cdots, A_p\}$에 대한 범주속성이라고 정의한다. 각 범주속성 $A_l(1\le l\le p)$는 도메인 값들이고 이를 $DOM(A_l)=\{a_l^{(1)}, a_l^{(2)}, \cdots, a_l^{(n_l)}\}$으로 표시한다. n_l은 범주속성 A_l의 범주값(category value)들의 개수를 의미한다. 이에 따라 x_j는 $[x_{j,1}, x_{j,2}, \cdots, x_{j,p}]$로 나타낸다. 따라서 x_j는 속성값들과 결합하여 논리적으로 다음과 같이 나타낼 수 있다.

$$[A_l=x_{j,1}]\wedge[A_j=x_{j,2}]\wedge\cdots\wedge[A_p=x_{j,p}],$$

여기서 $x_{j,l}\in DOM(A_l)$ 이며 $1\le l\le p$이다.

k-modes 알고리즘은 범주형 데이터 X를 k개의 군집으로 군집화하기 위하여 아래와 같은 함수의 최소화하는 기법을 사용한다.

$$J_m(V:X)=\sum_{i=1}^{k}\sum_{j=1}^{n}(\mu_{i,j})^m d_c(v_i, x_j). \tag{2.1}$$

여기서 $\mu_{i,j}$는 k-modes 알고리즘에서 x_j가 i번째 군집에 소속되었을 때 $\mu_{i,j}=1$이고, 다른 경우에는 $\mu_{i,j}=0$이다. 반면 퍼지 k-modes 알고리즘에서는 $\mu_{i,j}$는 소속도를 나타낸다. 이때 군집의 센트로이드를 $V=(v_1, v_2, \cdots, v_k)$라고 하면, 각 범주형 군집에서 센트로이드 v_i는 p개의 후보로써 $[v_{i,1}, v_{i,2}, \cdots, v_{i,p}]$로 나타낼 수 있다. (2.1)식에서 모수 m은 각 데이터의 소속을 제어하기위한 양(positive)의 계수이다.

범주형 데이터를 군집화하기 위하여 k-modes 알고리즘에서는 군집의 센트로이드와 범주형 데이터의 거리를 측정하고 군집화의 각 단계에서 군집 센트로이드를 갱신한

다. k-modes 알고리즘에서 센트로이드 v_i와 범주형 데이터 x_j 사이의 거리값 $d_c(v_i, x_j)$는 다음과 같이 정의한다.

$$d_c(v_i, x_j) = \sum_{l=1}^{P} \delta(v_{i,l}, x_{j,l}). \tag{2.2}$$

여기서 $v_{i,l} = x_{j,l}$일 경우 $\delta(v_{i,l}, x_{j,l}) = 0$이 되며, $v_{i,l} \neq x_{j,l}$일 경우 $\delta(v_{i,l}, x_{j,l}) = 1$이 된다. i번째 mode를 i번째 군집의 센트로이드 $v_l = [v_{i,l}, \cdots, v_{i,p}]$라고 할 때 이 센트로이드의 갱신은 각 $v_{i,l} \in v_i (1 \leq l \leq p)$에 대하여 다음과 같이 갱신한다.

$$v_{i,l} = a_l^{(r)} \in DOM(A_l). \tag{2.3}$$

여기서 $a_l^{(r)}$은 다음과 같은 조건을 만족한다.

$$|\{\mu_{i,j} | x_{j,l} = a_l^{(r)}, \mu_{i,j} = 1\}| \geq |\{\mu_{i,j} | x_{j,l} = a_l^{(t)}, \mu_{i,j} = 1\}|, \ 1 \leq t \leq n_l. \tag{2.4}$$

퍼지 k-modes 알고리즘에서는 다음과 같은 조건을 사용한다.

$$\sum_{x_{j,l} = a_l^{(r)}} \mu_{i,j}^m \geq \sum_{x_{j,l} = a_l^{(t)}} \mu_{i,j}^m, \ 1 \leq t \leq n_l. \tag{2.5}$$

k-modes 알고리즘에서 군집 센트로이드 v_i에 대한 속성 $v_{i,l}$의 군집은 i번째 군집에 소속된 데이터의 집합에서 속성 A_l의 군집에 대한 빈도 형태로 결정된다. 이에 비하여 퍼지 k-modes 알고리즘에서는 센트로이드 $v_{i,l}$은 모든 군집에서 i번째 군집에 대한 $\mu_{i,j}$값의 합을 구하여 이중 가장 높은 값을 가지는 군집에서 센트로이드를 정하게 된다.

