Journal of Korea Game Society 2018 Feb; 18(1): 105-114
http://dx.doi.org/10.7583/JKGS.2018.18.1.105

데이터마이닝 활용

빅 데이터의 자율 속성 감축을 위한 확장된 정보 엔트로피 기반 상관척도

박인규
중부대학교 게임 소프트웨어학과
fip2441g@gmail.com

Extended Information Entropy via Correlation for Autonomous Attribute Reduction of BigData

In-Kyu Park
Dept. of Game Software, College of Engineering Joongbu University

요 약

고객 유형 분석에 쓰이는 다양한 데이터 분석 방법은 고객들을 위한 맞춤형 콘텐츠를 기획하고, 보다 편리한 서비스를 제공하기 위하여 고객들의 유형과 특성을 정확히 파악하는 것이 매우 중요하다. 본 논문에서는 정보의 손실을 줄이기 위한 일환으로 정보 엔트로피를 확장하여 속성의 불확실성을 이용한 k-modes 군집분석 알고리즘을 제안한다. 따라서 속성에 대한 유사도의 측정은 두 가지의 측면에서 고려되어진다. 하나는 각 분할의 중심에 대한 각 속성간의 불확실성을 측정하는 것이고, 다른 하나는 각 속성이 가지는 불확실성에 대한 확률적 분포에 대한 불확실성을 측정하는 것이다. 특히 속성내의 불확실성은 속성의 엔트로피를 확률적 정보로 변환하여 불확실성을 측정하기 때문에 최종적인 불확실성은 비확률적인 척도와 확률적인 척도에서 고려되어 진다. 여러 실험과 척도를 통하여 제안한 알고리즘의 정확도가 최적의 초기치를 기반으로 군집분석을 수행한 결과에 준수함을 보인다.

ABSTRACT

Various data analysis methods used for customer type analysis are very important for game companies to understand their type and characteristics in an attempt to plan customized content for our customers and to provide more convenient services. In this paper, we propose a k-mode cluster analysis algorithm that uses information uncertainty by extending information entropy to reduce information loss. Therefore, the measurement of the similarity of attributes is considered in two aspects. One is to measure the uncertainty between each attribute on the center of each partition and the other is to measure the uncertainty about the probability distribution of the uncertainty of each property. In particular, the uncertainty in attributes is taken into account in the non-probabilistic and probabilistic scales because the entropy of the attribute is transformed into probabilistic information to measure the uncertainty. The accuracy of the algorithm is observable to the result of cluster analysis based on the optimal initial value through extensive performance analysis and various indexes.

Keywords : Information Entropy(정보 엔트로피), K-modes Clustering(*k*-모드 군집화), Categorical Data(범주형 데이터), Similarity(유사도), Uncertainty(불확실성)

Received: Dec. 18. 2017 Revised: Jan. 30. 2018
Accepted: Feb. 5. 2018
Corresponding Author: In-Kyu Park(Joongbu University)
E-mail: fip2441g@gmail.com

ISSN: 1598-4540 / eISSN: 2287-8211

1. 서 론

데이터의 단순한 저장이나 수집외에 대용량의 데이터를 저장, 수집, 발굴, 분석, 비즈니스화하는 빅데이터가 도래되었다. 최근 데이터의 영역은 각종 디지털 디바이스들을 통해 저장 수집된 데이터 속에서 가치 있는 정보를 찾아내어 알기 쉽게 전달하고, 정보를 원하는 사람이나 기관에 판매하는 비즈니스 과정을 포괄한다. 또한 여러 분야와의 융합을 통하여 무한한 형태의 패러다임을 형성하고 있다. 특히 군집(clustering)을 통한 데이터 마이닝은 비슷한 속성을 가지고 있는 데이터를 합치면서 의미 있는 군집을 형성하는 과정이며 다양한 기법들이 개발되어 왔다[1,2]. 특히, 게임에서 고객이 어떤 행동 패턴을 보이는지, 시간이 흐름에 따라 그 패턴이 어떻게 변화하는지, 그리고 긴밀한 관계를 맺고 있는 집단은 어떠한지를 통해 고객의 유형을 세분화와 같은 데이터 분석 방법에는 여러 가지 방법이 존재한다. 데이터 마이닝 기법의 일환으로서 군집분석의 효율성을 위한 내용은 데이터의 확장성, 속성이 다른 데이터 타입을 다루는 능력, 임의의 형태에서 군집 발견, 군집의 개수를 결정하는 방법, 이산치를 다루는 능력, 고차원, 입력순서, 상호 작동성에 있다. 데이터의 확장성은 대용량의 데이터를 다루는 능력으로써 수백만의 다양한 크기의 데이터에 대하여 일관성을 의미하고 있으며, 많은 알고리즘이 수치형 속성을 기반으로 하고 있기 때문에 다양한 속성(attributes)에 대한 일반화 능력이 군집분석에서 확보되어야 한다[3,4,5]. 하지만 실제데이터들은 범주형 속성(categorical attributes)및 혼합형 속성(mixed attributes)등으로 구성되어 있으며, 이러한 다양한 속성에 대한 분석방법이 필요하다.

