응용통계연구 제 17권 1호, 2004년, pp. 135-144

K-평균 군집화의 재현성 평가 및 응용

허명회[1] 이용구[2]

요 약

K-평균 군집화(K-means clustering)는 고객 세분화(customer segmentation) 등 데이터 마이닝에서 중요한 한 몫을 하는 비지도 학습방법(unsupervised learning method)이다. K-평균 군집화가 재현성(reproducibility)이 있는가를 보기위하여, 다수의 기존 연구에서는 관측 자료를 2개 셋으로 나눈 자료 분할(data partitioning) 방법이 활용되고 있다. 본 교신에서 우리는 이보다 개념적으로 명확한 새로운 자료 분할 방법을 제안한다. 이 방법은 관측 자료를 3개 셋으로 나누어 그 중 2개 자료 셋을 독립적인 군집화 규칙을 생성하는 데 사용하고 나머지 1개의 자료 셋을 규칙간 일치성을 테스트하는 데 사용한다. 또한 2개의 군집화 규칙간 일치성 평가를 위한 지표로서 엔트로피 기준의 활용 방법을 제시한다.

주요용어: K-평균 군집화, 데이터 마이닝, 재현성, 자료 분할, 훈련 자료, 테스트 자료, Rand 지수, 엔트로피 기준.

1. 연구 배경과 제안

군집화는 고객 세분화(customer segmentation) 등 데이터 마이닝에서 중요한 한 몫을 하는 비지도학습(unsupervised learning) 방법이다. 대표적인 것이 K-평균 군집화(K-means clustering)인데 이것을 하려면 군집 수를 몇 개로 할 것인지를 미리 정해야 한다. 자료 분석자로서는 곤혹스러운 일이다. 면밀한 자료탐색(data exploration)을 하더라도 다변량 자료의 분포적 특성을 알아내는 것은 결코 쉽지 않기 때문이다.

재현성(reproducibility; 타당성, validity)은 과학적 방법에 대한 주요 요건 중 하나이다. 어떤 방법으로 만들어졌건 형성된 군집화 결과가 재현성을 갖는가를 확인할 필요가 있다. 즉 동일한 메커니즘에서 생성된 독립적인 새 데이터 셋을 동일한 방식으로 군집화한 결과가 기존 군집화 결과와 유사하다면 재현성이 있는 것이므로 그런 군집화는 바람직하다. 그러나 두 결과가 상당부분 불일치한다면 그런 군집화는 재현성의 결여라는 문제를 갖는다.

이와 같은 2단계 작업에는 시간이 걸리고 더욱이 시간적 정상성(stationarity)를 가정하여야 하므로 자료 분석자가 활용하는데 제약이 걸린다. 따라서 군집화의 재현성을 일시에 평가할 필요가 있겠는데, 이를 위하여 자료 분할(data partitioning) 방법을 활용할 수 있다. 자료 분할은 지도학습(supervised learning) 모형의 적합과 평가를 위하여 흔히 쓰여 왔다.

1) (136-701) 서울특별시 성북구 안암동 5-1, 고려대학교 통계학과 교수.
E-mail: stat420@korea.ac.kr
2) (156-756) 서울특별시 관악구 흑석동 221, 중앙대학교 응용통계학과 교수.
E-mail:leeyg@cau.ac.kr

여태까지 비지도학습(unsupervised learning)에서는 별로 활용되지 않았으나 전혀 없었던 것은 아니다. 몇몇 기존 연구에서 군집화 평가를 위해 제시된 자료 분할 방식은 다음과 같다 (McTintyre and Blashfiled, 1980; Morey, Blashfield and Skinner, 1983; Milligan, 1996; Gordon, 1999, p.184).

0) 주어진 자료를 임의로 2개로 분할한다. 그 중 하나를 자료 셋 1이라고 하고 다른 하나를 자료 셋 2라고 하자.

1) 자료 셋 1에 대한 군집화를 수행한다. 이 때 생성된 군집화 규칙을 규칙 1이라고 하자. 자료 셋 2의 각 개체를 규칙 1에 따라 분류한다. [K-평균 군집화에서 K 개의 군집 중심이 산출된다. 군집화 규칙은 각 개체를 가장 가까운 군집 중심을 찾아 해당 군집 레이블을 부여하는 것이다.]

2) 자료 셋 2를 동일한 방식으로 군집화하여 규칙 2를 생성해 낸다. 이에 따라 자료 셋 2의 각 개체를 분류한다.

