논문 19-44-01-18 The Journal of Korean Institute of Communications and Information Sciences '19-01 Vol.44 No.01
https://doi.org/10.7840/kics.2019.44.1.124

Fuzzy C-Means 클러스터링 잡음 처리 방법 연구

이 광 규◆, 우 정 현°

A Study of Fuzzy C-Means Clustering Noise Processing Method

Kwang-Kyu Lee◆, Jung-Hyun Woo°

요 약

FCM(Fuzzy C-Means)알고리즘은 반복 최적화 기법을 통해 최적해를 찾는다. 특히, 클러스터링 초기 중심과 잡음(noise)의 위치, 개수에 따라 실행시간 차이가 난다. 본 논문에서는 FCM의 전처리 과정으로 실행속도를 줄이기 위한 들로네 삼각-FCM(Delaunay Triangulation-FCM:DT-FCM)잡음 제거 방법을 제안한다. 들로네 삼각형은 평면 위의 점들을 삼각형으로 연결하여 공간을 분할할 때, 삼각형들의 내각의 최소값이 최대가 되도록 하는 분할이다. 이 방법은 FCM으로 클러스터링된 점들 중에서 가장 저밀도의 점들을 외접원의 들로네 삼각형을 만들어 잡음을 제거하는 효과적인 방법이다. 실험 결과 제안된 방법이 기존 FCM보다 실행시간이 감소됨을 보였다.

Key Words : Fuzzy C-Means, FCM, Clustering, Data Mining, Delaunay Triangulation-FCM(DT-FCM)

ABSTRACT

FCM(Fuzzy C-Means) algorithm finds the optimal solution through iterative optimization technique. In particular, the execution time differs depending on the initial center of clustering and the location and number of noise. In this paper, we propose a Delaunay Triangular-FCM(DT-FCM) noise canceling method to reduce the execution speed of FCM preprocessing. The triangle triangles are those that divide the space by connecting the points on the plane with triangles so that the minimum value of the interior angle of the triangles is the maximum. Experimental results show that the proposed method reduces execution time compared to existing FCM.

Ⅰ. 서 론

클러스터링은 주어진 데이터의 유사성을 기준으로 클러스터링을 형성하는 무감독 학습 방법(unsupervised learning)중 하나이다[8]. 퍼지 클러스터링(fuzzy clustering) 알고리즘은 오랜 역사를 가지고 있으며,Fuzzy C-Means(FCM)은 유클리디안 공간상에서 거리를 이용하여 퍼지 멤버쉽을 할당함으로서 클러스터링을 수행한다[11]. FCM 역시 유클리드 거리를 사용하므로 잡음에 민감하며 이를 해결하기 위한 방법은 크게 비유클리드 거리를 사용하는 방법과 유클리드 거리를 사용하면서 FCM을 변형하는 두 가지가 있다[4]. 개발된 여러 가지 알고리즘은 FCM의 잡음 문제를 해결하기 위하여 PCM(Probabilistic C-Means), FPCM(Fuzzy Probabilistic C-Means)[9], PFCM (Probabilistic FCM)[8] 등으로 발전하였다. 하지만 이 알고리즘들은 효과적인 잡음 제거에 대처하지 못하였으며, 초기화 문제, 계산량의 증가, 파라미터들의 증가

◆ First Author : Shinhan University Department of IT Convergence, kklee@shinhan.ac.kr, 정회원
° Corresponding Author : Yuhan University Department of Information Security, wjhsun03@naver.com, 정회원
논문번호 : 201809-285-C-RN, Received September 19, 2018; Revised December 19, 2018; Accepted December 27, 2018

