http://dx.doi.org/10.14400/JDC.2015.13.11.157

k-Modes 분할 알고리즘에 의한 군집의 상관정보 기반 빅데이터 분석

박인규
중부대학교 컴퓨터•게임공학과

A Big Data Analysis by Between-Cluster Information using k-Modes Clustering Algorithm

In-Kyoo Park
Dept. of Computer · Game Engineering, College of Engineering

요 약 본 논문은 융복합을 위한 범주형 데이터의 부공간에 의한 군집화에 대해서 다룬다. 범주형 데이터는 수치형 데이터에만 국한되지 않기 때문에 기존의 범주형 데이터들의 평가척도들은 순서화(ordering)의 부재와 데이터의 고차원성과 희소성으로 인하여 한계를 가지기 마련이다. 따라서 각각의 군집에 존재하는 범주형 속성들의 상호 유사도을 보다 근접하게 측정할 수 있는 조건부 엔트로피 척도를 제안한다. 또한 군집의 최적화를 위하여 군집내의 발산을 최소화하고, 군집간의 독립성을 향상시킬 수 있는 새로운 목적함수를 제안한다. 제안된 알고리즘의 성능을 4개의 알고리즘과 비교검증하기 위하여 5가지의 데이터에 대하여 실험을 수행하였다. 비교검증을 위한 평가척도는 정확도, f-척도와 적응된 Rand 색인이다. 실험을 통하여 제안된 방법이 평가척도에 의한 결과에서 기존의 방법들보다 좋은 성능을 보였다.

주제어 : 융복합, 부공간 분할, 범주형 데이터, 군집화, 엔트로피

Abstract This paper describes subspace clustering of categorical data for convergence and integration. Because categorical data are not designed for dealing only with numerical data, The conventional evaluation measures are more likely to have the limitations due to the absence of ordering and high dimensional data and scarcity of frequency. Hence, conditional entropy measure is proposed to evaluate close approximation of cohesion among attributes within each cluster. We propose a new objective function that is used to reflect the optimistic clustering so that the within-cluster dispersion is minimized and the between-cluster separation is enhanced. We performed experiments on five real-world datasets, comparing the performance of our algorithms with four algorithms, using three evaluation metrics: accuracy, f-measure and adjusted Rand index. According to the experiments, the proposed algorithm outperforms the algorithms that were considered int the evaluation, regarding the considered metrics.

Key Words : Convergence and Integration, Subspace Partition, Categorical Data, Clustering, Entropy

* 본 논문은 2015년 중부대학교의 교내연구비에 의하여 지원되었음
Received 21 September 2015, Revised 28 October 2015
Accepted 20 November 2015
Corresponding Author: In-Kyoo Park(Joongbu University)
Email: fip2441g@gmail.com

ISSN: 1738-1916

1. 서론

임의의 객체들을 임의의 특성에 의해서 여러 부분으로 분할하는 군집화를 통해서 하나의 군집에 존재하는 객체들과 다른 여러 군집에 존재하는 객체들의 유사도를 파악할 수 있다. 이러한 유사도는 기하학적인 특성을 기반으로 평가가 이루어기 때문에 비교가 불가능한 범주 값을 가지는 경우에는 적용이 곤란하다. 또한 범주형 데이터의 고차원성으로 인하여 모든 쌍의 데이터에서 하나의 객체에 대하여 가장 가까운 객체와의 비유사도는 가장 멀리 있는 객체와의 비유사도와 거의 유사하게 되는 차원의 저주(the curse of dimensionality)가 존재한다. 따라서 군집화의 최적화에는 많은 변수가 존재한다[1].