3. 확장된 알고리즘의 제안

3.1 k-population의 정의

퍼지 k-modes 알고리즘에서 군집속성에 대한 센트로이드는 소속도를 측정하여 결정하게 되는데 한 갱신 단계에서 정해진 센트로이드는 그 단계에서 군집을 대표하게 된다. 이러한 방법은 자칫 군집 속에서 다른 데이터들의 특성을 반영하지 못하고 2차, 3차 대표성을 가지는 데이터들의 특성을 무시하게 되어 최종 단계에서 잘못된 군집으

로 분류할 가능성을 내포하고 있다. 이는 중간 갱신단계에서 국소적 극소화 오류가 발생할 수 있음을 의미한다. 따라서 퍼지 k-modes 방법은 다음 갱신반복에서 현재 센트로이드가 가지는 성질을 반영할 수 없게 된다. 예를 들면 $DOM(A_l)=\{yes, no\}$ 라고 하고, 세 개의 데이터 x_1, x_2, x_3가 있고 각각의 소속도는 $\mu_{i1}=0.70$, $\mu_{i2}=0.80$, $\mu_{i3}=0.15$라고 할 때 각 데이터 포인트에 대한 l번째 속성값은 $x_{1,l}=yes$, $x_{2,l}=no$, $x_{3,l}=yes$ 결정된다.

i번째 군집의 센트로이드에 대한 l번째 속성을 $v_{i,l}$라고 할 때, (2.3)식과 (2.5)식에 의하여 $\sum_{x_{j,l}=yes} \mu_{ij}^{m} = 0.70^{m} + 0.15^{m}$ 그리고 $\sum_{x_{i,l}=no} \mu_{ij}^{m} = 0.80^{m}$의 계산에 의하여 $v_{i,l}$은 yes 혹은 no로 결정된다. 따라서 m=1.0일 때 $v_{i,l}$은 "yes"가 되며, m=2.0일 때 $v_{i,l}$은 "no"가 된다. 결정단계에서 둘 중에 하나는 다음 갱신반복에서 영향을 줄 수도 있지만 나머지는 폐기되어 버린다. 비록 폐기된 데이터의 소속도가 갱신 시 센트로이드를 정할 때 영향을 주도록 선택된 데이터보다 소속도가 근접하게 작다 하더라도 다음 갱신반복에서 어떠한 영향을 줄 수도 없게 된다.

이러한 방법은 데이터를 잘못 분류 하게 되며 국소적 극소화 오류가 발생할 가능성이 있다. *k*-modes 알고리즘에서도 군집의 속성 A_l의 분포를 나타내는 가장 높은 발생 빈도를 가지는 단일의 속성값 $a_l^{(r)}$이 퍼지 *k*-modes 알고리즘처럼 비슷한 문제점을 만든다. 이러한 문제를 해결하기 위하여 군집 속성들을 반영할 수 있는 군집 센트로이드를 결정하도록 한다. 또한 데이터 중에서 센트로이드로서 가지는 불확실성을 최종 갱신반복이 완성될 때까지 유지 하도록 한다. 본 연구에서는 이러한 기법이 가능하도록 population이라는 개념을 도입하였다.