범주형 데이터를 여러 개의 공간으로 분할하여 서로 유사한 특성을 가지는 객체들끼리 구성되는 공간을 형성하기 위하여 임의의 속성들에 의하여 여러 부공간이 형성되어 진다. 일반적으로 이러한 공간분할은 속성에 임의의 가중치를 부여하는 과정이 수반되어 지기 마련인데, 하나의 공간에 존재하는 하나의 속성은 하나의 가중치를 가지게 되어 공간과 속성의 연관성을 나타낸다. 이와 같은 속성의 가중치를 결정하는 과정에는 목적함수를 통하여 각 공간의 속성이 가지는 정수 가중치를 통하여 공간내에 존재하는 데이터를 재배치하여 공간을 재구성하는 소프트 가중치 (soft weighting)방법이 있다[6,7]. 이러한 방법은 데이터가 많을 경우에는 복잡성이 증가하게 된다. 반면에 하드 가중치(hard weighting)방법에서는 속성의 가중치가 부동소수점으로 운용되어 진다. 따라서 이러한 방법을 이용하여 EWKM, LAC, ESSC 등등 고차원의 데이터에 대한 많은 연구가 수행되어 왔다[8,9]. 결국 이러한 방법에서 속성이 가지는 값의 의미는 속성의 가지는 범주형 데이터들의 발산정도에 반비례하는 척도로 사용되어 진다. 속성의 가중치를 최적화하는 방법에 의하여 범주에 따른 데이터의 빈도수를 이용하는 방법과 범주별 분포를 고려하여 상관관계를 통한 휴리스틱 방법으로 분류되어 진다[10,11,12,13,14]. 결국, 이러한 방법들은 모델기반에 기초하고 있기 때문에 특정 공간의 데이터에 국한하고 또한 그러한 데이터의 분포는 고려하지 않고 있다.

본 논문에서 제안하는 확장된 정보 엔트로피를 이용한 분할 알고리즘을 통하여 공간내의 부분적인 정보뿐만 아니라 공간내의 분포를 고려하여 최적화된 속성의 가중치를 측정하고자 한다. 속성이 가지는 불확실성을 정보 엔트로피를 이용하여 측정하고 여기에 잡음 데이터로 인한 속성의 정보를 필터링할 수 있는 확률적 정보 엔트로피를 혼합하여 불확실성을 최적화 한다. 제안되는 정보 엔트로피를 k-modes 알고리즘에서 속성에 대한 불확실성(uncertainty)을 측정하고 군집을 수행한다.

2. 범주형 속성 엔트로피

범주형 데이터의 정보시스템 $DT = (U, A, V, f)$에

서 동치관계(indispensible relation)는 다음과 같이 정의할 수 있다.

$$IND(P) = \{(x,y) \in U \times U \mid \vee a \in P, f(x,a) = f(y,a)\}$$

(eq.1)

여기서 *IND*(*P*)는 U에 대한 동치류이고 *IND*(*P*) = ∩a∈P *IND*(*a*)이다. A의 부분집합 P가 주어지면 *IND*(*P*)는 범주형 데이터를 분할을 나타내며 *U*/*IND*(*P*) 는 {[xp] | x∈U}이다. [xp]가 P에 대한 동치류를 나타낸다. 결국, [x]p는 {y∈U | (x, y) ∈*IND*(*P*)}이다. [Table 1]에서처럼 *IND*(*P*)에 (x,y)가 속하면 x와 y는 P에 의하여 분별이 불가능하다.