고차원의 데이터는 속성들의 부분집합에 의한 부공간(subspace)으로 변환을 통하여 군집화의 유연성을 확보할 수 있다. 따라서 각 군집에 존재하는 각각의 속성들에 서로 다른 가중치를 통하여 각 속성의 군집형성의 기여도를 나타낸다. 그러나 각각의 군집마다 서로 다른 가중치가 할당되기 때문에 속성의 가중치가 군집화에 주요한 핵심이라고 할 수 있다. 꾸준히 많은 군집화 알고리즘이 제안되고 있는데 예를 들자면 [2]에서 임의의 군집에 대한 객체들의 평균거리에 따라서 각 가중치가 계산되어진다. 즉, 군집내의 거리의 합이 적은 속성에는 가중치가 크게 할당되고, 합이 큰 경우에는 가중치가 작게 할당되어 진다. [3]에 의한 분석을 통하여 보면 이러한 전략은 매개변수 의 선택에 민감하고 주어진 군집에 대한 주어진 속성의 가중치는 주어진 속성의 군집 모드에 존재하는 범주 값의 빈도수의 함수이다. 이 방법은 가중치를 결정하기 위해서 세 가지의 매개변수 , 와 가 종속되어 있다. [4]에서는 속성의 엔트로피에 대한 보수개념을 이용하고 있다. 보수 엔트로피는 하나의 속성이나 속성의 집합에 대하여 객체집합이 가지는 불확실성을 반영하기 때문에, 보수 값이 크면 불확실성도 비례하게 된다. [5]에서는 군집에서 엔트로피가 감소하게 되면 군집을 결정하는 과정에서 큰 가중치에 대한 차원의 부분집합들의 신뢰도가 높아진다는 개념을 이용하고 있다. 따라서 이러한 방법은 군집 내의 발산을 최소화시킴과 동시에 하나의 군집의 동정에 많은 차원의 속성이 기여할 수 있도록 음의 가중치 엔트로피를 최대화시켜준다. k-modes 알고리즘들이 효율적인 정교한 군집화를 가능하도록 하였지만 최종적인 군집화의 데이터에는 오류를 포함하고 있다[6,7,8,9].

본 논문에서는 군집화의 최적화를 위하여 k-modes 알고리즘을 확장한 형태의 조건부 엔트로피 기반 k-modes 알고리즘(entropy based k-modes)을 제안하였으며 하나의 속성에 대한 불확실성(uncertainty)을 조건부 엔트로피(conditional entropy)를 이용하여 속성의 유사도(similarity)를 측정하는데 오류를 줄였다. 속성에서 범주 값에 의한 각 부분집합의 엔트로피를 계산하여 속성의 불확실성을 결정한다. 이러한 방법으로 속성의 유사도는 군집에 존재하는 각 속성에 대한 엔트로피 값의 평균에 반비례하게 된다. 제안된 방법은 다양한 데이터를 대상으로 실험을 통하여 그 성능을 입증하였다.

2. k-Modes 군집의 유사도 척도

2.1 범주형 정보시스템

정보 시스템을 하나의 확률분포로 가정할 경우에 각각의 분포에 대한 불확실성의 생성과 측정은 필수적이라고 할 수 있다. 범주형 정보 시스템은 $S=(U,V,A,f)$라 하자. 여기서, U는 전체집합에 해당하는 객체들의 집합, A는 속성들의 집합, V는 속성이 가지는 도메인의 합집합이다. 또한 $f: U\times A\rightarrow V$는 매핑함수로서 $x\in U$와 $a\in A$에 대하여 $f(x,a)\in V_a$이다. $P\subseteq A$이면, 식별불가능 관계(indiscernibility relation)인 $IND(P)$에 의하여 속성에 대하여 객체의 분포가 형성된다.

〈Table 1〉 Example of categorical data set

objects/attributes	a_1	a_2	a_3	a_4	a_5
x_1	a	k	n	q	s
x_2	b	k	n	r	s
x_3	c	k	n	q	s
x_4	d	k	o	r	t
x_5	e	l	o	q	t
x_6	f	l	o	r	t
x_7	g	l	o	q	t
x_8	h	m	o	r	u
x_9	i	m	p	q	u
x_{10}	j	m	p	r	u

<Table 1>의 범주형 데이터는 전체집합 $U = \{x_1, x_2, ,...,x_{10}\}$ 속성집합 $A = \{a_1, a_2...,a_5\}$ 그리고 속성에 대한 각

각의 정의역은 $Dom(a_1)$ = {a,b, ...,j}, $Dom(a_2)$ = {k,l,m}, $Dom(a_3)$ = {n,o,p}, $Dom(a_4)$ = {q,r}, $Dom(a_5)$ = {s,t,u}로 구성되어진다. <Table 1>에서 U는 다음과 같이 세 가지의 부류로 분할된다고 가정한다. c_1={x_1,x_2,x_3,x_4}, c_2={x_5,x_6,x_7}, c_3={x_8,x_9,x_{10}}이고 이들의 중심값은 c_1 = {d,k,n,r,s}, c_2 = {g,l,o,q,t}, c_3 = {j,m,p,r,u}이다.