군집 속에서 하나의 결정적인 데이터 포인트에 의하여 센트로이드가 결정되어 대표적 속성으로 정해지는 기존 방법에 비하여 제안하는 센트로이드의 속성은 군집의 분포 정보를 표현하기 위하여 범주형 값에 대한 population을 센트로이드로 정한다. $DOM(A_l)=\{a_l^{(1)}, a_l^{(2)}, \ldots, a_l^{(n_l)}\}$에 대하여 i번째 군집 센트로이드의 population은 다음과 같이 정의 한다.

$$v_i=[v_{i,1}, \cdots, v_{i,l}, \cdots, v_{i,p}],$$

여기서 $v_{i,l}=\{(a_l^{(t)}, \omega_l^{(t)}) | a_l^{(t)} \in DOM(A_l),\ 1 \le t \le n_l\}$ 이고,

다음 조건 $0 \le \omega_l^{(t)} \le 1$, $0 < \sum_{t=1}^{n_l} \omega_l^{(t)} < n$ 을 만족한다. (2.6)

따라서 $v_{i,l}$은 i번째 군집에 데이터가 소속될 때 속성 A_l의 군집 분포를 나타낸다.

$\omega_l^{(t)}$는 $a_l^{(t)}$가 $v_{i,l}$에 대한 신뢰도를 나타낸다.

3.2 거리도 계산 및 센트로이드 갱신

v_i와 x_j를 각각 i번째 군집과 데이터 포인트라고 할 때 군집에서는 $[v_{i,1}, v_{i,2}, \cdots, v_{i,p}]$와 $[x_{j,1}, x_{j,2}, \cdots, x_{j,p}]$로 나타낸다. v_i와 x_j의 거리 측정은 다음과 같이 정의 한다.

$$d_c(v_i, x_j) = \sum_{l=1}^{P} \delta(v_{i,l}, x_{j,l})$$

여기서 $\delta(v_{i,l}, x_{j,l}) = \frac{1}{\eta_i} \sum_{t=1}^{n_l} \tau(a_l^{(t)}, x_{j,l})$ 이고,

$$\tau(a_l^{(t)}, x_{j,l}) = \begin{cases} 0 & , a_l^{(t)} = x_{j,l} \\ \omega_l^{(t)} & , a_l^{(t)} \neq x_{j,l}, \end{cases} \text{ 이다.} \tag{3.1}$$

함수 δ는 $a_l^{(t)} \in DOM(A_l)$과 $x_{j,l}$ 사이의 비유사도의 합에 의하여 구할 수 있다. 함수 τ는 만약 두 값이 같을 경우 0.0, 다른 경우에는 신뢰값을 부여 한다. $\eta_i \left(= \sqrt{\sum_{t=1}^{n_l} (w_l^{(t)})^2} \right)$는 정규화 팩터 이다. 그러면 v_i에 대한 x_j의 소속도는 다음과 같은 식에 의하여 구할 수 있다.

$$\mu_{i,j} = [\sum_{z=1}^{k} (\frac{d_c(v_i, x_j)}{d_c(v_z, x_j)})^{1/(m-1)}]^{-1}. \tag{3.2}$$

이제 변경된 알고리즘에 대하여 i번째 센트로이드 v_i의 population을 갱신하는 방법을 고안한다. (2.6)식에서 속성 $v_{i,l}$의 갱신은 $\omega_l^{(t)} (1 \leq t \leq n_l)$에 대하여 다음과 같은 식에 의하여 계산할 수 있다.

$$\omega_l^{(t)} = \frac{1}{\lambda_i} \sum_{j=1}^{n} \gamma(x_{j,l}),$$

$$\text{여기서} \quad \gamma(x_{j,l}) = \begin{cases} m_{i,j}^m, & a_l^{(t)} = x_{j,l}, \\ 0, & a_l^{(t)} \neq x_{j,l}, \end{cases} \text{ 이다.} \tag{3.3}$$

$\lambda_i \left(= \sqrt{\sum_{j=1}^{n} \mu_{i,j}^{2m}} \right)$은 정규화 팩터이다. $v_{i,l}$은 범주의 값을 가지고 군집을 구성할 때 영향을 주며, 각 갱신반복 때 $\omega_l^{(t)}$에 의하여 갱신된다. 이러한 방법으로

k-population 알고리즘은 순차적으로 군집의 집합과 센트로이드를 갱신하고 $J_m(V:X)$함수에 도달할 때까지 반복 계산하게 된다.

