Journal of Digital Contents Society Vol. 16 No. 3 Jun. 2015(pp. 373-379)
http://dx.doi.org/10.9728/dcs.2015.16.3.373

# 삼각 부등식을 이용한 온라인 VQ 코드북 생성 방법

이현진*

## 요 약

본 논문에서는 실시간으로, 문서, 웹 페이지, 블로그, tweet 등 텍스트 정보와 센서, 머신데이터등 IoT의 데이터가 생성되는 상황에서 새로 추가되는 데이터들을 기존에 만들어진 VQ 코드북에 추가시키면서, 기존 VQ 코드북 모델을 실시간으로 갱신하기 위한 온라인 VQ 코드북 생성 방법을 제안한다. 기존에 일괄 작업으로 만들어진 VQ 코드북의 성능을 저하시키지 않으면서, 새로 추가된 데이터를 활용하여 VQ 코드북을 점진적으로 수정하는 방식으로 삼각 부등식을 활용하여 높은 정확도와 속도를 보일 수 있었다. 테스트 데이터에 적용한 결과 일괄 작업과 유사한 성능을 보이면서, 다른 온라인 K-Means 보다 빠른 속도를 보였다.

**키워드 : 실시간, 온라인, VQ 코드북 생성, 군집화, 점진적 학습**

# Online VQ Codebook Generation using a Triangle Inequality

Hyunjin Lee*

## Abstract

In this paper, we propose an online VQ Codebook generation method for updating an existing VQ Codebook in real-time and adding to an existing cluster with newly created text data which are news paper, web pages, blogs, tweets and IoT data like sensor, machine. Without degrading the performance of the batch VQ Codebook to the existing data, it was able to take advantage of the newly added data by using a triangle inequality which modifying the VQ Codebook progressively show a high degree of accuracy and speed. The result of applying to test data showed that the performance is similar to the batch method.

**Keywords : Real-Time, Online, VQ Codebook Generation, Clustering, Incremental Learning**

## 1. 서 론

벡터 양자화 (Vector Quantization, VQ)는 정보 탐색과 지식 분류, 시각화 분야등에서 활용되고 있는 데이터마이닝 방법 중 하나이다[1, 2]. 벡터 양자화는 데이터의 구조를 분석을 위해 데이터의 요소들을 유사한 그룹으로 나누는 방법이다[3]. 벡터 양자화는 단순하고, 쉽게 적용할 수 있기 때문에 패턴 인식, 이미지 압축, 웹 마이닝, 음성 인식, 얼굴 인식 등 다양한 분야에서 사용되고 있다[4, 5, 6].

Generalize Lloyd Algorithms (GLA)라고도 불리는 K-Means 알고리즘은 가장 많이 사용되는 VQ 코드북 생성 알고리즘중 하나이다[7]. GLA는 $d$-차원 공간의 원소들을 입력 데이터로 사용하여 속성이 유사한 데이터들을 $k$개의 군집으로 구분하는 것이 목표이다. GLA는 사용하기 쉽고, 효율적이라는 장점이 있지만, 군집의 개수 $k$를 미리 설정해야 하고, 대규모 데이터를 처리에는 비효율적이라는 단점을 가지고 있다.

삼각 부등식을 활용한 VQ 코드북 관련 연구는 다음과 같다. 김성재 등은 벡터 양자화를 빠

※ Corresponding Author : Hyunjin Lee
Received: April 28, 2015
Revised: May 17, 2015
Accepted: May 30, 2015
* Dept. of Computer Science & Software, Korea Soongsil Cyber University
Tel: +82-2-708-7863 , Fax: +82-2-708-7749
email: hjlee@mail.kcu.ac

르게 수행하기 위한 삼각 부등식을 이용한 빠른 VQ 코드북 탐색법을 제안했다[8]. 이현진은 VQ 코드북을 빠르게 생성하기 위하여 삼각 부등식을 이용한 빠른 VQ 코드북 생성 방법을 제안했다[9].