2.2 조건부 엔트로피에 의한 속성간의 유사도

정보이론에서 엔트로피는 임의의 랜덤변수가 가지는 불확실성의 척도로서 랜덤변수의 발생빈도가 균등할 수록 엔트로피가 커지게 되어 그와 관련된 변수의 불확실성도 커지게 되고, 반면에 편중될수록 엔트로피가 작아져서 불확실성이 줄어들게 된다. 결국 엔트로피는 랜덤변수 X의 각각의 상태 i의 확률을 p_i에 대한 정보량을 평균한 것으로 식 (1)과 같다[10].

$$\varepsilon(X) = -\sum_i p(x_i) log\ p(x_i) \tag{1}$$

확률변수 X, Y에 대해서 조건부 확률 $p(y|x)$에 대하여 조건부 엔트로피(conditional entropy) $ce(Y|X)$는 식(2)와 같다.

$$\begin{aligned} ce(Y|X) &= -\sum_{x \in A_x} p_X(x)\epsilon(Y|X=x) \\ &= -\sum_{x \in A_x}\sum_{y \in A_y} p_{XY}(x,y) log\, p_{XY}(y|x) \end{aligned} \tag{2}$$

조건부 엔트로피 $ce(Y|X)$는 X의 정보가 주어졌을 때 Y의 불확실성을 의미한다. 이것은 변수를 측정한 후나 이미 알고 있는 변수에 의해 얻어진 정보의 양을 말해준다. 결국 속성의 범주값에 대한 유사도를 측정을 측정하는 척도로 활용할 수 있다. 따라서 다음과 같이 속성과 객체간의 관계를 나타내는 함수를 나타낼 수 있다.

범주값 $a_h^{(l)} \in Dom(a_h)$을 $s \subseteq c_i$로 사상하는 함수 $\varnothing_i : (V \to 2^{c_i})$는 속성 a_h에 대하여 범주값 $a_h^{(l)}$을 가지는 분할 $c_i \in C$에 있는 모든 객체에 적용되고 식(3)과 같이 정의할 수 있다.

$$\varnothing_i(a_h^{(l)}) = \{x_q | x_q \in c_i \text{ and } x_{qh} = a_h^{(l)}\} \tag{3}$$

2^{c_i}는 c_i의 멱집합(powerset)으로 공집합과 c_i를 포함하는 c_i의 모든 부분집합을 의미한다. <Table 1>에서 $\varnothing_1(s)$는 $\{x_1, x_2, x_3\}$이다. 범주값 $a_h^{(l)}$을 $V' \subseteq V$집합에 대한 주어진 속성 $a_j \in A$로 사상하는 함수 $\alpha_i : V \times A \to 2^V$는 $c_i \in C$분할에서 범주값 $a_h^{(l)}$과 공존하는 속성 a_j의 범주값의 집합을 나타내며 식(4)와 같이 정의할 수 있다.

$$\alpha_i(a_h^{(l)}, a_j) = |\{a_j^{(p)} | \forall x_q \in \varnothing_i(a_h^{(l)}), x_{qj} = a_j^{(q)}\}| \tag{4}$$

<Table 1>에서 $\alpha_1(s, a_1)$는 $\{a,b,c\}$이다. 더욱이 a_j와 a_j가 동시에 주어질 경우에 $c_i \in C$에서 두 개의 범주값 $a_h^{(l)} \in Dom(a_h)$와 $a_j^{(p)} \in Dom(a_j)$를 객체의 개수로 변화하는 함수 $\psi_i : V \times V \to \mathbb{N}$는 식(5)와 같이 정의할 수 있고 <Table 1>에서 $\psi_i(s,q)$는 $|\{x_1, x_3\}|$로써 2이다.

$$\psi_i(a_h^{(l)}, a_j^{(p)}) = |\{x_q | x_q \in c_i \cap x_{qh} = a_h{}^{(l)} \cap x_{qj} = a_j^{(p)}\}| \tag{5}$$

본 논문에서는 주어진 범주값 $a_h^{(l)} \in Dom(a_h)$와 하나의 범주값 $a_j \in A$를 속성 a_j에 대하여 $\varnothing_i(a_h^{(l)})$로 사상하는 함수 $\varepsilon_i : V \times A \to \mathbb{R}$은 식(6)과 같이 정의하였다.

$$ce_i(a_h^{(l)}, a_j) = -P(\psi_i(a_h^{(l)}, a_j^{(p)}) * Q \tag{6}$$

$$Q = \sum_{a_j^p \in \alpha(a_h^{(l)}, a_j)} \frac{\psi_i(a_h^{(l)}, a_j^{(p)})}{|\varnothing_i(a_h^{(l)})|} log \frac{\psi_i(a_h^{(l)}, a_j^{(p)})}{|\varnothing_i(a_h^{(l)})|} \tag{7}$$

<Table 1>에서 $ce_1(s, a_1)$ = -(3/10) * (1/3*log1/3 + 1/3*log1/3 + 1/3*log1/3) = 0.33이고 반면에 $ce_1(s, a_3)$ = -(3/3*log3/3) = 0이다. 결국, a_5의 범주값 s에 대하여 a_1과 a_3간의 유사도를 고려할 경우에 a_1의 속성이 균일하고 a_3는 하나의 범주값 n으로 편중되어 있다.