로 실제적인 클러스터링 과정에 이용하기가 불편해졌다. 특히, 클러스터 주변의 애매한 경계점이나 클러스터에 속하지 않는 잡음은 실행시간에 많은 영향을 미친다. 잡음은 클러스터에 속하지 않는 점으로 측정 오류의 임의적 성분이다. 잡음은 값의 왜곡이나 가짜 객체의 추가와 관련되므로 CPU 실행시간의 상당부분 지연을 초래한다. 그러므로 잡음의 제거는 매우 중요한 사항이며, 데이터 마이닝에서는 상당량의 작업이 잡음의 존재 유무 검사와 제거하는데 치중하고 있다[7]. 따라서 본 논문에서는 들로네 삼각형을 이용한 효과적인 잡음 제거를 위한 DT-FCM 클러스터링 잡음 제거 중심값 설정 방법을 제안한다[2,10,12]. 제안 방법은 임의의 세 점을 기준으로 최대한 정삼각형을 만들어 가는 들로네 삼각형을 이용한다. 만들어진 삼각형에 속하는 일정 길이 이상의 긴 변들을 제거하고, 남은 연결선 성분들로 외접원을 구성한다. 그리고 최근접 이웃방식으로 점을 연결해가면 가능한 최대 내각의 형태가 좋은 삼각형을 만들어 낸다. 이와 같은 방법을 반복해가는 들로네 삼각형은 3차원 이상의 고차원으로도 확장이 가능하다[1,6].

본 논문의 구성은 다음과 같다. 2장에서는 FCM에 대해 설명한다. 3장에서는 들로네 삼각형을 이용한 DT-FCM 클러스터 잡음 처리 중심값 설정 방법을 제안한다. 4장에서는 실험 결과를 통해 제안 방법의 효율성을 보이고, 결론 및 향후 연구 방향은 5장에서 언급한다.

Ⅱ. Fuzzy C-Means(FCM)

일반적으로 퍼지 클러스터링은 데이터를 그룹화하기 위한 무감독 학습전략으로 사용되지만, 데이터로부터 퍼지 규칙(if ~then ~)을 생성할 때에도 유용하게 사용된다[3]. 퍼지 규칙의 구조는 사용한 데이터의 특성에 따라 달라진다. 가령, 고장 진단이나 패턴분류에서는 데이터가 어떤 클러스터링으로 분류되어야 하는지를 결정하도록 퍼지 규칙이 만들어지고 퍼지 제어, 시스템 인식이나 함수 근사화에서는 입출력 변수들 사이의 연속적인 관계를 기술할 수 있도록 만들어진다. 이미지 분석이나 인식에서는 원이나 타원과 같은 공간상의 기하학적 형체를 감지하고 분리하는데 사용되며, 이를 쉘 클러스터링 알고리즘이라 부른다. 퍼지 클러스터링의 목적은 임의의 데이터 집합을 특정한 수의 퍼지 클러스터링으로 분할하는 것이다. 현재까지 가장 널리 사용되고 있는 퍼지 클러스터링 방법은 FCM 알고리즘이다. 이 알고리즘은 n개의 항목으로 구성된 데이터 집합 X={$X_1, X_2, \cdots, X_n$}을 c개의 퍼지 클러스터링으로 분할하고자 할 때 식(1)의 목적함수를 최소화하는 퍼지 분할 $\hat{F}=\{\hat{F}_1, \hat{F}_2, \ldots \hat{F}_c\}$ 를 찾는 것이 목적이다. 하드 클러스터링과는 매개변수 m과 퍼지 분할 행렬의 값에 차이가 있다. 분할의 애매함(fuzziness)을 표시하는 매개변수인 m은 1보다 큰 값을 가질 수 있는데, 이 값이 클수록 모호함이 높아진다(m이 1인 경우는 애매하지 않는 명확한 하드 클러스터링을 표시). m의 값은 일반적으로 1.25나 2가 좋은 결과를 제공하는 것으로 알려져 있지만, 응용 분야에 따라 적합한 m의 값을 선택할 수 있다. μ_{ik}는 데이터 X_k가 퍼지 클러스터링에 대하여 속하는 정도를 나타내며, $(c \times n)$ 크기의 퍼지 분할 행렬 U=[μ_{ij}]의 원소로 식(1)의 조건을 만족한다. 특정한 데이터가 모든 클러스터링에 속하는 정도를 합하면 1이 되는 것은 하드 클러스터링의 경우와 동일하다.