[Table 1]> A Categorical Data Set

객체 \ 속성	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

범주형 데이터에서 유사한 그룹의 패턴을 효율적으로 분류하기 위해서는 여러 속성에서 중요하게 작용하는 속성을 추출할 필요가 있다. 결국 속성의 범주값에 대한 유사도의 측정을 측정하는 척도로 엔트로피를 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 $\varnothing_i : (V \rightarrow 2^{c_i})$ 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[15].

$$\varnothing_i(a_h^{(l)}) = \{x_q \mid x_q \in c_i \text{ and } x_{qh} = a_h^{(l)}\} \quad \text{(eq.2)}$$

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 $\alpha_i : V \times A \rightarrow 2^V$ 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.

$$\alpha_i(a_h^{(l)}, a_j) = |\{a_j^{(p)} \mid \forall x_q \in \varnothing_i(a_h^{(l)}), x_{qj} = a_j^{(q)}\}| \quad \text{(eq.3)}$$

<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 $\psi_i : V \times V \rightarrow \mathbb{N}$는 식(5)와 같이 정의할 수 있고 <Table 1>에서 $\psi_i(s,q)$는 $|\{x_1, x_3\}|$로써 2이다.

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

(eq.4)

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 $\varepsilon_i : V \times A \rightarrow \mathbb{R}$은 식(6)과 같이 정의하였다.

$$\beta_i(a_h^{(l)}, a_j) = -\sum_{a_j^{(p)} \in \alpha(a_h^{(l)}, a_j)} \frac{\psi_i(a_h^{(l)}, a_j^{(p)})}{|\varnothing_i(a_h^{(l)})|} log \frac{\psi_i(a_h^{(l)}, a_j^{(p)})}{|\varnothing_i(a_h^{(l)})|} \quad \text{(eq.5)}$$

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 범주 속성 $a_h \in A$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 경

우에 $z_i = \{z_{i1}, ..., z_{im}\}$ 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 $E_i(z_{ij}, a_h)$ 값의 평균으로 사상하는 함수 $\varepsilon_i : A \times \mathbb{R}$ 은 식(8)과 같이 정의할 수 있다.

$$\varepsilon_i(a_h) = \sum_{z_{ij} \in z_i} \beta_i(z_{ij}, a_h)/|A| \qquad \text{(eq.6)}$$

직관적으로 $\varepsilon_i(a_h)$는 모드 z_i의 모든 범주값을 고려하여 a_h와 연관된 불확실성, 즉 유사도의 평균을 측정한다. [Table 2]에서 $\beta_1(d, x_5)$ = 0, $\beta_1(k, x_5)$ = 0.56, $\beta_1(n, x_5)$ = 0, $\beta_1(r, x_5)$ = 0.69, $\beta_1(s, x_5)$ = 0이다. 결과적으로 $\varepsilon_1(a_5)$ = (0 + 0.56 + 0.69 + 0)/5 = 0.25이다. 같은 방법으로 $\varepsilon_1(a_1)$ = 0.86, $\varepsilon_1(a_2)$ = 0, $\varepsilon_1(a_3)$ = 0.25, $\varepsilon_1(a_4)$ = 0.39가 된다. 확장된 엔트로피 기반 적합도 지수(extended entropy based similarity index : *EESI*)는 분할 $c_i \in C$에 대한 주어진 범주속성 $a_h \in A$의 유사도를 측정한다. 이 척도는 식(9) 와 같이 함수 $\lambda_i(a_h)$: $A \rightarrow \mathbb{R}$을 통하여 측정할 수 있다.

$$\lambda_i(a_h) = \exp(-\varepsilon_i(a_h)) / \sum_{a_j \in A} \exp(-\varepsilon_i(a_h)) \qquad \text{(eq.7)}$$

[Table 1]의 경우에는 $\lambda_1(a_1)$ =0.12, $\lambda_1(a_2)$ =0.27, $\lambda_1(a_3)$ =0.21, $\lambda_1(a_4)$ =0.18, $\lambda_1(a_5)$ =0.21이다. $\lambda_1(a_h)$는 $e_i(a_h)$에 반비례하는 것을 알 수 있다. $e_i(a_h)$가 작을수록 $\lambda_1(a_h)$는 커지므로 해당하는 범주속성 $a_i \in A$의 중요성이 증가하게 된다.