〈Table 2〉 Attributes' Similarity by entropy

attributes	a_1	a_2	a_3	a_4	a_5
a_1	0	0	0	0	0
a_2	1.386	0	0.562	0.693	0.562
a_3	1.098	0	0	0.636	0
a_4	1.609	1.054	0.950	0	1.054
a_5	1.098	0	0	0.636	0

따라서 조건부 엔트로피의 경우에 균일한 경우보다 편중된 경우가 작게 나타나기 때문에 속성값의 분포가 유사도나 유사성이 보다 높다고 할 수 있다. <Table 2>와 <Table 3>에 엔트로피와 조건부 엔트로피에 의한 속성간의 유사도를 각각 나타내었다.

〈Table 3〉 Attributes' Similarity by conditional entropy

attributes	a_1	a_2	a_3	a_4	a_5
a_1	0	0	0	0	0
a_2	0.554	0	0.224	0.277	0.224
a_3	0.329	0	0	0.19	0
a_4	0.804	0.527	0.475	0	0.527
a_5	0.329	0	0	0.19	0

<Table 2>에서 각 속성의 인접속성과의 엔트로피는 a_1 = 0.107, a_2 = 0.2453, a_3 = 0.2238, a_4 = 0.2044, a_5 = 0.2192이다. 따라서 속성간의 우선순위는 a_1 - a_5 - a_4 - a_2 - a_3의 순서로 결정되었다. 반면에 <Table 3>에서 a_1 = 0.160, a_2 = 0.2155, a_3 = 0.2082, a_4 = 0.2099, a_5 = 0.2061로 속성의 우선순위는 a_1 - a_5 - a_3 - a_4 - a_2 의 순서로 결정되었다.

2.3 조건부 엔트로피에 의한 속성과 모드간의 유사도

모드의 범주값 z_{ij}와 속성 a_h와의 유사도를 측정하기 위하여 모드 z_i의 속성 $a_j \in A$의 중심값 z_{ij}에 대하여 주어진 분할 $c_i \in C$에서 주어진 범주속성 $a_h \in A$와 관계하는 조건부 엔트로피를 측정하는 조건부 엔트로피 함수 ce를 사용하였다. $1 \le i \le k$와 $1 \le q \le m$에 대하여 $z_{iq} \in Dom(a_q)$인 경우에 $z_i = \{z_{i1}, ..., z_{im}\}$ 과 같이 ce를 이용하여 분할 모드 $c_i \in C$로서 z_i을 고려하여, 모든 $a_j \in A$에 대하여 주어진 범주속성 $a_h \in A$을 $ce_i(z_{ij}, a_h)$ 값의 평균으로 사상하는 함수 $e_i : A \times \mathbb{R}$ 은 식(8)과 같이 정의할 수 있다.

$$e_i(a_h) = \sum_{z_{ij} \in z_i} \frac{ce_i(z_{ij}, a_h)}{|A|} \tag{8}$$

직관적으로 $e_i(a_h)$는 모드 z_i의 모든 범주값을 고려하여 a_h와 연관된 불확실성, 즉 유사도의 평균을 측정한다. <Table 1>에서 $ce_1(d, x_5)$ = 0, $ce_1(k, x_5)$ = 0.56, $ce_1(n, x_5)$ = 0, $ce_1(r, x_5)$ = 0.69, $ce_1(s, x_5)$ = 0이다. 결과적으로 $e_1(a_5)$ = (0 + 0.56 + 0.69 + 0)/5 = 0.25이다. 같은 방법으로 $e_1(a_1)$ = 0.86, $e_1(a_2)$ = 0, $e_1(a_3)$ = 0.25, $e_1(a_4)$ = 0.39가 된다. 조건부 엔트로피 기반 적합도 지수(conditional entropy based similarity index : *CESI*)는 분할 $c_i \in C$에 대한 주어진 범주속성 $a_h \in A$의 유사도를 측정한다. 이 척도는 식(9) 와 같이 함수 $CESI_i(a_h)$: $A \to \mathbb{R}$을 통하여 측정할 수 있다.

