Journal of Digital Contents Society Vol. 12 No. 3 Sep. 2011(pp.339 - 345)

외부 군집 연관 기준 정보를 이용한 군집수 최적화

이현진*, 지태창**

요 약

군집화는 주어진 데이터를 분할하여 데이터 속에 숨겨져 있는 의미를 자동으로 발견하는 방법이다. k-means는 간단하고 빠른 군집화 알고리즘 중의 하나이다. 군집의 수 k는 군집화를 수행하는데 매우 중요한 요소이며, k의 값에 의해 군집화 결과가 달라진다. 본 논문에서는 반복적인 k-means 수행과 군집의 품질을 평가하는 외부 군집 연관 기준 정보를 결합하여 최적의 군집수를 결정하는 방법을 제안한다. 실험 결과 기존의 방법들에 비하여 제안하는 방법이 군집수의 정확성 측면에서 우수한 성능을 보였다.

A Study on Optimizing the Number of Clusters using External Cluster Relationship Criterion

Hyunjin Lee*, Taechang Jee**

Abstract

The k-means has been one of the popular, simple and faster clustering algorithms, but the right value of k is unknown. The value of k (the number of clusters) is a very important element because the result of clustering is different depending on it. In this paper, we present a novel algorithm based on an external cluster relationship criterion which is an evaluation metric of clustering result to determine the number of clusters dynamically. Experimental results show that our algorithm is superior to other methods in terms of the accuracy of the number of clusters.

Keywords : Clustering, External Cluster Relationship Criterion, K-means, Number of Clusters

1. 서 론

군집화는 주어진 데이터를 여러개의 군집으로 분할하는 방법이다. 사전지식 없이 스스로 학습하는 무교사학습(unsupervised learning)에 해당한다. 군집화는 유사한 데이터 개체들의 군집으로 분할하여 주어진 데이터에 대한 이해와 활용을 효율적으로 할 수 있도록 하는 방법이다[1].

군집화에서 많이 사용하는 알고리즘 중의 하나인 k-means는 self-organizing map [2, 3]과 함께 큰 규모의 군집화에 자주 사용되고 있다.

※ 제일저자(First Author) : 이현진
접수일:2011년 9월 02일, 수정일:2011년 9월 15일
완료일:2011년 9월 16일
* 한국사이버대학교 컴퓨터정보통신학과
hjlee@mail.kcu.ac
** 연세대학교 컴퓨터과학과

데이터마이닝 분야에서 k-means를 위한 고성능 기법들도 개발되어 왔다[4, 5].

군집수를 자동적으로 선택하는 것은 군집화에서 가장 어려운 문제 중의 하나이다[6]. 보통 군집화 알고리즘은 여러 개의 k값에 대해 수행된 다음에 군집 평가 함수의 값이 가장 높은 k 값을 최적의 값으로 선택하게 된다. 군집수를 선택하는 여러 연구들이 수행되고 있지만, 의미있는 군집수를 결정하는 일은 쉬운 일이 아니다.

Figueiredo와 Jain은 최소 문자 길이 (minimum message length (MML)) 기준과 가우시안 혼합 모델 (Gaussian mixture model (GMM))을 결합하여 K 값을 추정하는데 사용했다[7]. 그들은 처음에 많은 개수의 군집에서 시작하여 군집을 결합하는 방법을 사용하였고, 결합하면서 MML 기준이 최소가 되는 K를 선택하였다.

Tibshirani 등은 Gap statistics라는 군집수를 결정하는 방법을 제안하였다[8]. 주요 가정은 데

이터가 최적의 군집수로 나뉘어 졌을 때, 개별 군집은 임의의 작은 변화(perturbation)에 탄력이 있다는 것이다.

Rasmussen은 군집수를 위한 비모수적 기준을 제시하였다[9]. 주요 아이디어는 비 모수적인 베이지안 (Byesian) 기준을 군집수를 추정하는데 제안한 것이다. 이는 군집수의 정확성을 평가하는 확률 모델로 많이 사용되고 있다.