3) 자료 셋 2의 개체들에 대한 규칙 1과 규칙 2의 분류 결과를 교차분류표로 나타낸다. 적용된 군집화가 일치적이라면 이 표에서 행과 열은 강한 대응성을 보일 것이다. 그러나 그렇지 않다면 행과 열의 대응성은 약하게 나타날 것이다.

신경망이나 나무형 분류 등 지도학습에서는 흔히 전체 자료를 훈련 자료(training data)와 테스트 자료(test data)로 분할하는데, 전자를 모형의 구축에, 후자를 모형의 평가에 활용함으로써 용도에 따라 역할이 구분된다. 이에 따라 군집화의 재현성 평가를 위한 앞의 절차를 검토해보면, 자료 셋 1이 훈련 자료에 해당한다고 볼 수 있지만 자료 셋 2의 역할이 모호하다. 규칙 2에 의한 분류가 자료 내부로부터 생성되기 때문에 엄격히 말하자면 자료 셋 2를 테스트 자료로 보기 어렵다.

이러한 문제를 해결하고자 우리는 전체 자료를 3개로 나누는 다음 자료 분할 방법을 제안하고자 한다.

0) 주어진 자료를 임의로 3개로 분할하여 자료 셋 1,2,3으로 명명한다.

1) 자료 셋 1에 대한 군집화를 수행하여 군집화 규칙 1을 생성해낸다.

2) 자료 셋 2를 동일한 방식으로 군집화하여 군집화 규칙 2를 생성해낸다.

3) 자료 셋 3의 각 개체를 규칙 1과 규칙 2에 따라 분류한다. 그리고 그 결과를 교차분류표로 나타낸다. 이 표에서 행과 열의 대응성이 뚜렷하게 보여야 한다.

이 절차에서 자료 셋 1과 2의 역할은 훈련용이고 자료 셋 3의 역할은 테스트용으로 자료의 용도에 따른 구분이 개념적으로 명확하다. 우리는 자료 셋 1,2,3의 비율을 대략 40%, 40%, 20%로 할 것을 제안 한다. 그러나, 이와 같은 자료 분할은 자료 크기가 크지 않은 경우 낭비적인 측면이 있다. 그런 경우엔, 한 심사자가 제안하였듯이, 자료 셋을 같은 크

기의 3개 셋으로 분할한 뒤 각 셋을 테스트 용도로 순환 활용하는 일종의 교차 타당성 확인(cross-validation) 방법을 생각할 수 있을 것이다.

2절에서 우리의 방식을 한 수치 예에 적용해보고 3절에서는 2개 군집화 분류간 일치도 측도에 관하여 고찰하면서 새로운 제안을 한다. 4절에서는 모의자료 사례에 적용하여 볼 것이다. 이하, 군집화 및 분류의 산출시 SPSS사의 클레멘타인 버전 7.1이 사용되었다.

2. 수치 예: Telco CAT 자료

제안 방법론의 데모로서 클레멘타인 Telco CAT 자료 셋 churn.txt에 적용해보도록 하겠다 (SPSS, 2000). 이 자료는 1477명의 고객에 관한 통화관련 기록으로 여기서 사용할 군집화 변수는 장거리 통화량(Longdist), 국제통화량(International), 시내통화량(Local) 등 3개이다. 단, Local이 지수분포 형태를 보였으므로 군집화 적용에 앞서 로그 변환을 취하였다: $Log\_local = log10(Local + 1)$.

앞 절에서 제안한 절차에 맞추어 K-평균 군집화에 대한 재현성을 평가한다:

0) 데이터 분할: 고객 번호 ID를 5로 나누어 나머지가 1이나 2면 자료 셋 1 (훈련 자료 1)로 하고, 나머지가 3이나 4면 자료 셋 2 (훈련 자료 2)로 하였으며 나머지가 0인 경우 자료 셋 3 (테스트 자료)으로 하였다. 이렇게 하여 전체 1477개 레코드가 571개 (39%), 604개 (41%), 302개(20%) 로 분할되었다. [자료 셋 구성비율이 정확히 40%, 40%, 20%가 아닌 이유는 ID 번호에 군데군데 빈 곳이 있기 때문이다.]

1) 훈련 자료 1에 군집수 K=4인 K-평균 군집화를 적용해보았다. 여기서 적용한 군집 수 4는 잠정 시도한 값일 뿐이다. 앞으로, 이 때 생성된 분류규칙을 Kmeans1으로 부를 것이다.