임의의 분할에서 속성이 가지는 적합성(relevance)을 엔트로피를 이용하여 알아보기 위하여 임의로 세 개의 분할(c1, c2, c3)을 가정한다. c1={x1, x2, x3, x4}, c2={x5, x6, x7}, c3={x8, x9, x10}. 각각의 중심은 c1=(d, k, n, r, s), c2=(g, l, o, q, t), c3=(j, m, p, r, u) 이다. [Table 2]에서 클러스터 1(c1)에서 a_2의 k에 대하여 a_1과 a_3와의 각 패턴의 유사도를 고려해보자. $E_1(k, a_1)$ = -(1/4 * log(1/4) + 1/4 * log(1/4) + 1/4 * log(1/4) + 1/4 * log(1/4)) = 1.3863이고 반면에 $E_1(k, a_3)$ = -(3/4 * log(3/4) + 1/4 * log(1/4)) = 0.5623이다. 결국, a_1은 범주값이 {a, b, c}로 불규칙적이고 a_3는 범주값이 n으로 균일하게 규칙적이다. 따라서 a1이 a3보다 엔트로피 값이 크기 때문에 더 불안정하다고 할 수 있다. 결국 각 속성간의 영향력을 고려한 엔트로피(between attribute entropy)는 a1 = 5.19043, a2 = 1.0549, a3 = 1.5126, a4 = 1.9661과 a5 = 1.6172임을 알 수 있다.

[Table 2] Attribute uncertainty by entropy

속성	a_1	a_2	a_3	a_4	a_5
a_1(“d”)	0	0	0	0	0
a_2(“k”)	1.3863	0	0.5623	0.6931	0.5623
a_3(“n”)	1.096	0	0	0.6365	0
a_4(“r”)	1.6094	1.0549	0.9503	0	1.0549
a_5(“s”)	1.0986	0	0	0.6365	0
불안정척도	5.1903	1.0549	1.5126	1.9661	1.6172

3. 확장된 엔트로피 k-modes 알고리즘

3.1 범주형 속성의 확장된 엔트로피

본 논문에서는 범주형 데이터의 불확실성을 측정하기 위하여 범주형 데이터에서 임의의 속성의 인접속성들에 대한 각각의 엔트로피와 각 엔트로피의 확률분포를 통하여 유효한 속성과의 관계는 늘리고 그렇지 않은 속성과의 관계는 줄이는 확장된 엔트로피(E)를 다음과 같이 구성하였다.

$$E_i(a_h^{(l)}, a_j) = -\{\omega + \sum_{a_j^p \in IND(a_h^{(l)}, a_j)} \beta_i(a_h^{(l)}, a_h)\} \qquad \text{(eq.8)}$$

여기서, Q는 임의의 범주형 데이터 $a_h^{(l)} \in dom(a_h)$와 임의의 범주형 속성 $a_j \in A$ 을 이용하여 a_j에 대하여 a_h속성의 범주형 데이터 $a_h^{(l)}$집합이 가지는 불확실성으로써 속성간의 불확실성을 측정할 수 있

다. 이와 같은 방법으로 특정한 속성의 다른 속성들과의 불확실성의 분포를 보면 해당 속성에 많은 기여도를 가지는 속성이 있을 수 도 있고 그렇지 않은 속성도 존재하기 마련이다. 따라서 속성간의 미비한 영향력을 가지는 속성을 배제하기 위하여 해당 속성의 엔트로피의 분포에 대한 불확실성을 다음과 같이 측정할 수 있다.

$$\omega = -p(A)log_{\ 2}p(A) - (1-p(A))log_{\ 2}(1-p(A)) \quad \text{(eq.9)}$$

여기서, P(A) = -m + 평균 * r / N이다. 여기서 m = num(≥평균), r = num(=평균)이다.