Pelleg와 Moore는 k-means를 확장한 방법인 X-means를 제안하였는데, 이 방법에 군집의 수를 추정하는 기능을 추가하였다[5, 10]. 여기서는 군집의 분리 여부를 판단할 때 베이지안 정보 기준을 사용하였다. 군집을 분리할 때 계산되는 정보 이득(Information Gain)이 군집을 유지할 때 계산되는 정보 이득보다 클 경우에 군집의 분리는 이루어지게 된다.

Salvador는 '군집의 수 vs. 군집 평가 척도' 그래프의 "knee"를 발견하는 L 방법을 제안하였다[11]. 이 방법의 군집화와 분류(segment) 알고리즘은 군집과 분류의 수를 결정하는데 좋은 결과를 보였지만, 이 방법은 계층적 알고리즘에만 적용 가능하다.

Lu는 최적의 군집수를 추정하는 진화(evolutionary) 알고리즘을 제안하였다[12]. 제안하는 진화 알고리즘은 새로운 엔트로피(entropy) 기반 적합 함수 (fitness function)와 군집을 분리하고, 결합하고, 제거하는 세 개의 새로운 유전 연산자 (operator)를 정의하였다. 이 방법을 사용하여 데이터 집합에서 최적의 군집수를 추정 할 수 있었다.

Boutsinas은 z-window 군집화 알고리즘을 제안하였다[13]. 이 방법은 윈도잉 (windowing) 기법을 사용해서 군집의 수를 결정하는 것을 목표로 한다. 주 아이디어는 충분히 많은 수의 초기 윈도우를 설정하고, 알고리즘을 수행하면서 윈도우들을 결합하는 것이다.

지태창 등은 다차원 척도법을 활용하여 기하학적인 방법으로 군집의 수를 결정하는 방법을 제안하였다[14]. 이 방법은 충분히 큰 초기 군집수에서 시작하여 군집수를 점점 축소하여 최적 군집수를 추정하는 방법이다.

최근에도 Skewed Distribution을 이용하는 방법[15]과 G-means를 확장한 방법[16], 인공 면역 시스템에서 지역 네트워크 이웃을 이용하는 방법[17] 등 군집수를 결정하기 위한 다양한 연구들이 수행되고 있다.

일반적으로 군집을 생성할 때 자신과 다른 군집의 중심에 대한 유사도를 이용하여 군집에 속할지 여부를 결정하게 된다. 하지만, 외부 군집에 대한 유사도도 존재하고, 동일 군집에 속하는 데이터들도 외부 군집에 대한 유사도는 서로 다른 경향을 보일 수 있기 때문에 군집의 충실도를 결정하는데 있어서 외부 군집에 대한 연관도를 고려해야 한다. 따라서 본 논문에서는 외부 군집 연관 기준을 정의하고 이를 군집화 알고리즘의 척도로 이용하여 군집수를 찾는 방법을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 외부 군집 연관 기준에 대해서 살펴보고 3장에서는 군집화 알고리즘과 외부 군집 연관 기준을 결합하는 방법을 살펴본다. 4장에서는 실험과 결과를 분석하고 5장에서 결론을 맺는다.

2. 외부 군집 연관 기준

좋은 군집의 선택은 군집 내의 데이터들은 서로 조밀하게(density) 연결되며 외부 군집의 데이터들과는 성기게(sparse) 연결되는 데이터의 그룹을 선택하는 것으로 정의된다. 즉, 같은 군집에 속한 데이터 간에는 조밀한 연결이 이루어져야 하고, 외부 군집 사이에도 성긴 연결이 존재하게 된다.

군집에 속한 데이터들의 외부 군집에 대한 반응을 분석한 것이 외부 군집 연관 기준(External Cluster Relationship Criterion, ECRC)이다. 군집화는 데이터와 군집 사이의 거리를 계산하여, 가까운 거리의 데이터들로 군집을 구성하는 것이다. 하지만, 데이터와 군집 사이에는 거리뿐만 아니라 방향도 존재하고 있다. 같은 군집에 속한 데이터들이 외부 군집에 대해서 같은 거리만큼 떨어져 있어도 방향이 동일할 것이라고 판단하기는 어렵다. 따라서 외부 군집과의 관계를 파악하기 위해서는 거리와 함께 방향도 고려해야 한다.

외부 군집 연관 기준을 구하는 방법은 <표1>과 같다.