$$CESI_i(a_h) = \frac{\exp(-e_i(a_h))}{\sum_{a_j \in A} \exp(-e_i(a_h))} \tag{9}$$

<Table 1>의 경우에는 $CESI_1(a_1)$ =0.12, $CESI_1(a_2)$ =0.27, $CESI_1(a_3)$ =0.21, $CESI_1(a_4)$ =0.18, $CESI_1(a_5)$ =0.21 이다. $CESI_1(a_h)$는 $e_i(a_h)$ 에 반비례하는 것을 알 수 있다. $e_i(a_h)$가 작을수록 $CESI_1(a_h)$는 커지므로 해당하는 범주속성 $a_i \in A$의 중요성이 증가하게 된다.

3. 조건부 엔트로피 k-modes 알고리즘

군집에 존재하는 각 속성들의 유사도를 측정하기 위하여 엔트로피 기반의 군집간의 정보를 고려한 k-modes 알고리즘을 제안한다. 제안된 알고리즘의 전략은 군집간의 독립적인 정보만을 이용하지 않고 군집간의 정보를 연관하여 부공간의 탐색공간을 최소화 한다. k-means 알고리즘의 구조를 토대로 U를 k개의 군집으로 분할하기 위한 W, Z와 Λ의 탐색공간의 최소화문제는 식(10)과 같이 정의할 수 있다[11,12,13,14,15].

$$\min_{W,Z,\Lambda} P(W, Z, \Lambda) \sum_{l=1}^{k} \sum_{i=1}^{n} w_{li} d(x_i - z_l) \tag{10}$$

단,

$$\begin{aligned}
& w_{li} \in \{0,1\} \qquad 1 \le l \le k, 1 \le i \le n \\
& \sum_{l=1}^{k} w_{li} = 1 \qquad 1 \le i \le n \\
& 1 \le \sum_{l=1}^{k} w_{li} \le n, \qquad 1 \le l \le k \\
& \lambda_{lj} \in [0,1], \qquad 1 \le l \le k, 1 \le j \le m \\
& \sum_{j=1}^{m} \lambda_{lj} = 1 \qquad 1 \le j \le k
\end{aligned} \tag{11}$$

여기서 W=[w_{li}]는 k*n 이진 멤버쉽 행렬이고 w_{li} =1이라는 것은 x_i가 c_l에 할당됨을 의미한다. Z=[z_{lj}]는 k개의 군집 중심을 포함하는 k*m 행렬이고 Λ는 군집된 객체들이다. 본 논문에서는 목적함수의 수렴성을 향상시키기 위하여 비유사도 함수 $d(x_i, z_l)$는 식(12)와 같이 확장하였다. 식(12)의 유사도 함수에서 첫 번째 항은 군집내부의 발산을 최소화하기 위함이고, 두 번째 항은 군집간의 독립성을 향상시키기 위함이다. 제안된 유사도 척도는 비확률적인 정보와 확률정보를 결합하여 정보의 손실을 최소화하였다. 군집내부의 발산은 조건부 엔트로피라는 비확률 척도를 이용하였고 군집간의 독립성은 군집의 빈도수에 대한 확률척도를 이용하였다.

$$d(x_i, z_l) = \sum_{j=1}^{m} (\theta_{a_j}(x_i, z_l) + \lambda_l(a_j^h)) \quad (12)$$

$$\text{단, } \theta_{a_j}(x_i, z_l) = \begin{cases} 1, & x_{ij} \neq z_{lj} \\ 1-\lambda_{lj}, & x_{ij} = z_{lj} \end{cases}, \lambda_{lj} = CESI_l(a_j) \quad , \quad (13)$$

$$\gamma_l(a_j^h) = \frac{\phi_l(a_j^h)}{\sum_{c=1}^{k} \phi_c(a_j^h)}, a_j^h \in Dom(a_j)$$

```
1  Input: A set of objects  U and k of clusters
2  Output: The objects in  U with in k clusters
3  initialize the oldmodes as a k×|P|-ary empty array:
4  randomly choose k distinct objects x_1, x_2, ..., x_k from U
5  and assign [x_1, x_2, ..., x_k] to the k x |P|-ary newmodes:
6  for l = 1 to k
7    for j = 1 to m
8      set all initial weights λ_lj to 1/|A|;
9    end
10 end
11 While oldmodes ≠ newmodes
12  for i = 1 to |U|
13     for l = 1 to k
14        get the dissimilarity between the ith object and the lth mode
15        classify the ith object into the cluster whose mode is closest to it;
16     end
17  end
18  for l = 1 to k
19     find the mode z_l of each cluster and assign to newmodes;
20     for j = 1 to m
21     a_h ∈ A  of the l-th cluster,
22        using CECI_l(a_h) ;
23     end
24  end
25 end
```