여기서 a5의 경우에 클러스터의 각 중심값(c1의 경우 {“d”,“k”,“n”,“r”,“s”})에 대한 a5속성의 엔트로피의 분포는 {0, 0.5623, 0, 1.0549, 0}이다. 이 경우에 각각의 중심값이 a5속성에 대한 엔트로피는 {“r”}중심값에 의한 영향력(1.0549)이 가장 크고, 다음으로 {“k”}의 경우 (0.5623)이고 나머지 중심값들은 영향력이 없다. 따라서 임의의 속성내에서의 엔트로피를 긍정적인 영향력을 가지는 중심값과 부정적인 영향력을 가지는 중심값으로 확률분포를 고려하여 속성내의 엔트로피(within attribute entropy)를 식(10)에 의하여 [0.5004024, 0.5004024, 0.67301166, 0.67301166, 0.67301166]와 같이 구할 수 있다. 결국 속성간의 엔트로피와 속성내의 엔트로피를 결합하여 얻어진 각 속성의 엔트로피는 a1이 0.60762155, a2이 0.74570245, a3이 0.8968547, a4이 0.8774428 이고 a5이 0.89221835이다. <Table 2>에서 각 속성의 인접 속성간의 엔트로피는 a1 = 0.107, a2 = 0.2453, a3 = 0.2238, a4 = 0.2044, a5 = 0.2192이다. 따라서 속성간의 우선순위는 a2 - a3 - a5 - a4 - a1의 순서로 결정되었다. 반면에 속성간의 엔트로피와 속성내의 엔트로피를 결합한 엔트로피는 a1 = 0.6076, a2 = 0.7457, a3 = 0.8968, a4 = 0.8774, a5 = 0.8922로 속성의 우선순위는 a3 - a5 - a4 - a2 - a1 의 순서로 결정되었다.

3.2 확장된 엔트로피 k-modes 알고리즘

n개의 범주형 데이터를 k개로의 분할은 목적함수의 최적화 문제로 귀착되며, n개의 범주형 데이터를 k개의 군집에 대한 분할은 W, Z와 Λ의 탐색공간의 최소화문제는 식(10)과 같이 정의할 수 있다[11,12,13,14,15].

$$F(W,Z,\Lambda) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{li}\, d(x_i - z_l) \quad \text{(eq.10)}$$

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 유사도 함수 $d(x_i, z_l)$는 식(11)과 같이 확장하였다. 유사도 함수에서 첫 번째 항은 군집내부의 발산을 최소화하고, 두 번째 항은 군집간의 독립성을 향상시킨다. 제안된 척도는 비확률적인 정보와 확률정보를 결합하여 정보의 손실을 최소화하였다. 군집내부의 발산은 조건부 엔트로피라는 비확률 척도를 이용하였고 군집간의 독립성은 군집의 빈도수에 대한 확률척도를 이용하였다.

$$\begin{aligned} d(x_i, z_l) &= \sum_{j=1}^{m} (\theta_{a_j}(x_i, z_l) + \lambda_l(a_j^h)) \\ \theta_{a_j}(x_i, z_l) &= \begin{cases} 1, & x_{ij} \neq z_{lj} \\ 1-\lambda_{lj}, & x_{ij} = z_{lj} \end{cases}, \lambda_{lj} = EESI_l(a_j) \\ \gamma_l(a_j^h) &= \frac{\phi_l(a_j^h)}{\sum_{c=1}^{k}\phi_c(a_j^h)}, a_j^h \in Dom(a_j) \end{aligned}$$

(eq.11)

,

Input:: A number of clusters k and a categorical data set U
initialize the oldmodes as a $k \times_{|P|}$-ary empty array:
randomly choose k distinct objects $x_1,x_2,...,x_k$ from U
and assign $[x_1,x_2,...,x_k]$ to the $k \times_{|P|}$-ary newmodes:
and set all initial weights to 1/ $|A|$;
While oldmodes $\neq$ newmodes
 for i = 1 to $|U|$
  for i = 1 to k
   get the dissimilarity between the i th object and the i th mode
   classify the i th object into the cluster; the closest to it;
  end
 end
 for i = 1 to k
  find the mode z_l of each cluster and assign to newmodes;
  for j = 1 to m
  $A_h \in A$ of the i-th cluster,
   using $EIECI_i(a_h)$;
  end
 end
end
Output:: The objects in U with in k clusters