기계학습 알고리즘들 중 일괄처리 방식들은 학습이 이루어지기 전에 확정된 데이터에 대해 적용할 수 있는 방법이다. 하지만 현실세계에서는 데이터는 끊임없이 생성되고, 갱신되고, 삭제된다. 따라서 데이터가 변화될 때 변화를 반영할 수 있는 온라인 군집화 방식이 연구되고 있으며, VQ 코드북 생성에 있어서도 데이터 변화를 반영할 수 있도록 생성된 VQ 코드북의 정보를 온라인으로 업데이트하는 방법들이 연구되고 있다.

Fokatis 등은 트리(Tree)구조를 활용한 Online Sum-Radii Clustering을 제안했다[10]. Barbakh 등은 온라인 군집화 시 GLA의 초기값에 따른 민감도를 극복하기 위한 방안을 제시했다[11]. Aggarwal 등은 다양한 Stream 군집화 알고리즘의 동향을 조사, 정리하였다[12]. 한경아 등은 모바일 사용자들에 대한 온라인 군집화에 대한 연구를 수행했다[13]. King은 온라인 GLA의 중심 갱신 규칙의 변화량을 휴리스틱하게 조절하는 알고리즘을 제안했다[14]

본 논문에서는 군집의 중심에 가까운 데이터는 군집의 코드북 갱신에 반영하고, 군집의 중심에서 먼 데이터는 군집의 코드북 갱신에 영향을 주지 않도록 군집과 군집 사이에 존재하는 데이터에 대한 거리 계산과 업데이트를 수행한다. 삼각 부등식을 활용한 GLA로 생성된 군집에 적용하여 군집의 성능을 높이고, 생성속도를 개선하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 삼각부등식을 이용한 코드북 생성 알고리즘을 설명하고, 3장에서는 제안하는 온라인 코드북 생성 알고리즘을 살펴본다. 4장에서는 합성 데이터를 사용하여 수행을 한 실험 결과를 분석하고 5장에서 결론을 맺는다.

## 2. 삼각 부등식을 이용한 VQ 코드북 생성 알고리즘

GLA의 속도를 향상시키는 알고리즘은 삼각 부등식에 기반을 두고 있다. $x$, $y$, $z$의 세 점이 있을 때 $d(x,z) \leq d(x,y) + d(y,z)$이다. 이 수식은 모든 거리 계산 방법인 $d$에 상관없이 성립한다[8].

**Lemma 1**: $x$가 데이터의 한 점이고, $b$와 $c$가 코드워드일 때, $d(b,c) \geq 2d(x,b)$ 면, $d(x,c) \geq d(x,b)$ 이다[15].

**Lemma 2**: $x$가 데이터의 한 점이고, $b$와 $c$가 코드워드이면, $d(x,c) \geq \text{abs}(d(x,b) - d(b,c))$이다.

이를 이용한 벡터 양자화 코드북 생성 알고리즘은 다음과 같다[8].

먼저, 임의로 초기 코드북 $CB_0$를 생성한다. i = 1 로 초기화 한다. 모든 입력 데이터 $x$를 가장 가까운 코드워드 $c(x) = \text{argmin}_c d(x,c)$ 로 할당한다. Lemma 1을 사용하여 불필요한 계산을 줄인다.

다음으로 아래 과정을 수렴할 때 까지 반복한다.