<표 1> 외부 군집 연관 기준 계산

1단계: 군집 간 기준 벡터 $\widehat{\overrightarrow{C_iC_k}}$를 구한다.

$$\widehat{\overrightarrow{C_iC_k}} = \frac{\overrightarrow{C_iC_k}}{\| \overrightarrow{C_iC_k} \|} \quad (1)$$

여기서, $i,k(i \neq k)$는 군집을 의미하고, C_i는 i 군집의 중심 좌표를 의미한다.

2단계: i군집에 속한 데이터 x_j의 외부 군집 연관도(External Cluster Relationship) $\overrightarrow{ER_j}$ 를 구한다.

$$\overrightarrow{ER_j} = \sum_{k \in K, k \neq i} d_{jk} \widehat{\overrightarrow{C_iC_k}} \quad (2)$$

여기서, d_{jk}는 x_j와 C_k 사이의 거리로, 군집화를 수행할 때 계산되는 값이다.

3단계: 군집 i의 외부 군집 연관 기준 ER_i를 구한다.

$$ER_i = \sqrt{\frac{1}{N} \sum_{j \in w_i} d^2\left(\overrightarrow{ER_j}, \overrightarrow{ER_{w_i}}\right)} \quad (3)$$

여기서, w_i는 i 군집에 속한 데이터들이고, N은 i 군집에 속한 데이터의 개수이다. $\overrightarrow{ER_{w_i}}$는 i 군집에 속한 데이터들의 외부 군집 연관도의 평균이다. 즉, ER_i는 i 군집의 개별 외부 군집 연관도의 표준 편차를 의미한다.

4단계: 마지막으로 외부 군집 연관 기준 ER을 구한다.

$$ER = \frac{\sum_{k \in K} ER_k}{K} \quad (4)$$

그림 1은 외부 군집 연관도를 도식화 한 것이다. C_k는 군집 중심의 좌표를 의미하고, x_i은 군집 1에 속한 데이터의 좌표이다. a)는 군집 1의 기준 벡터를 구하는 방법이고, b)는 데이터 1의 외부 군집 연관도를 의미한다. 군집간 기준 벡터를 계산하여 군집들의 방향에 대한 기준을 계산할 수 있다. 다음으로 데이터에 대한 외부 군집 연관도를 계산하게 된다. 외부 군집 연관도는 데이터가 외부 군집에 주로 영향을 받는 방향을 의미하며, 먼저 데이터가 각 외부 군집에 영향 받는 방향을 계산한다. 이는 기준 벡터와 데이터와 군집의 거리를 곱하여 계산하게 된다. 각 군집과의 방향 벡터를 계산하면, 데이터의 외부 군집 연관도를 계산할 수 있으며, 이는 군집과의 방향 벡터의 벡터합으로 계산 된다.