[Fig. 1] Conditional Entropy k-Modes Algorithm (CEKM)

제약조건 (13)에 의한 목적함수 (12)의 최소화는 비선형의 최적화문제에 해당한다. 따라서 k-modes 알고리즘의 최적화는 첫째로, $Z^{(t)}$와 $\Lambda^{(t)}$를 고정하고 $W^{(t)}$에 대한 필요조건을 찾아서 $F(W^{(t+1)}, Z^{(t)}, \Lambda^{(t)})$을 부분 최소화한다. 둘째로, $W^{(t)}$와 $\Lambda^{(t)}$를 고정하고 $Z^{(t)}$에 대하여 $F(W^{(t)}, Z^{(t+1)}, \Lambda^{(t)})$을 부분 최소화한다. 마지막으로, $\Lambda^{(t)}$에 대하여 $W^{(t)}$와 $Z^{(t)}$를 고정하고 $F(W^{(t)}, Z^{(t)}, \Lambda^{(t+1)})$을 부분 최소화하고, $F(W^{(t+1)}, Z^{(t+1)}, \Lambda^{(t+1)}) = F(W^{(t)}, Z^{(t+1)}, \Lambda^{(t+1)})$이면 정지하고, 아니면 t=t+1로 정하고 둘째 단계로 간다. 결국 이러한 부분최소화(partial optimization)를 수행하는 과정을 구현한 k-means를 기반으로 구현한 것으로 조건부 엔트로피를 기반으로 군집간의 정보를 이용하여 군집내의 각 속성의 유사도를 측정할 수 있는 k-modes 알고리즘을 [Fig. 1]에 나타내었다.

알고리즘에서 제안된 비유사도 척도에 의한 k-modes 알고리즘의 시간 복잡도(time complexity)는 다음과 같이 계산되어 진다. 각각의 속성과 다른 하나의 속성간의 유사도를 계산하는 시간 복잡도는 $O(|U||P|k)$로 7-12줄에 나타나 있다. 또한 i 번째 객체를 l 번째 클러스터에 할당하는 시간 복잡도는 $O(|U|k)$로 14-19줄에 나타나 있다. 모든 클러스터의 갱신에 필요한 시간복잡도는 $O(|U||P|k)$로 20와 27줄에 나타나 있다. $CESI_l(a_h)$의 시간복잡도는 $O(|U|)$로 24줄에 나타나 있다. 반복횟수를 t로 가정하면 제안된 비유사도 척도에 의한 k-Modes 알고리즘의 전체적인 시간복잡도는 $O(|U||P|k) + t(O(|U|k + O(|U||P|k)))$로써 $O(t|U||P|k)$가 된다. 이는 객체, 속성과 클러스터의 개수와 선형적인 관계라는 것을 알 수 있다.

4. 실험 및 결과고찰

제안된 방법의 성능을 검증하기 위하여 기존의 방법들에 대하여 정확도(accuracy), f-척도와 ARI(adjusted rand index)척도를 비교하여 평가하였다. 실험에 사용된 데이터는 Congressional Voting Records, Mushroom, Breast Cancer, Soybean 과 Genetic Promotors이다. 이러한 데이터는 UCI저장소에서 이용하였다. <Table 4>에 실험에 데이터의 특성이 나타나 있다. 실험에서 제안된 알고리즘을 Standard k-modes (KM)[7], New Weighting k-modes (NWKM)[4], Mixed Weighting k-modes (MWKM)[3]와 k-modes의 기존의 알고리즘과 비교하였다. NWKM의 경우에 β는 2로 하였고 MWKM의 경우도 β는 2로 T_v는 1, T_s는 1로 설정하였다.