1) 코드워드에 속한 데이터들의 평균을 구하여 새로운 코드북 $CB_i$를 계산한다.
2) 코드북 $CB_i$의 모든 코드워드들인 $c$, $c'$에 대하여 둘 사이의 거리 $d(c,c')$을 계산한다. 모든 코드워드 $c$에 대하여 수식 (1)과 같이 $s(c)$를 계산한다.

$$s(c) = \frac{1}{2} min_{c' \neq c} d(c,c') \qquad (1)$$

3) 코드북 $CB_i$의 코드워드 $c$와 이전 단계의 코드북 $CB_{i-1}$의 코드워드 $c^*$ 사이의 거리 $d(c,c^*)$ 를 계산한다.
4) 모든 코드워드 $c$에 대하여, 입력데이터와의 거리 $d(x,c)$를 계산한다. Lemma 2에 의하여 $x$가 이전 단계의 코드워드 $c^*$에 속해 있었다면, 수식 2와 같이

$$d(x,c) = d(x,c^*) + d(c,c^*) \qquad (2)$$

이고, 이전 단계의 코드워드 $c^*$에 속해 있지 않다면, 수식 3과 같이

$$d(x,c) = \text{abs}(d(x,c^*) - d(c,c^*)) \qquad (3)$$

이다.
5) 모든 입력 데이터 $x$에 대하여 가장 가까운 코드워드를 할당한다. 새로 할당된 코드워드 $c(x)$가 이전 단계의 코드워드 $c^*(x)$와 같은 모든 입력 데이터를 찾아내어 새로운 코드워

드와 거리 계산을 줄인다.

6) 남아 있는 입력 데이터 $x$에 대하여 $d(x,c) \le s(c(x))$인 입력 데이터 $x$를 찾는다.
7) 모든 남아 있는 입력 데이터 $x$와 코드워드 $c$에 대하여, 즉,

$$c(x) \neq c^*(x) \quad (4)$$

$$d(x,c) \ge d(c(x), c)/2 \quad (5)$$

일 때, 입력 데이터 $x$와 코드워드 사이의 거리 $d(x,c)$를 계산하고, 가장 가까운 코드워드 $c(x)$에 할당한다.
8) 코드북 $CB_i$를 평가하여 이전 단계와 비교하여 결과가 좋으면 멈추고, 그렇지 않으면, $i = i + 1$을 수행한다.

## 3. 온라인 VQ 코드북 생성 방법

온라인 환경은 이론적으로 데이터의 제한이 없기 때문에, 모든 데이터를 저장하고 있는 것이 사실상 불가능하다. 따라서, 한 시점에서 다른 시점으로 넘어갈 때 보통 군집의 중심에 대한 정보만을 저장한다. King이 제안하는 온라인 GLA 알고리즘은 <표 1>과 같다[14].

<표 1> King의 온라인 GLA 알고리즘

initialize the $k$ cluster centers $c_1,\ldots,c_k$ in any way
crete counters $n_1,\ldots,n_k$ and initialize them to zero
loop
  get new data point $x$
  determine the closest center $c_i$ to $x$
  update the number of point in that cluster : $n_i \leftarrow n_i + 1$
  update the cluster center : $c_i \leftarrow c_i + \alpha(x - c_i)$ for $\alpha \in (0,1)$
end loop

<Table 1> King's Online GLA Algorithm

온라인 코드북 생성을 위한 일괄 알고리즘에서 추가로 저장해야 하는 값은 코드북의 크기 $n(c)$와 코드북에 속한 데이터의 최대거리인 $max(c)$ 이다. 온라인 VQ 코드북 생성 알고리즘은 <표 2>와 같다.

<표 2> 온라인 VQ 코드북 생성 알고리즘

1) 먼저, 새로운 데이터 $x$와 코드워드 사이의 거리 $d(x,c)$를 계산하고, 가장 가까운 코드워드 $c(x)$에 할당한다. 만약, $d(x,c)$가 $s(c)$ 보다 크면, 노이즈 데이터로 코드북 갱신을 수행하지 않고, 종료한다.
2) 새로운 데이터가 속한 코드워드 $c$에 대하여 새로운 코드북 $CB'(c)$를 수식 6과 같이 계산한다.