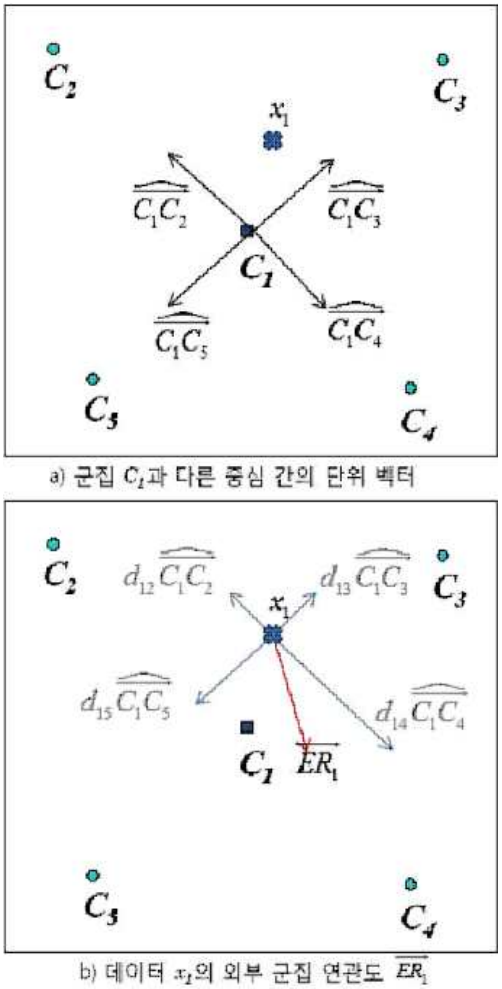

그림 1 외부 군집 연관의 개념도

3. 군집화 알고리즘과 외부 군집 연관 기준

3.1 군집화 알고리즘

대상들을 군집화 하는 방법은 매우 다양하지만 모든 방법이 공통적으로 가지고 있는 기본전제는 군집 내의 객체들 간의 유사성을 극대화하고, 군집간의 유사성은 극소화하는 것이다. 군집화 알고리즘에는 k-means 군집화 알고리즘과 같은 분할 기법(Partitional) 군집화 알고리즘과 계층적(Hierarchical) 군집화 알고리즘 등이 존재한다[18].

k-means의 수행 절차는 <표 2>와 같다[18].

<표 2> k-means 알고리즘

1단계:	데이터 집합에서 k 개의 초기 데이터를 선택한다. 이들은 하나의 원소로 이루어진 군집이 된다.
2단계:	남아있는 데이터를 가장 가까운 군집 중심이 있는 군집에 할당한다.
3단계:	군집 중심을 계산하고, 새로운 중심으로 설정한다.
4단계:	모든 데이터를 가장 가까운 군집 중심에 할당할 때까지 반복한다.

대부분의 k-means 수행 절차들은 군집 중심의 변화가 없을 때까지 반복적으로 군집 중심을 계산해야한다.

3.2 외부 군집 연관 기준과 결합

Pelleg와 Moore는 [5, 10]는 x-means라는 2분할 알고리즘을 제안했다. x-means는 군집의 추가 분할 여부를 평가할 때 BIC(Bayesian Information Criterion)을 사용하였다. BIC는 군집의 형태의 왜곡(distortion)을 평가하는 기준으로 본 논문에서 제안하는 기준과는 차이가 있다.

k-means의 시간 복잡도는 $O(kN)$이고, x-means의 시간 복잡도는 k-means의 2배의 시간이 걸린다[10]. 하지만, x-means는 k의 개수를 증가시키면서 실험을 하는 것이기 때문에, 이와 비교하면, 각 단계마다 수행해야 하는 k-means의 시간 복잡도는 $\sum_{k=2}^{K} O(kN)$로 x-means가 훨씬 빠르다.

제안하는 알고리즘은 <표 3>과 같다. 이 알고리즘은 x-means를 응용한 것으로 2 분할 k-means를 반복적으로 수행하게 된다. 2 분할 k-means의 수행 여부를 결정할 때 외부 군집 연관 기준(ECRC)을 기준으로 하며, 외부 군집 연관 기준(ECRC)이 가장 큰 군집에 대해서 2분할 k-means를 수행한다. 최종적으로 외부 군집 연관 기준(ECRC)의 최소값을 찾을 때까지 반복한다.

<표 3> k-means와 외부 군집 연관 기준의 결합

1단계:	$k=2(=k_0)$로 k-means 군집화 알고리즘을 수행한다. (초기 값을 늘려도 상관없다.) 군집의 이름은 $C_1, C_2, ..., C_{k_0}$ 으로 설정한다.
2단계:	수식 (3)과 (4)에 의해서 각 군집의 외부 군집 연관 기준 ER_i와 전체 외부 군집 연관 기준 ER을 계산한다.
3단계:	$\max(ER_{k_0})$인 군집 C_i를 선택해서 3단계부터 7단계를 반복 수행한다.
4단계:	군집 C_i에 대해서 $k=2$로 군집화를 수행한다. 생성된 군집의 이름은 $C_i^{(1)}, C_i^{(2)}$이다.
5단계:	각 군집의 외부 연관 기준 $ER_i^{(1)}, ER_i^{(2)}$와 전체 외부 연관 기준 ER'을 계산한다.
6단계:	$ER > ER'$이면, 군집 수행 결과를 확정하고, 인덱스들을 수정한다. $k_0 = k_0 + 1$ $ER = ER'$ $C_i = C_i^{(1)}$ $ER_i = ER_i^{(1)}$ $C_{k_0} = C_i^{(2)}$ $ER_{k_0} = ER_i^{(2)}$
7단계:	$ER \le ER'$면, 군집 수행 결과를 복원하고, 반복 수행 작업(iteration)을 종료한다.
8단계:	군집수 k_0, 각 군집 $C_1, C_2, ..., C_{k_0}$등을 출력하고 종료한다.