〈Table 4〉 Specifications of the data sets

Dataset	Turples	Attributes	Classes
Vote	435	17	2
Mushroom	8124	23	2
Breast Cancer	286	10	2
Soybean	683	36	19
Genetic promotors	106	58	2

<Table 3>의 데이터에서는 알고리즘의 결과가 임의적인 초기치 설정에 지배적이기 때문에 100번의 실행을 통한 결과를 평균하였다. k는 각 데이터의 군집의 수로 정하였다. 세 가지의 평가척도에 대한 실험결과가 <Table 5,6,7>에 나타나 있다. 각 데이터에서 알고리즘별로 아래 부분의 수치는 평균치를 나타내고, 위의 수치는 가장 양호한 수치를 나타낸다. 또한 굵은 수치는 각 데이터의 가장 우수한 결과를 나타낸다.

<Table 5,6,7>를 통하여 알 수 있는 것처럼 제안된 방법이 평가 척도에 따라서 양호한 결과를 보여주었다.

〈Table 5〉 Comparison of the accuracy

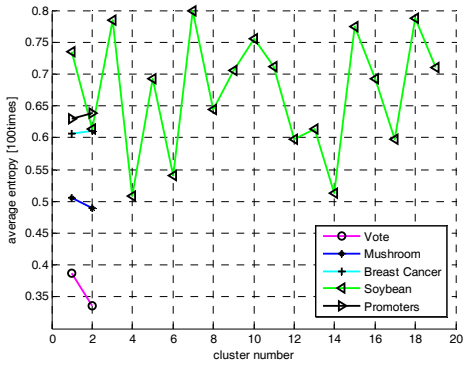
Algorithm	Vote	Mushroom	Breast Cancer	Soybean	Promotors	Average
KM	0.86 0.86	0.89 0.71	0.73 0.70	0.70 0.63	0.80 0.59	0.80 0.70
NWKM	0.88 0.86	0.89 0.72	0.75 0.70	0.73 0.63	0.77 0.61	0.80 0.70
MWKM	0.87 0.86	0.89 0.72	0.71 0.70	0.72 0.63	0.81 0.61	0.80 0.70
WKM	0.88 0.87	0.89 0.73	0.74 0.70	0.74 0.65	0.78 0.62	0.81 0.71
EBKM	0.88 0.87	0.89 0.76	0.74 0.70	0.75 0.66	0.83 0.62	0.82 0.72
CEKM	0.89 0.86	0.89 0.74	0.74 0.71	0.75 0.66	0.84 0.62	0.82 0.82

〈Table 6〉 Comparison of the f-measure

Algorithm	Vote	Mushroom	Breast Cancer	Soybean	Promotors	Average
KM	0.77 0.76	0.81 0.64	0.68 0.54	0.52 0.42	0.69 0.53	0.69 0.58
NWKM	0.80 0.78	0.81 0.64	0.70 0.56	0.55 0.42	0.66 0.54	0.70 0.59
MWKM	0.78 0.76	0.81 0.64	0.67 0.54	0.53 0.42	0.70 0.54	0.70 0.58
WKM	0.88 0.87	0.81 0.66	0.69 0.55	0.55 0.45	0.68 0.55	0.70 0.60
EBKM	0.88 0.87	0.81 0.68	0.69 0.56	0.57 0.47	0.73 0.55	0.72 0.61
CEKM	0.88 0.87	0.82 0.69	0.69 0.56	0.57 0.47	0.76 0.55	0.74 0.74

〈Table 7〉 Comparison of the ARI

Algorithm	Vote	Mushroom	Breast Cancer	Soybean	Promotors	Average
KM	0.52 0.51	0.61 0.26	0.19 0.01	0.48 0.37	0.36 0.06	0.43 0.24
NWKM	0.56 0.54	0.62 0.26	0.21 0.02	0.51 0.37	0.30 0.07	0.44 0.25
MWKM	0.56 0.52	0.62 0.28	0.14 0.01	0.49 0.37	0.39 0.07	0.44 0.25
WKM	0.57 0.54	0.62 0.28	0.18 0.02	0.51 0.41	0.32 0.08	0.44 0.27
EBKM	0.57 0.54	0.62 0.33	0.18 0.03	0.53 0.42	0.44 0.09	0.47 0.28
CEKM	0.58 0.54	0.62 0.33	0.19 0.02	0.53 0.42	0.44 0.08	0.47 0.47



[Fig. 2] Clustering error vs. different number of clusters

또한, 원래의 데이터가 가지는 각각의 군집의 개수에 대하여 각각의 알고리즘의 군집오차를 비교하였다.