$$CB(c') = \frac{CB(c) \times n(c) + x}{n(c) + 1} \quad (6)$$

3) 새로운 코드북 $CB(c')$와 기존 코드북 $CB(c)$ 와의 차인 $var$와 새로운 데이터 $x$와 새로운 코드북 까지의 거리 $d(x,c')$을 수식 7과 같이 계산한다.

$$var = d(BC(c), BC(c')) \quad (7)$$

4) 새로운 코드북 $CB(c')$에 속한 데이터의 최대 거리를 삼각 부등식에 의해 수식 (8)과 같이 추정한다.

$$max(c') = max(max(c) + var, d(x,c')) \quad (8)$$

5) 코드북 사이의 거리인 $s(c')$를 계산한다.

$$s(c') = \frac{1}{2} min_{c^* \neq c'} d(c', c^*) \quad (9)$$

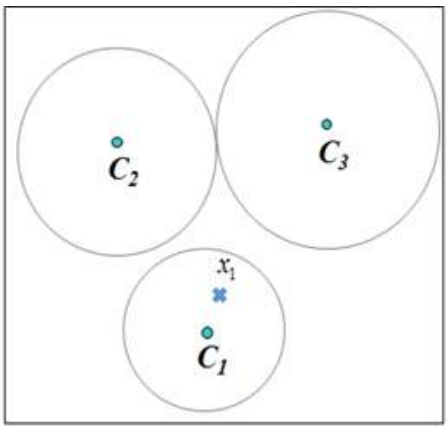
6) $max(c')$이 $s(c')$보다 작으면 새로운 데이터에 대한 코드북 적용을 승인한다.
7) $max(c')$이 $s(c')$보다 크면, 새로운 데이터에 대한 코드북 적용을 취소하고, 다른 코드워드들에 대하여 2)~7)을 반복한다.
8) 모든 코드워드에 대하여 최적의 코드워드가 할당되지 않을 경우 해당 데이터를 가장 가까운 코드워드에 할당하지만, 코드북에 대한 갱신은 하지 않고, 종료한다.

<Table 2> Online VQ Codebook Generation Algorithm

(그림 1,2,3)은 새로운 데이터 $x_1$이 기존 군집 $C_1$, $C_2$, $C_3$ 사이에 배치되는 경우에 대해서 제안하는 알고리즘의 처리방식을 도식화 한 것이

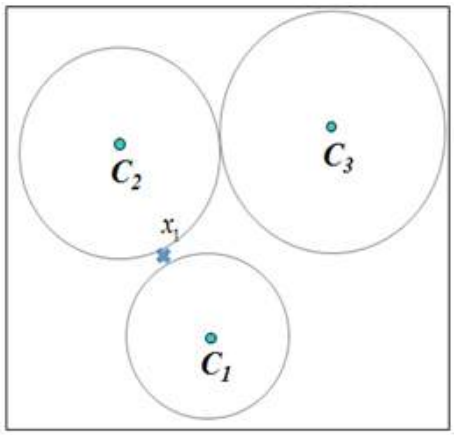
다. (그림 1)의 경우 새로운 데이터 $x_1$이 기존 군집 중 $C_1$의 범위 안에 포함되어 있다. 이 경우 제안하는 알고리즘에 의하여 코드북에 대한 업그레이드를 진행하여, 새로운 코드북을 생성하게 된다.

(그림 1) 새로운 데이터 $x_1$이 군집 $C_1$에 포함될 경우




(Figure 1) The case when new data $x_1$ is included in cluster $C_1$

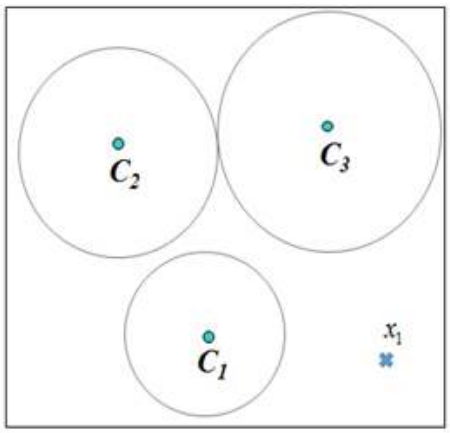
(그림 2) 새로운 데이터 $x_1$이 군집 $C_1$과 $C_2$사이에 있을 경우