$$J_m(U, V: X) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m |X_k - V_i|^2 \quad (1)$$

위 식(1)에서 $V=(V_1, V_2, \ldots, V_c)$는 c개의 클러스터링의 중심 벡터 집합이다. 즉, V_i는 i번째 클러스터링의 중심 벡터이다. $|X_k - V_i|$은 데이터 X_k와 i번째 클러스터링 중심과의 기하학적 거리를 나타낸다. V와 X_k의 차원이 p이면 $V_i = (V_{i1}, V_{i2}, \ldots, V_{ip})$이고 $X_k = (x_{k1}, x_{k2}, \ldots, x_{kp})$이다. 매개 변수 m은 데이터가 클러스터링이 속하는 정도를 조정하는 값으로 하드 클러스터링의 경우 1이다. μ_{ik}는 데이터 X_k가 클러스터링 $\hat{F}_i$에 대하여 속하는 정도를 나타내는 것으로 0 또는 1 가운데 하나의 값을 가지며, $(c \times n)$ 크기의 분할 행렬 $U=[\mu_{ij}]$의 원소로 식(2)의 조건을 만족한다. 특정한 데이터가 모든 클러스터링에 속하는 정도를 합하면 1이어야 하므로, 데이터는 하나의 클러스터링에만 속해야 한다.

$$\mu_{ik} \in [0,1], \qquad \sum_{i=1}^{c}\mu_{ik} = 1 \quad (2)$$

FCM 클러스터링 알고리즘을 정리하면 다음과 같다.
① 분할 수 c($2 \le c \le n$)와 m의 값을 선택한다.

② 퍼지 분할 행렬 $U^{(t)}$의 초기값을 결정한다. 통상 식(2)을 만족시키는 랜덤 값을 사용한다.

③ 식(3)을 이용하여 클러스터링의 중심 V를 계산한다.

$$V_i^{(t+1)} = \frac{\sum_{i=1}^{c} (\mu_{ik}^{(t)})^m X_k}{\sum_{i=1}^{c} \mu_{ik}^{(t)}} \qquad m > 1 \tag{3}$$

i=1,⋯,c

④ 식(4)를 사용해서 퍼지 분할 행렬을 갱신한다.

$$\mu_{ik} = \frac{1}{\sum_{j}^{c} \left(\frac{|X_k - V_i|^2}{|X_k - V_j|^2} \right)^{1/(m-1)}} \tag{4}$$

i=1,⋯,c k=1,⋯,n

⑤ $|U^{(t+1)} - U^{(t)}| < \delta$ 을 만족하면 이 과정을 종료하고 그렇지 않으면 단계 ③으로 복귀해서 이 과정을 반복한다.

Ⅲ. 들로네 삼각형 FCM(Delaunay Triangulation-FCM:DT-FCM)

