$k-$평균무리짓기알고리듬에서 불명확성을 고려한 표본무게결정의 한가지 방법

현철민, 윤룡한

자료발굴의 중요한 분야인 무리짓기에서 $k-$평균무리짓기알고리듬은 계산량이 적고 효과적인것으로 하여 널리 쓰이고있다.[1]

$k-$평균무리짓기알고리듬의 무리짓기정확도를 개선하기 위하여 표본들의 무게를 결정하는 여러가지 방법들이 연구되고있다.[1－3]

선행연구[2]에서는 표본과 무리중심사이의 거리에 의해서 표본들의 무게를 결정하였으며 선행연구[3]에서는 표본과 무리중심들사이의 거리와 함께 각을 리용하여 표본의 명확성, 불명확성을 정의하고 그것에 의해 무게를 결정하는 방법을 제기하였다.

그러나 선행연구[3]의 방법은 불명확한 표본에 대하여 불명확성만 고려하고 그것의 불명확한 정도는 고려하지 못하였으며 무리중심근방의 표본들도 불명확한 표본으로 되는 제한성이 있다.

론문에서는 표본들의 명확성과 불명확성을 새롭게 정의하고 불명확한 표본에 대하여 그것의 불명확한 정도까지 고려하여 무게를 결정함으로써 무리짓기정확도를 높이였다.

1. 표본의 명확성과 불명확성, 무게결정방법

$X=\{x_i \mid i=1,\ 2,\ \cdots,\ n\}$ 을 n 개의 표본, X 가 m 개의 무리 $C_j\ (j=1,\ 2,\ \cdots,\ m)$ 로 나누어지며 $c_j\ (j=1,\ 2,\ \cdots,\ m)$ 를 매 무리의 중심이라고 하자. $\overrightarrow{c_jc_q}$ 를 두 무리중심점 c_j 와 c_q 를 지나는 벡토르, $\prod_{jq}$ 를 중심점 c_j 를 지나며 $\overrightarrow{c_jc_q}$ 를 법선벡토르로 하는 초평면

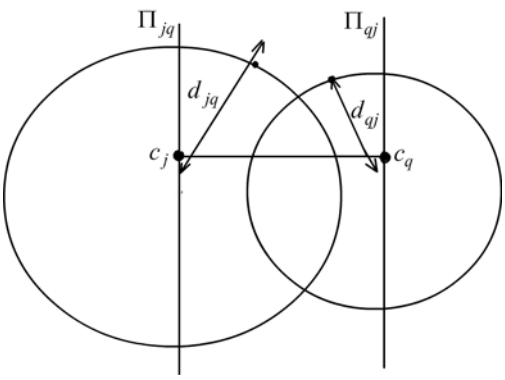

그림. 표본의 명확성판정

$$A_{jq}=\{x_i\in C_j \mid (\overrightarrow{c_jx_i},\ \overrightarrow{c_jx_i})>0 \wedge d(x_i,\ c_j)<d(c_j,\ c_q)\}$$

$$d_{jq}=\frac{1}{|A_{jq}|}\sum_{x_i\in A_{jq}}\| x_i-c_j\|^2$$

라고 하자.(그림)

정의 $x\in C_j$ 일 때 모든 $q\in\{1,\ 2,\ \cdots,\ m\}\backslash\{j\}$ 에 대하여 $x\notin V(c_j,\ d_{jq})\bigcap V(c_q,\ d_{qj})$ 이면 x 는 C_j 에서 명확하다고 말하고 그렇지 않으면 불명확하다고 말한다. 여기서 $V(c,\ d)$ 는 중심이 c 이고 반경이 d 인 구이다.

무리 C_j 에 대하여 $\alpha_q^{(j)}$, $X_q^{(j)}$ 를

$$\alpha_q^{(j)}=\frac{d(c_j,\ c_q)-d_{qj}}{d_{jq}} \tag{1}$$

$$X_q^{(j)} = \{x \in C_j \mid x \in V(c_j,\ d_{jq}) \cap V(c_q,\ d_{qj})\}$$

로 놓고 C_j의 불명확한 표본 x_i에 대하여 모임 $X_q^{(j)}$가 x_i를 포함하는 $q \in \{1,\ 2,\ \cdots,\ m\} \setminus \{j\}$들의 모임을

$$Q_q^{(ij)} = \{q \in \{1,\ 2,\ \cdots,\ m\} \setminus \{j\} \mid x_i \in X_q^{(j)}\}$$

라고 하자. 이에 기초하여 불명확한 표본의 무게 w_{ij}를 다음과 같이 결정한다.

$$w_{ij} = \prod_{q \in Q_q^{(ij)}} \alpha_q^{(j)} \exp(-\| x_i - c_j \|^2) \tag{2}$$

또한 명확한 표본의 무게 w_{ij}를 다음과 같이 결정한다.

$$w_{ij} = \exp(-\| x_i - c_j \|^2) \tag{3}$$

식 (2), (3)으로부터 다음의 사실들이 성립한다는것을 쉽게 알수 있다.