그림 2는 2차원 데이터에 제안하는 알고리즘을 적용한 결과를 보여준다. 4개의 Gaussian 분포의 결합에 의한 입력 데이터를 생성하였다. k 값을 2부터 실험하였고, 그림에는 3의 것을 보여준다. 실험 결과 외부 군집 연관 기준(ECRC)는 각각 25.7, 19.8, 24.6으로 $k=4$일 때 가장 작은 값을 보였다. 즉, 제안하는 알고리즘에 의해서 군집수는 4로 결정되었다.

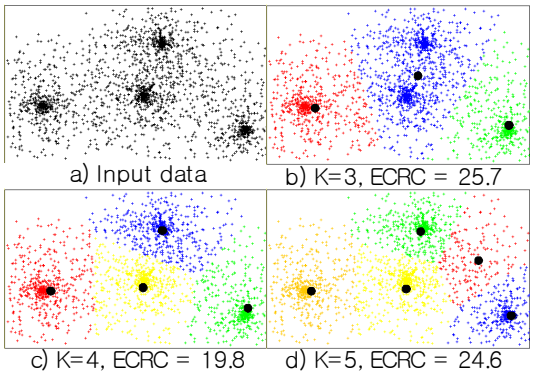

그림 2 제안하는 알고리즘에 의한 자동적인 군집수 선택 결과. a) 4개의 Gaussian 분포의 결합에 의한 입력 데이터 생성 b),c),d) K를 각각 3, 4, 5로 하고 수행한 군집화 결과.

4. 실험환경 및 결과

본 연구는 Pentium 4 2.8GHz CPU 시스템에서 C# 언어로 구현하였다. 사용된 컴파일러 버전은 .Net Framework 4 이다. 본 연구에서 제안하는 외부 군집 연관 기준(ECRC)을 이용한 군집화 방법과 BIC와 AIC를 이용한 x-means와 군집화 결과의 군집수를 비교하였다[10].

4.1 군집수에 대한 실험

데이터는 Gaussian 분포의 결합에 의하여 임의로 생성했다. 240개의 2차원의 정규 분포를 가지는 데이터이며, 4개의 군집을 형성한다. 각 군집은 60개의 요소를 지닌다. 초기에 $k_0=2$로 설정하고, 1,000번의 실험을 수행했다.

<표 4>는 실험에 의하여 생성된 군집수에 대한 요약이다. 1,000번의 실험 동안 가장 많은 경우는 4개의 군집이 생성되었을 경우이고, 479번 발생했다. 그 다음 선택된 군집수는 5개로 241번 발생했다. 두 번째 실험은 BIC를 이용한 x-means이다. ECRC를 이용해서 제안하는 방법과 유사한 경향을 보이고 있지만, 제안하는 방법이 BIC를 사용한 경우 보다 군집수가 4일 경우에는 약 10%, 군집수가 3~5일 경우에는 약 1% 정도 더 많은 실험 횟수를 보이고 있기 때문에 더 정확하게 군집수를 선택하고 있다.

<표 4> 표준 분포를 가지는 250개의 2차원 데이터에 대한 군집수

군집수	2	3	4	5
ECRC	5	184	479	241
x-means (BIC)	7	193	436	269
x-means (AIC)	6	237	329	226

군집수	6	7	8+	total
ECRC	62	23	6	1,000
x-means (BIC)	70	18	7	1,000
x-means (AIC)	106	60	36	1,000

마지막은 AIC를 이용한 x-means이다. AIC를 사용한 방법은 군집이 과생산되는 경향을 보이고 있는 것을 확인할 수 있다. BIC와 AIC는 유사한 정보 기준(Information Criteria)이다. 두 기준 모두 모델을 선택할 때 과적합(overfitting)되는 경향을 보일 수 있는데, 이는 파라메터의 숫자에 대한 벌칙(penalty)을 둠으로써 해결했다. 이 때 BIC가 AIC 보다 더 큰 벌칙을 가하게 되는데, 이에 의해서 차이가 발생된다[19].