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

식(11)의 목적함수 최소화는 비선형의 최적화문제에 해당한다. 따라서 k-modes 알고리즘의 최적화는 첫째로, $Z^{(t)}$와 $\Lambda^{(t)}$를 고정하고 $W^{(t)}$에 대한 필요조건을 찾아서 $F(W(t+1),Z(t),\wedge(t))$을 부분 최소화한다. 둘째로, $W^{(t)}$와 $\Lambda^{(t)}$를 고정하고 $Z^{(t)}$에 대하여 $F(W(t),Z(t+1),\wedge(t))$을 부분 최소화한다. 마지막으로, $\Lambda^{(t)}$에 대하여 $W^{(t)}$와 $Z^{(t)}$를 고정하고 $F(W(t),Z(t),\wedge(t+1))$을 부분 최소화하고, $F(W(t+1), Z(t),\wedge(t+1)) = F(W(t),Z(t+1),\wedge(t+1))$ 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.

4. 실험 및 결과고찰

제안된 방법의 성능을 검증하기 위하여 기존의 방법들에 대하여 정확도(accuracy), 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 3]에 실험에 데이터의 특성이 나타나 있다. 실험에서는 제안된 알고리즘을 검증하기 위하여 Standard k-modes [7], Entropy Weighting k-modes와 제안된 Extended Entropy Weighting k - modes 알고리즘에 대하여 실험을 수행하였다.

[Table 3] 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 Promoters	106	58	2

분할이 {C1, C2, ..., Ck}이고 n개의 데이터를 가지는 집합 U에 대하여 분할의 정확성을 검증하기 위한 척도는 다음과 같다.

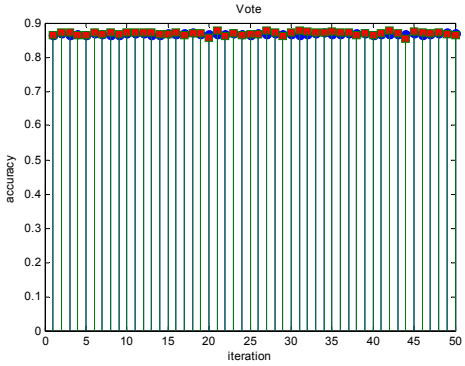
$$AC = \sum_{i=1}^{k} s_i / |U| \qquad \text{(eq.12)}$$

s는 분할된 데이터의 개수이다. [Table 4]는 5개의 데이터 집합에 대하여 각각 100번의 실험을 통하여 분할의 평균 정확도를 집계한 것이다.

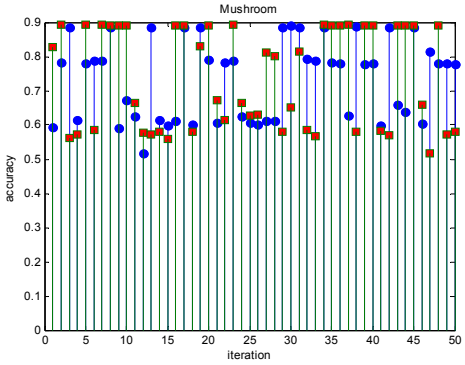
[Table 4] Specifications of the data sets

Dataset	Standard k-modes	Entropy k-modes	Extended Entropy k-modes
Vote	0.86 0.86	0.88 0.87	0.88 0.869
Mushroom	0.89 0.71	0.89 0.76	0.89 0.74
Breast Cancer	0.73 0.70	0.74 0.70	0.74 0.70
Soybean	0.70 0.63	0.75 0.66	0.73 0.70
Promoters	0.80 0.59	0.83 0.62	0.79 0.58

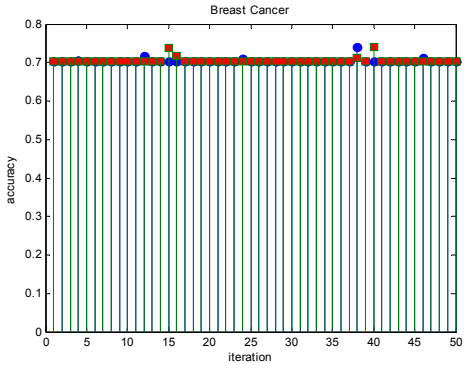
확장된 엔트로피에 의한 방법이 기존의 k-modes 알고리즘보다 우수한 결과를 보였다. 또한 제안된 방법이 기존의 엔트로피에 의한 방법보다 몇 가지의 데이터 집합에서 약간의 우수한 결과를 보였다.