4. 실험 및 결과

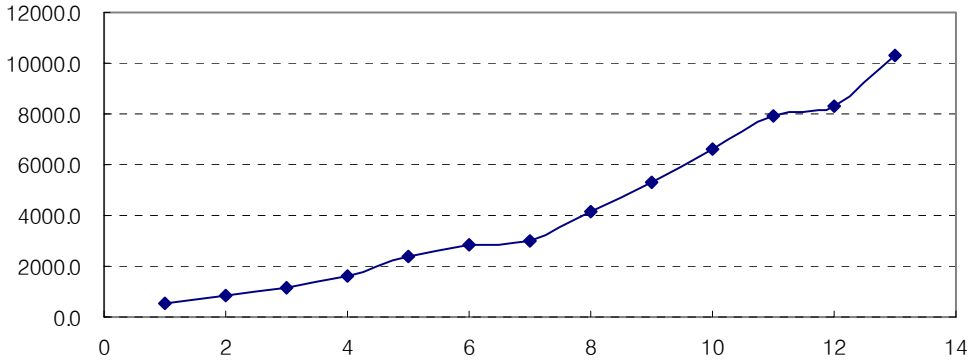
k-population 알고리즘의 효용성을 시험하기 위하여 범주형 데이터에 대해서 제안된 알고리즘과 계층적 알고리즘, *k*-modes 알고리즘, 퍼지 *k*-modes 알고리즘의 성능을 비교하였다. 각 알고리즘의 초기 센트로이드는 *k*개의 서로 다른 데이터를 무작위 추출하여 사용하였다. 퍼지 *k*-modes 알고리즘과 *k*-population 알고리즘에서 *m*은 1.1에서 2.0 사이의 값을 사용하였다.

각 알고리즘의 성능을 비교하기 위하여 테스트 데이터는 UCI Repository에서 Soybean, Zoo, Credit, Hepatitis 데이터 등 4개의 데이터 셋을 사용하였다. 군집화 결과를 분석하기 위하여 Huang의 정확도인 (r) 측정값을 사용하였는데, 만약 높은 (r)값은 보다나은 군집화 결과를 나타내며 완벽한 군집화의 경우 100.0%의 값을 가진다.

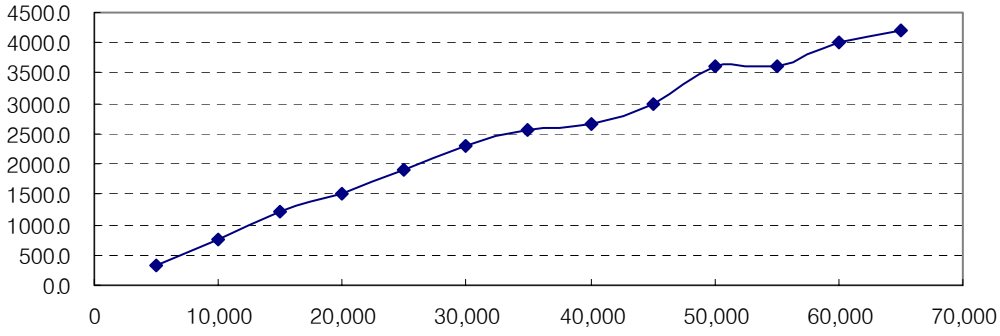
<도표 1> 테스트 데이터 셋에 대한 각 군집화 방법의 평균 정확도 (*r*)

Data set	Hierarchical clustering	*k*-modes clustering	Fuzzy *k*-modes clustering	*k*-population clustering
Soybean data	85.11	79.90	78.26	100.00
Zoo data	69.31	67.66	68.65	86.58
Credit approval data	53.73	73.30	73.30	84.83
Hepatitis domain data	78.17	70.57	70.57	79.61