2) 훈련 자료 2에 군집수 K=4인 K-평균 군집화를 적용해보았다. 이 때 생성된 분류규칙을 Kmeans2라고 하자.

3) 테스트 자료의 각 개체에 Kmeans1과 Kmeans2를 적용하여 2개의 군집 레이블을 산출하였다. 다음은 Kmeans1와 Kmeans2간의 연관성을 보여주는 교차분류표이다.

K=4		Kmeans2			
		군집1	군집2	군집3	군집4
Kmeans1	군집1	112	0	0	1
	군집2	0	21	0	0
	군집3	6	3	143	0
	군집4	0	0	0	16

앞의 표에서 행(Kmeans1)과 열(Kmeans2)이 강하게 대응하고 있다. [군집 레이블은 임의적(arbitrary)이기 때문에, 여기서 '강한 대응'이란 두 군집화의 레이블간 거의 1:1인 대응

적 관계가 존재하는 것을 뜻한다. 다시 말하여, 군집 레이블이 꼭 순서대로 일치해야 하는 것은 아니다.] 따라서 K=4인 K-평균 군집화의 재현성이 크다고 높다고 볼 수 있다. 그렇다고 더 이상의 시도를 해 볼 필요가 없는 것은 아니다. 이보다 큰 군집 수 K=5, 6, ··· 인 경우에서도 재현성이 확보된다면 대부분의 연구자들은 보다 세밀한 군집화를 선호할 것이기 때문이다.

1절의 방식에 따라 군집 수 K=5, 6, 7에 대한 재현성을 평가하였다. <표 2.1>에 그 결과를 정리하였다. K=5의 군집화는 행과 열의 대응성이 뚜렷하지만 K=6, 7의 군집화는 그렇지 않은 것으로 나타났다.

<표 2.1> Telco CAT 자료에 대한 K-평균 군집화의 재현성 (K=5, 6, 7)

K=5		Kmeans2				
		군집1	군집2	군집3	군집4	군집5
Kmeans1	군집1	45	0	0	0	0
	군집2	0	20	0	0	0
	군집3	0	4	106	0	12
	군집4	0	0	0	16	0
	군집5	5	0	0	0	94

K=6		Kmeans2					
		군집1	군집2	군집3	군집4	군집5	군집6
Kmeans1	군집1	45	0	0	0	4	0
	군집2	0	8	0	1	0	13
	군집3	0	0	51	0	19	3
	군집4	0	0	0	15	0	0
	군집5	0	0	1	0	77	0
	군집6	0	0	63	0	0	2

K=7		Kmeans2						
		군집1	군집2	군집3	군집4	군집5	군집6	군집7
Kmeans1	군집1	37	0	0	0	0	0	0
	군집2	0	10	0	2	0	10	0
	군집3	0	0	48	0	2	2	0
	군집4	0	0	0	10	0	0	5
	군집5	7	0	0	0	51	0	2
	군집6	0	0	48	0	0	3	0
	군집7	0	0	15	0	47	1	2

<표 2.1>에서 볼 수 있듯이, 어떤 교차분류표에서는 2개 군집화간 일치적 경향이 뚜렷하게 나타나지만 어떤 교차분류표에서는 그렇지 않다. 다음 절에서 2개 군집화간 일치성에 대한 수치적 측도에 대하여 고찰하기로 한다.

3. 2개 군집화간 일치도 지표

2개 군집화간 일치도를 재는 지표로 널리 받아들여지고 있는 것은 Rand Index (Rand 1971)와 Hubert와 Arabie의 수정 Rand Index (Hubert and Arabie, 1985)이다 (Gordon, 1999, p.198). 국내에서 채성산(1997; Chae, 1995; Chae and Warde, 1991)이 Rand Index의 성질 및 활용에 대하여 집중적인 연구를 한 바 있다.

Rand Index는 다음과 같이 산출된다. 행 i 와 열 j 로 분류된 개체의 수를 n_{ij} 라고 하자 $(i, j = 1, \cdots, K)$. 임의의 두 개체 a와 b의 군집 분류를 살펴보면 다음 세 경우 중 어느 하나가 될 것이다. n은 분류된 개체의 총 수이다.

1) a와 b가 훈련 자료 1의 군집화에 의하여 한 군집으로 분류된다. 동시에 훈련 자료 2의 군집화에 의하여도 한 군집으로 분류된다.