(Figure 2) The case when new data $x_1$ is located between cluster $C_1$ and $C_2$

(그림 2)의 경우 새로운 데이터 $x_1$은 기존 군집 $C_1, C_2$ 사이에 존재하고 있다. 일괄 모드의 경우에는 경계선에 존재하는 데이터도 반복 과정을 거쳐서 최적의 코드북을 생성할 수 있지만, 온라인 환경의 경우 반복 과정이 존재하지 않아, 오류가 발생시 오류를 보정이 어렵다. 따라서, 기존 코드북의 성능을 저하시킬 요인은 코드북 갱신에서 배재한다. (그림 3)의 경우와 같이 기존 군집과 다른 성향을 보이는 새로운 데이터 $x_1$은 기존 코드북의 갱신에서 배재한다.

(그림 3) 새로운 데이터 $x_1$이 기존 군집과 다른 경우




(Figure 3) The case when new data $x_1$ is different from other clusters

제안하는 온라인 방법은 군집의 경계에 있어서 소속 군집을 결정하기 모호한 경우(일괄 작업하에서도 진동하는 데이터 형태)와 기존 군집과 크게 떨어져 있어서 다른 형태인 경우(수가 작은 경우는 노이즈)는 코드북 갱신에서 배재하여 코드북의 안정성을 고려하였다.

## 4. 실험환경 및 결과

본 논문에서는 두 개의 데이터 집합을 이용하였다. 기계학습(Machine Learning)에서 잘 알려진 데이터인 iris 데이터와 covtype 데이터에 time series 속성을 부여하였다[16]. 이렇게 수정된 iris 데이터는 4차원에 3개의 군집을 가지며, 군집은 지속적으로 변화하도록 하였다. covtype

은 54개의 차원의 7개의 군집을 가지며, 마찬가지로 군집은 지속적으로 변화 하도록 하였다. 원 데이터의 크기는 각각 150개, 581,012개 이며 반복적으로 데이터를 적용하였다.

성능 비교를 위하여 일괄 학습 방법인 GLA와 King의 Online K-means[14]과 제안하는 방법을 비교하였다. 성능 비교는 전통적인 코스트(cost) 값을 수정한 King의 방법을 사용하였다[14]. 일괄 k-means의 코스트 함수는 수식 10과 같다.

$$Cost(CB_1,...,CB_k,c_1,...,c_k) = \sum_{k}\sum_{i:\, x_i \in CB_k} \| x_i - c_k \|_2^2 \quad (10)$$

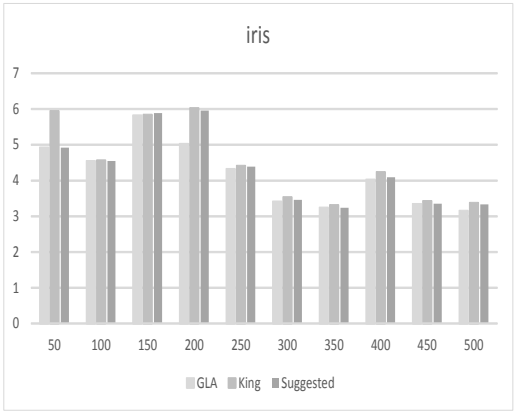
여기서, $CB_1,...,CB_k$는 $k$개의 군집이고, $c_1,...,c_k$는 군집의 중심이다. 온라인 환경에서는 시간의 변화에 따른 모든 데이터를 고려해야 한다. 특정 시점 $t$에 군집의 중심은 $c_1^t,...,c_k^t$이고, 각 군집은 $CB_1^t,...,CB_k^t$이다. 따라서, 온라인 코스트 함수는 수식 11과 같다. 코스트는 낮을수록 더 좋은 결과를 보인다.

$$Cost(CB_1^t,...,CB_k^t,c_1^t,...,c_k^t) = \sum_{k}\sum_{i:\, x_i \in CB_k^t} \| x_i - c_k^t \|_2^2 \quad (11)$$

(그림 4)는 iris 데이터에 대한 실험 결과이고 (그림 5)는 covtype 데이터에 대한 실험 결과이다. 그래프에서 x축은 데이터의 양이고, y축은 코스트 값이다.