<도표 1>은 각 알고리즘을 4가지 데이터 셋을 200번씩 적용하여 평균 정확도를 나타낸 것이다. Soybean 데이터는 콩의 질병에 대한 47개의 데이터 포인트를 가지고 있다. 각 데이터 포인트는 35가지의 군집 특성을 가지고 있으며 4가지의 질병 중에 하나로 판별된다. <도표 1>에서 보는 바와 같이 k-population 알고리즘이 다른 알고리즘보다 나은 군집화 성능을 볼 수 있다. 계층적 알고리즘은 85.11%의 정확도를 가져서 k-modes 알고리즘의 79.90%나 퍼지 k-modes 알고리즘의 78.26% 보다도 더 정확함을 알 수 있다. k-population 알고리즘은 Soybean 데이터 셋에서 계층적 알고리즘보다 14.9% 앞서는 100.0%의 정확도를 보여 주어 의미하는 바가 크다. 다른 세 가지 종류의 데이터 셋에서도 k-population 알고리즘은 나머지 알고리즘보다도 더 나은 성능을 보여주어 일반화된 실제 데이터에서도 그 효용성이 있음을 알 수 있다.



(a)



(b)

<그림 1> 대량의 데이터 셋에 대한 범용성 실험 결과

이와 더불어 제안된 알고리즘의 범용성 즉, 일반적인 특성을 가진 대량의 데이터 집합에서도 동작 속도나 군집화의 효율성을 검증하는바 이는 데이터가 대량인 경우 알고리즘의 동작시간이 비선형으로 늘어나 효율성이 떨어지는지 그 여부를 검증한 것이다. 사용한 데이터 셋은 65,000개의 데이터 포인트를 포함하고 있고 각 데이터 포인트는 42가지의 범주 속성을 갖고 있는 Connect데이터 셋이다. 주어진 데이터에 대한 군집의 개수와 주어진 군집에 대한 데이터의 개수에 대한 두 가지 범용성 시험을 하였다. <그림 1>의 (a)는 가로축은 군집의 개수(k)를, 세로축은 처리시간을 나타내었다. (a)실험에서 사용된 데이터의 개수는 65,000개이며 이 데이터를 서로 다른 군집으로 분류할 때 소요되는 시간을 나타내고 있으며 보는 바와 같이 군집 수에 대하여 처리시간이 선형으로 늘어남을 알 수 있다. <그림 1>의 (b)는 가로축이 데이터의 수, 세로축은 처리시간을 의미한다. 군집의 개수는 10개로 고정하고 데이터의 수가 증가함에 따른 처리시간을 살펴본 실험의 결과이다. <그림 1>에서 보는 바와 같이 k-population 알고리즘은 군집의 개수나 데이터의 개수가 증가함에 따라 처리시간이 선형으로 증가함을 볼 수 있다. 이는 제안한 알고리즘이 데이터의 크기나 군집의 수

에 대하여 처리시간이 선형으로 증가함으로 효율성이 있음을 알 수 있다.

5. 결론

기존 k-modes 방법을 사용하는 알고리즘들은 범주형 데이터를 비록 효율적으로 군집화 할 수 있지만, 단일의 결정적 센트로이드를 사용하고 있고 간단한 거리도 측정을 하여 정확도나 대량의 범주형 데이터에 적용하기에는 문제점을 내포하고 있었다. 이는 군집속의 데이터 특성을 반영하지 못한 결과이다. 따라서 본 연구에서는 population이라는 개념을 도입하여 각 군집의 센트로이드를 표현함으로 이러한 문제점을 개선하게 된 것이다. Population이란 개념은 범주의 값과 각 속성에 대한 신뢰도를 함께 내포하고 있다. 본 연구에서는 제안된 알고리즘이 다른 k-modes 방법을 사용하는 알고리즘보다 최종 군집화된 결과가 정확도에서 우수함을 보였다.

참고문헌

1. Gowda K. C., E. Diday, (1991). Symbolic clustering using a new dissimilaity measure, *Pattern Recognition* 24(6), 567-578
2. Huang Z., (1998). Extensions to the k-modes algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discriminations* 2(3), 283-304
3. Huang Z., Ng M. K., (1999). A fuzzy k-modes algorithm for clustering categorical data, *IEEE Tr. Fuzzy Systems.* 7(4), 446-452
4. Rencher A., (2002). Methods of Multivariate Analysis, Wiley-Interscience
5. Alan A., (2002). *Categorical Data Analysis*, Wiley

[2005년 4월 접수, 2005년 7월 채택]