2) a와 b가 훈련 자료 1의 군집화에 한 군집으로 분류된다. 그러나 훈련 자료 2의 군집화에 의하여는 다른 군집으로 분류된다.

3) a와 b가 훈련 자료 2의 군집화에 의하여 한 군집으로 분류된다. 그러나 훈련 자료 1의 군집화에 의하여는 다른 군집으로 분류된다.

4) a와 b가 훈련 자료 1의 군집화에 의하여 다른 군집으로 분류된다. 동시에 훈련 자료 2의 군집화에 의하여도 다른 군집으로 분류된다.

$\binom{n}{k}$ 개의 가능한 개체 쌍 중에서 1), 2), 3), 4)에 해당하는 쌍의 수는 각각 다음과 같다.

1) 의 경우의 수 $\sum_{i=1}^{K} \sum_{j=1}^{K} \binom{n_{ij}}{2}$ $(= N_1)$.

2) 의 경우 수: $\sum_{i=1}^{K} \binom{n_{i\cdot}}{2}$ - $\sum_{i=1}^{K} \sum_{j=1}^{K} \binom{n_{ij}}{2}$ $(= N_2)$. (여기서, $n_{i\cdot} = \sum_{j=1}^{K} n_{ij}$)

3) 의 경우 수: $\sum_{j=1}^{K} \binom{n_{\cdot j}}{2}$ - $\sum_{i=1}^{K} \sum_{j=1}^{K} \binom{n_{ij}}{2}$ $(= N_3)$. (여기서, $n_{\cdot j} = \sum_{i=1}^{K} n_{ij}$)

4) 의 경우 수: $\binom{n}{2}$ - $\sum_{i=1}^{K} \binom{n_{i\cdot}}{2}$ - $\sum_{j=1}^{K} \binom{n_{\cdot j}}{2}$ + $\sum_{i=1}^{K} \sum_{j=1}^{K} \binom{n_{ij}}{2}$ $(= N_4)$.

따라서 일치 쌍은 모두 $N_1 + N_4$개, 비일치 쌍은 모두 $N_2 + N_3$ 개가 되고 총 합계는 $\binom{n}{2}$개이다. Rand Index는 총 개체 쌍 중 일치하는 쌍의 비율인

$$\text{Rand Index} = (N_1 + N_4) \;/\; \binom{n}{2}$$

로 정의된다. 그런데 우연하게 일치하는 개체 쌍이 다수 나올 수 있음을 감안하여 Hubert와 Arabie (1985)는 다음과 같이 Rand Index를 수정하였다.

$$\text{Corrected Rand Index} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}\binom{n_{ij}}{2} - \sum_{i=1}^{K}\binom{n_{i\cdot}}{2}\sum_{j=1}^{K}\binom{n_{\cdot j}}{2}/\binom{n}{2}}{[\sum_{i=1}^{K}\binom{n_{i\cdot}}{2} + \sum_{j=1}^{K}\binom{n_{\cdot j}}{2}]/2 - \sum_{i=1}^{K}\binom{n_{i\cdot}}{2}\sum_{j=1}^{K}\binom{n_{\cdot j}}{2}/\binom{n}{2}}$$

이 수정 지수는 모든 쌍이 경우 1)이나 4)인 경우 1의 값을 취하고 두 군집화가 완전히 무관할 때 평균적으로 0의 값을 취하도록 되어 있다. 이 같은 Rand Index의 수정은 교차 분류표의 행과 열의 합인 $n_{i\cdot}$와 $n_{\cdot j}$에 조건화하였을 때 우연적인 일치 쌍 개수 기댓값을 감안한 것이다. 그런데, 군집화 재현성이 군집크기에도 나타나야 하므로 이 조건화가 과연 합당한지에는 의문의 여지가 있다.

Rand Index와 병행하여 사용할 수 있는 2개 군집화간 일치도 지표로서, 우리는 엔트로피(entropy) 기준에 의한 지표 *Ent*를 제안하고자 한다.

$$AveEnt = (RowEnt + ColEnt)/2,$$

여기서 *RowEnt*와 *ColEnt*는 각각 행 다항 분포와 열 다항 분포의 엔트로피를 결합한 것으로

$$RowEnt = \sum_{i=1}^{K}\frac{n_{i\cdot}}{n}\left\{\sum_{j=1}^{K}-\frac{n_{ij}}{n_{i\cdot}}\log\left(\frac{n_{ij}}{n_{i\cdot}}\right)\right\},\ ColEnt = \sum_{j=1}^{K}\frac{n_{\cdot j}}{n}\left\{\sum_{i=1}^{K}-\frac{n_{ij}}{n_{\cdot j}}\log\left(\frac{n_{ij}}{n_{\cdot j}}\right)\right\}$$

로 표현된다 ($0\ log0 = 0$). 엔트로피는 불확실성(uncertainty)에 대한 측도로 다항형 범주자료에 적용되는 경우 등확률 분포에서 최대가 되며 비균등성이 심화될수록 작아진다. 엔트로피는 이미 C5.0 모형에서 확고히 활용되고 있다 (Quilan, 1993).