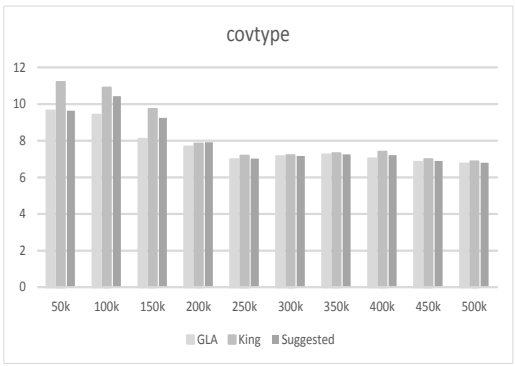
데이터가 실시간으로 증가할수록 일괄 학습 방법인 GLA보다 성능이 떨어지지만, 온라인 GLA 방법인 King의 결과보다는 우수한 결과를 보이는 것을 알 수 있었다. 제안하는 온라인 방법은 기존 군집을 벗어나게 되는 노이즈 데이터는 판단을 유보하고, 군집간의 경계에 속하는 모호한 데이터에 대해서는 코드북 갱신을에서 배재하여 King의 방법보다 우수한 결과를 얻을 수 있었다.

(그림 4) iris 데이터에 대한 실험 결과




(Figure 4) Experimental Result of iris

(그림 5) covtype 데이터에 대한 실험 결과




(Figure 5) Experimental Result of covtype

$k$는 군집수, $n$은 데이터수, $l$은 반복회수라고 했을 때 GLA의 복잡도는 $O(knl)$이고, King의 방법과 제안하는 방법은 $O(k)$이다. GLA는 데이터의 크기가 증가 할수록 전체 데이터를 사용하여 일괄 작업을 수행하기 때문에 수행 시간이 기하급수적으로 증가한다. 하지만 King의 방법과 제안하는 방법은 데이터 크기가 증가해도 증가된 데이터만을 사용하여 학습을 수행하였기 때문에 온라인 환경에서 군집에 학습을 실시간으로 할 수 있는 상수 값의 수행시간을 보인다.

## 5. 결 론

본 논문에서는 새로운 데이터가 유입되는 환경에서 VQ 코드북의 성능을 향상시키는 방법에

대하여 연구하였다. 기존 군집의 체계 변경에 크게 영향을 끼칠 수 있는 데이터를 정의하고, 이 데이터에 처리 방안을 제시하였다. 기존 군집의 구조를 군집의 중심 (코드워드)과 군집의 경계로 정의하고, 코드워드의 변화에 따른 군집 경계의 이동을 삼각부등식을 이용하여 예측하였다. 군집 간의 경계에 속하는 데이터는 모호한 데이터로 정의하고, 기존 군집들의 경계를 벗어나는 데이터는 노이즈 데이터로 정의하였다. 온라인 군집화의 성능을 높이기 위하여 모호한 데이터와 노이즈 데이터에 대한 처리 방안을 제시하여 실시간 학습이 이루어지면서 좋은 성능을 보일 수 있었다.