주어진 데이터 분포에서 유의미한 클러스터를 뽑아내는 것은 데이터 마이닝, 분류 등에서 매우 어려운 문제 중의 하나이다[3]. 들로네 삼각형은 주어진 데이터에서 최대한 정삼각형이 되게 하면서, 이은 선들이 모여 만든 삼각형이다. 가장 중요한 것은 어떤 삼각형의 외접원도 그 삼각형의 세 꼭지점을 제외한 다른 어떤 점도 포함하지 않는다는 것이다. 즉, 길고 뾰족한 삼각형이 나오지 않는다. 일반적으로 이 알고리즘은 매개 변수나 사전 지식을 설정할 필요 없이 복잡한 형상의 클러스터와 비균질 밀도를 공간 데이터베이스에서 자동으로 발견 할 수 있으며, 사용자는 특수 응용 프로그램에 맞게 매개 변수를 수정할 수도 있다[5]. 들로네 삼각형은 평면위의 점들을 삼각형으로 연결하여 공간을 분할할 때, 이 삼각형들의 내각의 최소값이 최대가 되도록 하는 분할이다. 이렇게 구성된 전체 데이터에 들로네 삼각형은 각 삼각형의 외접원 내부에 다른 점들이 포함되지 않도록 공백의 원을 형성함으로써, 모든 삼각형이 이루는 각을 최대한 최소화하므로 전체 공간은 더욱 촘촘하게 분할된다. DT-FCM 제안은 들로네 삼각형에 기반한 새로운 공간 근접 정의를 사용하여 공간 클러스터를 생성한다. 그림 1은 클러스터가 두 개(C_1, C_2)로 결정되고 세 점을 기준으로 최대한 정삼각형을 만들어 가는 들로네 삼각형이다. 임의의 최근접 이웃 점을 검색 후, 삼각형을 수정하여 점을 삽입하거나 제거해나간다. 경계점의 d_2는 클러스터 C_2로 편입시키고 경계점의 애매한 위치에 있는 d_1과 d_3는 잡음으로 분류되어 클러스터링 계산 시 제외된다.

그러므로 들로네 분할은 최소각을 최대로 하는 방법이다. 구체적으로 그림 2에서와 같이 주어진 세 점(d_1,d_2,d_4)을 지나는 원에 대하여 다른 한 점 d_3와의 관계를 조사한다. 즉, 점 d_3가 외접원의 외부에 존재하게 되면 삼각형(d_1,d_2,d_4)의 최소각 α는 삼각형(d_2,d_3,d_1) 최소각 β보다 크게 된다[2,5]. 따라서, 사각형(d_1,d_2,d_3,d_4)에서 대각선은 d_2d_4로 정해진다. 만일 대각선이 최초에 d_1d_3로 정해져 있었다면 최소각 검사에

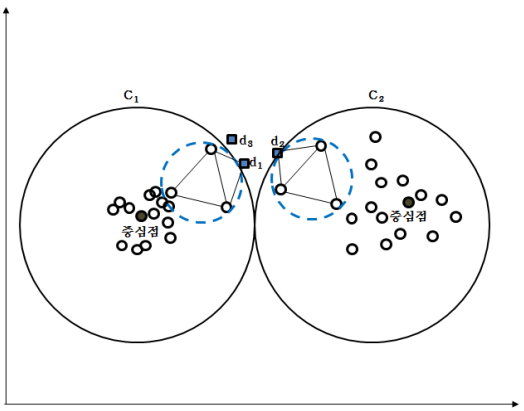

그림 1. 들로네 삼각형의 잡음 처리 방법
Fig. 1. Noise Processing Method of Delaun Triangulation

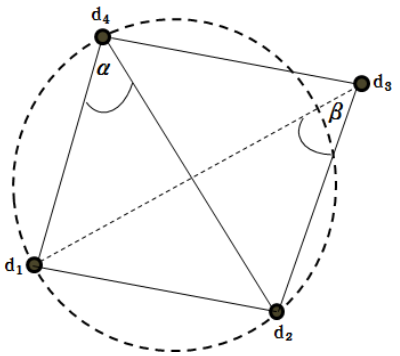

그림 2. 들로네 삼각형의 기본 개념
Fig. 2. Basic Concepts of Delaunay Triangulation

의해 d_2d_4로 결정되어 진다. 두 삼각형 사이의 최소각을 비교해 대각선을 결정하고 이웃 삼각형들에 대해 위의 과정을 계속적으로 수행하게 된다.