4.2 실제 데이터 집합을 이용한 실험

이 실험은 다양한 실제 데이터 집합을 사용하여 이루어졌다. 데이터 집합은 Reuter-21578 문서 집합(Reuters)[20]과 UCI 기계 학습 Repository로부터 네 개의 데이터베이스를 선택하여 사용하였다[21]. 이 네 개의 데이터베이스는 각각 Australian Credit Approval (Australian), Pima Indians Diabetes Database (Diabetes), Heart disease dataset (Heart), Iris

Plants Database (Iris) 이다. <표 5>는 이 다섯 개의 데이터 집합에 대한 통계수치의 요약 자료이다.

<표 5> 실험에 사용된 데이터

	instances	features	clusters
Australian	690	14	2
Diabetes	768	8	2
Heart	270	13	2
Iris	150	4	3
Reuter	2,094	8,031	10

실험 결과는 <표 6>과 같다. 제안하는 외부 군집 연관 기준(ECRC)를 이용한 방법은 5개의 데이터 집합 중 Australian, Heart, Iris 3개의 데이터 집합의 군집수를 정확하게 선택하였으며, 다른 2개의 데이터 집합은 약간의 오차를 보였다. BIC는 Heart, Iris의 경우에 정확한 군집수를 선택했으며, 다른 3개의 경우는 오차를 보였다. AIC의 경우는 Iris의 경우에만 정확한 군집수를 선택했다. Reuter 데이터 집합을 살펴보면, 제안하는 ECRC는 9개로 1개를 덜 선택했고, BIC는 11개로 1개를 더 선택하여 두 방법 모두 10%의 오차를 보이고 있다. 반면에, AIC는 12개로 군집수를 2개 더 선택해서 20%의 오차를 보이는 것을 확인할 수 있다. Reuter의 경우 문서 데이터로 희박한(Sparse) 분포의 데이터 형태를 가지게 되며, 제안하는 방법이 BIC와 AIC 보다 희박한 분포의 데이터에 대해 우수한 성능을 보이고 있다.

<표 6> 다섯개 데이터 집합에 대한 실험 결과

	ECRC	BIC	AIC
Australian	2	3	3
Diabetes	3	3	3
Heart	2	2	3
Iris	3	3	3
Reuter	9	11	12

임의의 데이터와 실 데이터에 대한 실험결과를 정리하면, 제안하는 ECRC를 사용한 방법은 BIC를 사용한 x-means와 유사하거나 근소하게 우수한 결과를 보이고 있고, AIC를 사용한 x-means 보다는 더 좋은 결과를 보이는 것을 확인할 수 있었다.

5. 결 론

본 논문에서는 군집화를 수행할 때 어려운 문제 중의 하나인 군집수를 자동적으로 결정하기 위한 방법으로 외부 군집 연관 기준(ECRC)를 제안하였다. 외부 군집 연관 기준과 k-means를 변형한 x-means를 일부 수정한 방법과 결합한 군집화 알고리즘을 제안하였다. 군집수를 증가시킨 후 전체 데이터에 대해 k-means를 수행하는 것이 아니라 최대 외부 군집 연관 기준을 보이는 군집에 대해서만 부분 군집화를 수행함으로써 군집화 속도를 저하시키는 것을 줄일 수 있었다. 임의의 데이터와 실 데이터에 대한 실험결과 제안하는 방법은 최적의 군집 개수를 찾아주는 것을 확인할 수 있었다.

제안하는 방법은 특정 군집에 대하여 부분 군집화를 수행함으로써 군집수가 점점 증가하는 경향을 보이게 된다. 군집화를 수행할 때는 군집을 결합시키는 것도 하나의 방법이 될 수 있기 때문에 결합 방법에 대한 연구도 필요할 수 있다. 단, 이런 경우 수행 시간이 증가될 수 있고, 군집의 분할과 결합 방법을 어떻게 조화시킬 것이며, 해당 반복 수행 작업에 대한 종료 조건 등에 대한 연구가 필요하다.