향후, 노이즈 데이터의 처리에 대한 연구를 더 수행할 필요가 있다. 노이즈 데이터의 개수가 적을 때에 새로운 코드북에 할당하는 것은 큰 의미가 없지만, 노이즈 데이터가 충분히 많이 쌓이면, 해당 데이터에 대해서만 새로운 코드북에 할당하는 방법에 대한 연구가 필요하다. 이를 위해 코드북의 개수가 고정적이지 않고 변화되는 방법에 대한 연구와 기존 코드북에 큰 영향을 끼치지 않으면서 새로운 코드북을 생성하는 방안에 대한 연구가 필요하다.

## References

[1] K. Cox, S. Hibino, L. Hong, A. Mockus and G. Wills, “A Multi-Modal Natural Language Interface to an Information Visualization Environment,” International Journal of Speech Technology, Vol. 4, No. 3, pp. 297-314, 2001.

[2] T. Li, S. Feng and L.X. Li, “Information Visualization for Intelligent Decision Support Systems,” Knowledge-Based Systems, Vol. 14, pp. 259-262, 2001.

[3] G.S. Linoff and M.J. Berry, “Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management”, Wiley Computer Publishing, New York, 2011.

[4] C. W. Tsai, C. Y. Lee, M. C. Chiang, and C. S. Yang, “A Fast VQ Codebook Generation Algorithm via Pattern Reduction, Pattern Recognition Letters”, vol. 30, pp. 653-660, 2009.

[5] Xiao-Gang W, and Yue L, “Web mining based on user access patterns for web personalization,” ISECS International Colloquium on Computing, Communication, Control, and Management. 1: 194-197, 2009.

[6] Hyunjin Lee, “Decombined Distributed Parallel VQ Codebook Generation Based on MapReduce,” Journal of Digital Contents Society, Vol. 15, No.3, pp.365-371, 2014.

[7] Krishnamoorthy R, Kalpana J, “Minimum distortion clustering technique for orthogonal polynomials transform vector quantizer,” Proc. 2011 Inter. Conf. Communication, Computing & Security. pp.443-448, 2011.

[8] S.J. Kim, C.W. Ahn, S.H. Kim, “Fast Codebook Search Method using Triangle Inequality for Vector Quantization,” Proceedings of the Korean Information Science Society Conference , Vol.25 No.2 (2), pp.526-528, 1998.

[9] Hyunjin Lee, “An Efficient Vector Quantization Codebook generation using a Triangle Inequality,” Journal of Digital Contents Society, Vol. 13, No.3, pp.309-315, 2012.

[10] D. Fotakis and P. Koutris, “Online Sum-Radii Clustering,” Theoretical Computer Science, Vol. 540-541, pp. 27-39, 2014.

[11] W. Barbakh and C. Fyfe, “Online Clutering Algorithms,” International Journal of Neural Systems(IJNS), Vol. 18, No. 3, pp. 1-10, 2008.

[12] C.C. Aggarwal and I.K. Chandan, “Data Clustering: algorithms and applications,” CRC Press, 2013.

[13] J.H, Yoo, K.A. Han, D.H. Jeong, H.J. Lee “Cluster-Based Routing Mechanism for Efficient Data Delivery to Group Mobile Users in Wireless Ad-Hoc Networks,” The Journal of Korea Information and Communications Society,Vol. 38C, No.11, pp. 323-324, 2013.

[14] A. King, “Online k-Means Clustering of Nonstationary Data,” Prediction Project Report, 2012.

[15] Charles Elkan, “Using the Triangle Inequality to Accelerate k-Means,” Proceedings of the Twentieth International Conference on Machine Learning, 147-153, 2003.

[16] A.W. Moore, “The anchors hierarchy: Using the triangle inequality to survive high dimensional data,” Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence, pp. 397-405, 2000.

**이 현 진**

1996년: 순천향대학교 전산학과 공학사
1998년: 연세대학교 대학원 컴퓨터과학과 공학석사
2002년: 연세대학교 대학원 컴퓨터과학과 공학박사
2003년~현재: 숭실사이버대학교 컴퓨터소프트웨어학과 부교수
관심분야 : 이러닝, 기계학습, 빅데이터