Ⅳ. 실험결과

클러스터링 초기값 결정에 따른 실행시간을 평가하기위해 FCM의 150개 데이터를 매트랩 을 사용하여 들로네 삼각형을 실험하였다. 데이터는 2차원의 삼각형을 이루는 세 점에 대한 인덱스로 구성되어 들로네 삼각분할로 플로팅 된다. 2차원 들로네 삼각분할은 각 삼각형에 대한 외접원이 내부에 다른 점을 포함하지 않도록 하여 자연적으로 더 높은 차원으로 확장된다. 작업 수행은 임의의 점을 기준으로 삼각형의 삼각분할을 검색하고 중심점을 계산한다. 중심점으로 삼각분할을 플로팅 한 후 최근접이웃 점을 검색한다. 삼각분할 내의 다른 점을 포함하는지 검색하여 삽입과 제거를 반복해가며 모든 점에 제약 조건을 적용한다. 그림 3은 150개 데이터를 기준으로 들로네 삼각형을 생성했다. 들로네 삼각형은 가장 중요한 개념이 세 점의 외접원내에 다른 점이 포함되어 있지 않아야 된다. 그림 3에서 보듯이 삼각분할내의 점들과 중복을 제거하고 추가해가며 레이블을 플로팅 한다.

그림 4는 들로네 삼각형을 적용하기 전의 FCM의 결과이다. 클러스터가 2개 결정되고 원점선의 잡음이나 중복이 존재해 중심이 치우쳐 국부 최적해를 제대로 찾지 못하는 경우가 발생하고 있다. 또한, 전체적인 모양이 어지러우며, 실행시간도 모든 점들을 계산하기 때문에 비교적 많이 걸린다.

반면에 제안된 DT-FCM을 사용하여 얻은 그림 5는 잡음이나 이상치가 깔끔히 제거되어

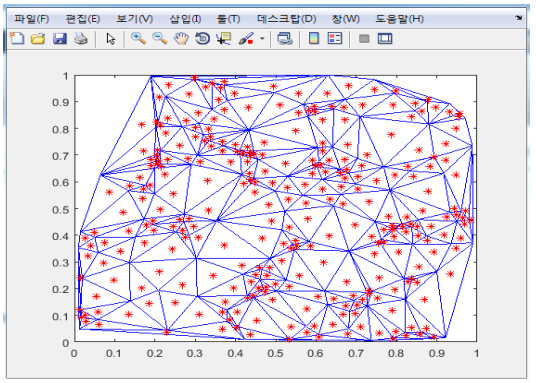

그림 3. 150개 데이터에 대한 2차원 들로네 삼각형
Fig. 3. Two-dimensional Delaunay Triangulation for 150 data

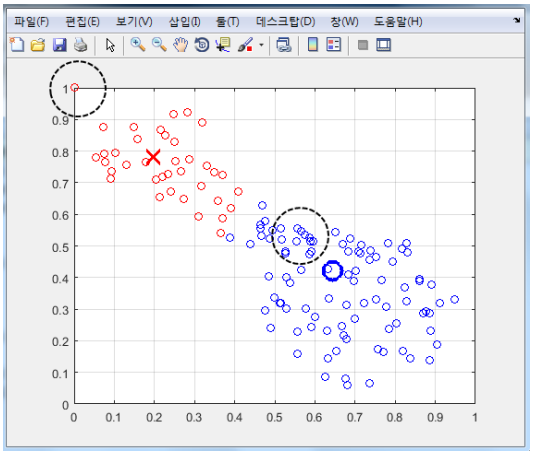

그림 4. FCM 클러스터(C=2)로 결정된 중심값
Fig. 4. The center value determined by the FCM cluster (C=2)

중심의 초기값을 명확하게 결정하였다. 육안으로도 중심의 전역 클러스터링은 FCM보다 정확한 클러스터 중심을 찾아내고 있으며, 영역에 애매하게 위치한 잡음을 제거하여 클러스터링 중심을 찾아냈다.

그림 6은 FCM과 DT-FCM의 CPU 실행시간 성능 비교이다. 초기 잡음 문제를 해결하지 못하는 FCM에 비해 제안된 DT-FCM은 들로네 삼각형의 내각의 최소값이 최대가 되도록 하는 분할이다. 이 방법은 FCM으로 클러스터링된 점들 중에서 가장 저밀도의 점들을 외접원의 들로네 삼각형을 만들어 잡음을 제거하므로서, 임의의 150개 데이터를 실험 시 18.7%실행시간을 줄일 수 있었다.