5. 결론

본 논문에서는 군집화에 필수적인 속성간의 유사도를 계측할 수 있는 척도를 비확률적인 조건부 엔트로피를 기반으로 제안하였다. 이에 대한 적용을 두 단계로 구분하여 속성들 간에 적용하였고 속성과 모드 간에 적용하였다. 결과적으로 군집내의 분산도를 평가하는 엔트로피의 변형과 군집간의 상관 정보의 확률정보를 결합하여 부공간의 군집화를 위한 k-modes 분할 알고리즘에 적용하였다.

각 군집에 존재하는 각 속성의 유사도에 대한 척도로써 조건부 엔트로피에 기반한 유사도 지수(CESI)를 구하

였다. 결과적으로 임의의 속성의 적합도 지수는 해당 군집 모드에 있는 각 속성값에 대하여 얻어진 엔트로피의 평균에 반비례하였다. 이러한 접근법을 실제적인 데이터에 적용하여 정확도, f-척도와 ARI의 세 가지의 척도에 대하여 성능을 비교분석한 결과, 기존의 방법보다 부분적인 우위를 유지하였다. 이러한 군집 유사도는 범주 값의 크기(cardinality)가 각각 다른 속성들로 구성된 데이터 객체 집합의 분산도를 평가할 때 아주 유용할 것으로 기대된다.

ACKNOWLEDGMENTS

This work was also supported by Joongbu University of South Korea.

REFERENCES

[1] Sang-Hyun Lee, "A Study on Determining Factors for Manufacturers to Distributors Warehouse in Supply Chain", Journal of the Korea Convergence Society, Vol. 4, No. 2, pp. 15-20, 2013.

[2] E. Y. Chan, W. K. Ching, M. K. Ng and J. Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures", Pattern Recognition, Vol. 37, No. 5, pp. 943-952, 2004.

[3] L. Bai, J. Liang, C. Dang, and F. Cao, "A novel attribute weighting algorithm for clustering high-dimensional categorical data", Pattern Recognition, Vol. 44, No. 12, pp. 2843-2861, 2011.

[4] F. Cao, J. Liang, D. Li and X. Zhao, "A weighting k-modes algorithm for subspace clustering of categorical data", Neurocomputing, Vol. 108, pp. 23-30, 2013.

[5] L. Jing, M.K. Ng, and J. Z. Hunag, "An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparce data", Knowledge and Data Engineering, IEEE Transactions on, Vol. 19, No. 8, pp. 1026-1041, 2007.

[6] D. Barbara, Y. Li, and J. Couto, Coolcat: "an entropy-based algorithm for categorical clustering", in Proceedings of the 11th international conference on Information and knowledge management, ACM, pp. 582-589, 2002.

[7] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data mining and Knowledge Discovery, Vol.2, No. 3, pp. 283-304, 1998.

[8] F. Cao, J. Liang, D. Li, L. Bai and C. Dang, "A dissimilarity measure for the k-Modes clustering algorithm, Knowledge-Based Systems", Vol. 26, pp. 120-127, 2012.

[9] In-Kyu Park. "The generation of control rules for data mining", The Journal of Digital Policy & Management, Vol. 11, No.1, pp.343-349, 2013.

[10] J. L. Carbonera and M. Abel, "Categorical data clustering: a correlation-based approach for unsupervised attribute weighting", in Proceedings of ICTAI, 2014.

[11] J. L. Carbonera and M. Abel, "An entropy-based subspace clustering algorithm for categorical data", 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, pVol. 48, No. 26, pp. 272-277, 2014.

[12] G. Gan and J. Wu, "Subspace clustering for high dimensional categorical data", ACM SIGDD Explorations Newsletter, Vol. 6, No. 2, pp.87-94, 2004.

[13] M. J. Zaki, M. Peters I. Assent, and T. Seidl, "Clicks: An effective algorithm for mining subspace clusters in categorical datasets", Data & Knowledge Engineering, Vol. 60, No. 1, pp. 51-70, 2007.

[14] E. Cesario, G. Manco and R. Ortale, "Top-down parameter-free clustering fo high-dimensional categorical data", IEEE Trans. on Knowledge and Data Engineering, Vol. 19, No. 12, pp. 1607-1624, 2007.

[15] H.-P. Kriegel, P. Kroger and A. Aimek, "Subspace clustering", Wisley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 2, No. 4, pp. 351-364, 2012.

박 인 규(Park, In Kyoo)

- 1985년 2월 : 연세대학교 공학석사
- 1997년 2월 : 원광대학교 공학박사
- 1997년3월 ~ 현재 : 중부대학교 컴퓨터학과 교수
- 관심분야 : 소프트컴퓨팅, 데이터마이닝
- E-Mail : fip2441g@gmail.com