<표 3.1>은 Telco CAT 자료의 군집화에서 산출된 엔트로피 지수를 보여준다. K=5에서 K=6으로 바뀜에 따라 *AveEnt*가 급증하는 것을 볼 수 있다. 이것은 K=5의 군집화가 적절한 것을 말한다.

<표 3.1> Telco CAT 자료에 대한 K-평균 군집화의 재현성 평가 지표 (K=4, 5, 6, 7)

	Rand	C. Rand	*Row Ent*	*Col Ent*	*Ave Ent*
K =4	0.95	0.90	0.151	0.121	0.136
5	0.92	0.80	0.252	0.214	0.233
6	0.85	0.59	0.330	0.551	0.440
7	0.81	0.45	0.453	0.790	0.621

4. 모의자료 예

모의자료에 적용하여 일반적인 결론을 유도하는 것은 대개 가능하지 않은 일이다. 따라서 여기서는 알려진 구조의 자료에 앞에서 제안된 방법론을 적용하여 봄으로써 방법론에 대한 제한된 지지를 얻고자 한다.

<표 4.1> 모의자료에 대한 K-평균 군집화의 재현성 (K=6, 7, 8)

K=6		Kmeans2					
		군집1	군집2	군집3	군집4	군집5	군집6
Kmeans1	군집1	98	0	0	0	0	0
	군집2	0	88	0	0	0	0
	군집3	0	0	0	0	110	0
	군집4	0	0	105	0	0	96
	군집5	0	0	1	90	0	1
	군집6	0	0	0	111	0	0

K=7		Kmeans2						
		군집1	군집2	군집3	군집4	군집5	군집6	군집7
Kmeans1	군집1	98	0	0	0	0	0	0
	군집2	0	88	0	0	0	0	0
	군집3	0	0	0	0	110	0	0
	군집4	0	0	0	0	0	97	0
	군집5	0	0	0	1	0	0	89
	군집6	0	0	0	111	0	0	0
	군집7	0	0	106	0	0	0	0

K=8		Kmeans2							
		군집1	군집2	군집3	군집4	군집5	군집6	군집7	군집8
Kmeans1	군집1	98	0	0	0	0	0	0	0
	군집2	0	88	0	0	0	0	0	0
	군집3	0	0	0	0	58	0	52	0
	군집4	0	0	0	0	0	97	0	0
	군집5	0	0	0	1	0	0	0	89
	군집6	0	0	0	111	0	0	0	0
	군집7	0	0	60	0	0	0	0	0
	군집8	0	0	46	0	0	0	0	0

모의자료는 7변량 개체 3,500개로, 7개의 소그룹으로 구성되었다: 독립적으로 그룹 j $(= 1, \cdots, 7)$의 개체들이

$$(x_1, \cdots, x_7) \sim \mathrm{N}\,(m_j, I_7)$$

에 따라 500개가 생성되었다. 여기서 평균 m_j는 j번째 요소만 5이고 나머지 요소들은 모두 0인 7x1 벡터이고 공분산 (I_7은 대각요소가 1인 7x7 대각행렬이다. 따라서 이 자료에 대하여

는 K=7이 참 값이다. 훈련 자료 1과 2는 전체 자료의 3,500개 개체 중 임의 추출한 1,400개의 개체로 구성하였으며 테스트 자료는 남은 700개 개체로 구성하였다.

<표 4.1>이 제안된 방법론을 적용하였을 때 얻는 테스트 개체들에 대한 군집 분류 결과이다. 지면의 제약상 K=6, 7, 8인 경우만 제시하였다. 그리고 <표 4.2>는 K=5, 6, 7, 8, 9의 군집화에 대하여 계산된 Rand Index, 수정 Rand Index 및 엔트로피 값이다.

<표 4.1>에서 K=7의 군집화가 안정적임을 시각적으로 볼 수 있고 <표 4.2>에서도 K=7에서 엔트로피가 최소임을 확인할 수 있다.