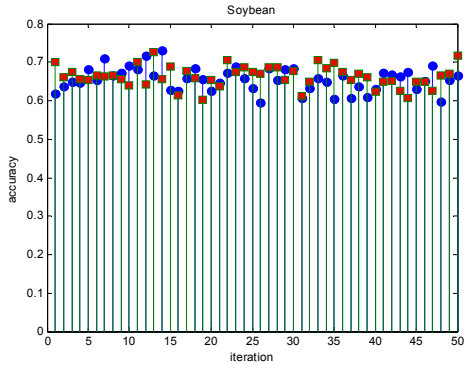
[Fig. 1] The subspace dimensions associated with each cluster on the vote



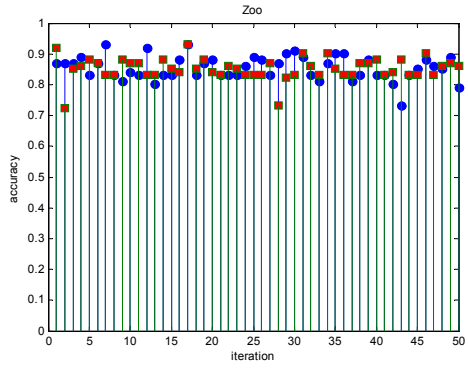
[Fig. 2] The subspace dimensions associated with each cluster on the Mushroom



[Fig. 3] The subspace dimensions associated with each cluster on the Breast Cancer



[Fig. 4] The subspace dimensions associated with each cluster on the Soybean



[Fig. 5] The Subspace Dimensions associated with Each Cluster on the Genetic Promotors

또한 가중치 값과 분할 데이터를 이용하여 분할된 부공간(subspace)을 [Fig. 4,5,6,7,8]과 같이 분

할의 정확도를 기준으로 살펴보았다. 각 분할직후에 속성의 개수의 역수(1/m)보다 큰 가중치를 가지는 각각의 데이터 집합에 대하여 서로 다른 분할에 존재하는 부공간의 정보를 추출하여 [Table 4]에 나타내었다. 각 데이터에 따른 100번의 실험결과에서 가장 우수한 결과를 토대로 추출하였다. 제안된 분할 방법에 의한 부공간을 살펴보면 다른 방법에 비하여 비교적 적은 차원이 형성된 것을 알 수 있다. 이는 분할과정에서 수반되는 오버헤드를 줄일 수 있는 근거를 제시하고 있음을 알 수 있다.

5. 결 론

본 논문에서는 군집화에 필수적인 속성간의 유사도를 계측할 수 있는 척도를 비확률적인 조건부 엔트로피를 기반으로 제안하였다. 이에 대한 적용을 두 단계로 구분하여 속성들간에 적용하였고 속성과 모드간에 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(EIESI)를 구하였다. 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 고객 유형 분석을 평가할 때 유용할 것으로 사료된다.

ACKNOWLEDGEMENTS

This paper was supported by Joongbu University Research & Development Fund, in 2017.

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] G. Gan and J. Wu, “Subspace clustering for high dimensional categorical data”, ACM SIGDD Explorations Newsletter, Vol. 6, No. 2, pp.87-94, 2004.
[12] 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.
[13] 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.
[14] 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.
[15] 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.

박인규(Park, In Kyu)

약 력 : 1985 연세대학교 전기과 전자계산기 응용(공학석사)
1997 원광대학교 전자과 마이크로 프로세서 응용(공학박사)
1997-현재: 중부대학교 게임소프트웨어학과 교수
관심분야 : 데이터 마이닝, 소프트웨어컴퓨팅