무리중심 c_j까지의 거리가 같은 명확한 표본 x_i와 불명확한 표본 x_k에 대하여 $w_{ij} > w_{kj}$이다.

두 불명확한(명확한) 표본 x_i, x_k에 대하여 $d(x_i,\ c_j) \geq d(x_k,\ c_j)$이면 $w_{ij} \leq w_{kj}$이다.

두 불명확한 표본 x_i, x_k에 대하여 $|Q_q^{(ij)}| > |Q_q^{(kj)}|$이면 $w_{ij} < w_{kj}$이다.

2. 제안한 무게붙은 k-평균무리짓기알고리듬

입력: n개의 표본 $X = \{x_i \mid i = 1,\ 2,\ \cdots,\ n\}$, 무리의 개수 m, 중심의 변동 ε, 최대반복수 T

출력: 무리들의 대표점 c_j $(j = 1,\ 2,\ \cdots,\ m)$

걸음 1 m개의 무리중심 c_j $(j = 1,\ 2,\ \cdots,\ m)$를 임의로 초기화하고 $t = 1$로 놓는다.

걸음 2 매 표본들에 대하여 모든 무리중심까지의 거리를 계산하고 거리가 최소인 무리에 표본을 소속시킨다.

걸음 3 매 무리에 대하여 무리내의 표본들의 명확성을 판정한다.

걸음 4 불명확한 표본에 대해서는 식 (2)로, 명확한 표본에 대하여서는 식 (3)에 의하여 그것의 무게 ω_{ij}를 결정하고 무리중심

$$c_j = \sum_{i=1}^{n} g_{ij}\omega_{ij}x_i \Big/ \sum_{i=1}^{n} g_{ij}\omega_{ij}$$

를 갱신한다. 여기서 $g_{ij} = \begin{cases} 1, & x_i \in C_j \\ 0, & x_i \notin C_j \end{cases}$ 이다.

걸음 5 $\max\limits_{1 \leq j \leq m} \| c_j - c_j' \| < \varepsilon$이거나 $t = T$이면 걸음 6으로 이행하고 그렇지 않으면 반복수를 하나 증가하고$(t = t + 1)$ 걸음 2로 이행한다. 여기서 c_j'는 전단계에서 얻어진 무리중심이다.

걸음 6 c_j $(j = 1,\ 2,\ \cdots,\ m)$를 출력한다.

3. 성 능 평 가

알고리듬의 성능을 KDD CUP99자료기지와 weak 3.9의 표준자료화일인 diabetes.arff를 가지고 평가하였다. 무게붙은 $k-$평균알고리듬(KM)과 선행연구[3]의 방법(SWKMA), 론문의 방법으로 무리짓기를 진행하였을 때 그 정확도는 표와 같다.

표. 알고리듬의 성능평가

자료기지	자료수/개	무리수/개	KM	SWKMA	론문의 방법
KDD CUP99	25192	5	72.28%	76.21%	78.07%
diabetes	768	2	67.8%	68.2%	70.1%

참 고 문 헌

[1] R. Xu et a1.; IEEE Transactions on Neural Networks, 16, 3, 645, 2005.

[2] P. F. Huang et a1.; Neurocomputing, 73, 16, 2935, 2010.

[3] Jianyuan Li et a1.; IEEE Transactions on Neural Networks, 27, 2, 589, 2015.

주체 108(2019)년 6월 10일 원고접수

A Sample's Weight Decision Method for $k-$Means Clustering Algorithm Considering Ambiguity of the Samples

Hyon Chol Min, Yun Ryong Han

We propose a new weighted $k-$means clustering algorithm based on ambiguity of the samples. The experimental results indicate that the proposed algorithm have better effect than KM and SWKMA.

Key words: $k-$means clustering algorithm, clustering algorithm, sample weighting scheme