<표 4.2> 모의자료의 군집화에 재현성 평가 지표 (K=5, 6, 7, 8, 9)

	Rand	C. Rand	*Row Ent*	*Col Ent*	*Ave Ent*
K =5	0.83	0.51	0.435	0.407	0.421
6	0.92	0.72	0.215	0.214	0.214
7	1.00	1.00	0.008	0.008	0.008
8	0.98	0.89	0.117	0.112	0.114
9	0.96	0.79	0.226	0.200	0.213

5. 맺음 말

본 교신에서는 K-평균 군집화의 재현성을 평가하는 새 자료 분할 방법을 제안하고 2개 군집화에 대한 일치성 측도로 엔트로피 기준의 활용을 제시하였다. 그러나 기존의 방법들과의 통계적 측면에서의 비교는 다루어지지 않았다. 한편, 채성산 (1997)은 군집화의 타당성 (재현성) 검토를 위하여 붓스트랩 방법의 활용을 제안한 바 있는데 좋은 연구방향일 것으로 생각한다. 향후 이들 방법론간 비교 연구를 기대해 본다.

K-평균 군집화가 요즘의 데이터 마이닝 방법론에서 중요한 역할을 하고 있기 때문에 본 교신에서 그것에 바탕을 두었을 뿐, 우리들의 방법론이 K-평균 군집화에만 적용 가능한 것은 아니다.

K-평균 군집화시 군집 수를 막연하게 정하여 왔다면, 최소한 재현성 기준에 근거하여 본 교신의 방법론을 활용해 볼 것을 데이터 마이너들에게 권한다.

참고문헌

[1] 채성산 (1997). "재표본 추출 및 검정을 통한 집락 수의 예측 ", 자연과학 (대전대학교 기초과학연구소 논문집), 8집 1호, 73-88.

[2] Chae, S.S. (1995). "An asymptotic result concerning a comparative statistic in cluster analysis," *Natural Science* (Taejon University), 6. 125-138.

[3] Chae, S.S. and Warde, W.D. (1991). "A method to predict the number of clusters," *Journal of Korean Statistical Society* , 20. 162-176.

[4] Gordon, A.D. (1999). *Classification* (2nd Edition). Chapman and Hall, Boca Raton. Chapter 7 (pp.183-211).

[5] Hubert, L. and Arabie, P. (1985). "Comparing partitions," *Journal of Classification*, 2. 193-218.

[6] McIntyre, R.M. and Blashfield, R.K. (1980). "A nearest-centroid technique for evaluating the minimum-variance clustering procedure," *Multivariate Behavioral Research,* 15. 225-238.

[7] Milligan, G.W. (1996). "Clustering validation: Results and implications for applied analyses," in *Clustering and Classification* (Edited by P. Arabie et al.) World Scientific, Singapore. 341-375.

[8] Morey, L.C., Blashfield, R.K. and Skinner, H.A. (1983). "A comparison of cluster analysis techniques within a sequential validation framework," *Multivariate Behavioral Research,* 18. 309-329.

[9] Quilan, J.R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann. CA: San Mateo.

[10] Rand, W.M. (1971). "Objective criteria for the evaluation of clustering methods," *Journal of American Statistical Association,* 66. 846-850.

[11] SPSS (2000). Clementine Application Templates for Telecommunication Industries (Telco CAT). Chicago, SPSS Inc.

[2003년 4월 접수, 2003년 6월 채택]

Reproducibility Assessment of K-Means Clustering and Applications

Myung-Hoe Huh [1] Yong Goo Lee [2]

ABSTRACT

We propose a reproducibility (validity) assessment procedure of K-means cluster analysis by randomly partitioning the data set into three parts, of which two subsets are used for developing clustering rules and one subset for testing consistency of clustering rules. Also, as an alternative to Rand index and corrected Rand index, we propose an entropy-based consistency measure between two clustering rules, and apply it to determination of the number of clusters in K-means clustering.

Keywords: K-means clustering, data mining, cluster reproducibility (validity), data partitioning, training data, testing data, Rand index (Hubert-Arabie corrected), entropy criterion.

1) Professor, Dept. of Statistics, Korea University. Anam-Dong 4-1, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr

2) Professor, Dept. of Applied Statistics, Chung-Ang University. HukSuk-Dong 221, Seoul 156-756, Korea. E-mail: leeyg@cau.ac.kr