참 고 문 헌

[1] R. O. Duda, P. E. Hart and Da. G. Stork, “Pattern Classification (2nd Edition)”, Wiley-Interscience, Oct., 2000.

[2] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parkankangas, "Self-Organizing Map in Matlab: the SOM Toolbox", Proceedings of the Matlab DSP Conference, pp. 34-40, 1999.

[3] M. H. Yang and N. Ahuja, "A Data Partition Method for Parallel Self-Organizing Map", Proceeding of the IJCNN 99, pp. 1929-1933, 1999.

[4] Z. Huang, "Extension to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery, Vol 2, pp. 283-304, 1998.

[5] D. Pelleg and A. Moore, "Accelerating Exact K-means Algorithms with Geometric Reasoning", International Conference on Knowledge Discovery and Data

mining '99, pp. 277-281, 1999.
[6] A. K. Jain, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, Vol. 31, pp. 651-666, 2010.
[7] M. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models", IEEE transactions on pattern analysis and machine intelligence, Vol. 24, pp. 381-396, 2002.
[8] R. Tibshirani, G. Walther and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic", Journal of the royal statistical society, Vol. 63, pp. 411-423, 2001.
[9] C. Rasmussen, "The infinite gaussian mixture model", Advances in neural information processing systems, Vol. 12, pp. 554-560, 2000.
[10] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters", In Proc. of the Seventeenth International Conference on Machine Learning (ICML2000), June, pp. 727-734, 2000.
[11] S. Salvador and P. Chan, "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms", In Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence, Nov., pp. 576-584, 2004.
[12] W. Lu and I. Traore, "Determining the optimal number of clusters using a new evolutionary algorithm", In Proc. Of the 17th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 05), Nov., 2 pp., 2005.
[13] B. Boutsinas, D. K. Tasoulis and M. N. Vrahatis, "Estimating the number of clusters using a windowing technique", Journal of Pattern Recognition an Image Analysis, Vol. 16, No. 2, April, pp. 143-154, 2006.
[14] 지태창, 이현진, 이일병, "온라인 문서 군집화에서 군집 수 결정 방법", 정보처리학회지, Vol. 117, pp. 513-522, 2007.
[15] O. Satoshi and T. Katsumi, "How Many Objects?: Determining the Number of Clusters with a Skewed Distribution", Proceeding of the 18th European Conference on Artificial Intelligence, pp. 771-772, 2008.
[16] R. V. Ranga, "Incremental Clustering Algorithm for Earth Science Data Mining", Proceeding of the 9th International Conference on Computational Science, pp. 375-384, 2009.
[17] A. J. Graaff and A. P. Engelbrecht, "Using sequential deviation to dynamically determine the number of clusters found by a local network neighbourhood artificial immune system", Journal of Applied Soft Computing archive, Vol. 11, pp. 2698-2713, 2011.
[18] Earl Gose, Richard Johnsonbugh and Steve Jost, "Pattern Recognition and Image Analysis", Prentice Hall, 1996.
[19] Y. Yang, "Can the strength of AIC and BIC be shared?", Biometrika, Vol. 92, pp. 937-950, 2005.
[20] D. D. Lewis, "Reuters-21578 text categorization test collection distribution 1.0", http://www.research.att.com/~lewis, 1999.
[21] S. Hettich and S. D. Bay, "The UCI KDD Archive [http://kdd.ics.uci.edu]", Irvine, CA: University of California, Department of Information and Computer Science, 1999.

이현진

1996년: 순천향대학교 전산학과 학사
1998년: 연세대학교 대학원 컴퓨터과학 석사
2002년: 연세대학교 대학원 컴퓨터과학 박사
2003년~현재: 한국사이버대학교 컴퓨터정보통신학과 부교수
관심분야 : 이러닝, 기계학습, 데이터마이닝

지태창

1997년: 연세대학교 컴퓨터과학과학사
1999년: 연세대학교 컴퓨터과학과석사
2004년 ~ 현재 연세대학교 컴퓨터과학과 박사과정
1999년 ~현재: LG CNS 책임 연구원
관심분야 : 패턴인식, 데이터마이닝, 스마트폰 기술