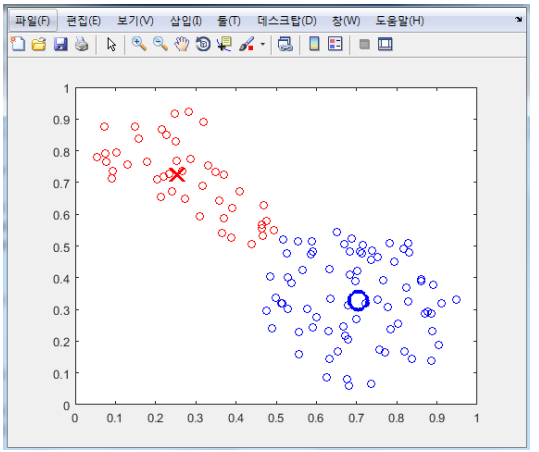

그림 5. DT-FCM 클러스터(C=2)로 결정된 중심값
Fig. 5. Center value determined by DT-FCM cluster (C = 2)

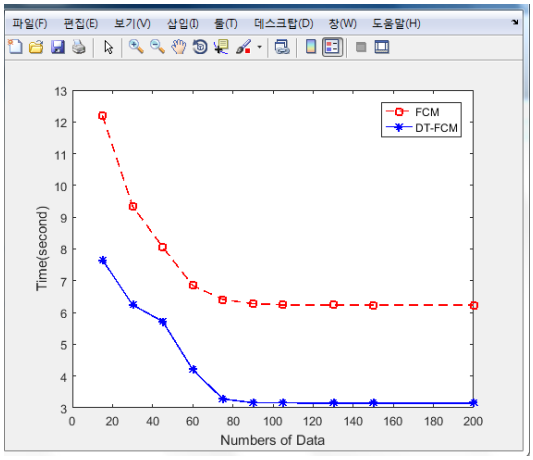

그림 6. FCM vs DT-FCM 실행시간 비교
Fig. 6. Comparison of FCM vs DT-FCM execution time

V. 결론 및 향후 연구과제

본 논문에서는 효율적인 DT-FCM 클러스터링 잡음 처리 방법을 제안하였다. 유클리디안 거리를 이용하여 클러스터링 중심값을 결정하는 FCM은 초기 클러스터 중심에 상당히 종속적이다. 특히, 잡음이 섞여 있는 데이터를 클러스터링 하는 경우는 실행시간에 많은 영향을 미친다. 제안 내용은 들로네 삼각형을 이용하여 DT-FCM 클러스터링 전처리 단계에서 거리 계산에 부담이 되는 잡음들을 제거한 후 FCM 클러스터링을 수행한다. 임의의 데이터 세 점을 선택하여 들로네 삼각형을 만들어 삼각형의 외접원을 구한 후, 경계점을 포함해서 외접원 내의 다른 데이터가 포함되는지 검사해서 포함되면 클러스터 데이터로 편입시키고, 그렇지 않은 경우에는 잡음으로 처리하여 계산에서 제외시킨다. 잡음이 제거됨으로서 전체적인 실행시간을 줄일 수 있었다. 제안된 DT-FCM은 실험 결과 기존의 FCM보다 18.7% CPU 실행시간이 감소하여 성능이 우수하다는 것을 보였다. 향후에는 중복되는 데이터를 선 제거하는 퀵정렬을 이용하여 보다 우수한 방법의 빅데이터 FCM을 연구하고자 한다.

References

[1] B. Geiger, "*Three-dimensional modeling of human organs and its application to diagnosis and surgical planning*," Report 2105, INRIA Sophia-Antipolis France, 1993.

[2] D. T. Lee and A. K. Lin, "Generalized delaunay triangulation for planar graphs," *Discrete Computational Geometry*, vol. 1, no. 3, pp. 201-217, 1986.

[3] G. Beliakov and M. King, "Density based fuzzy c-means clustering of non-convex patterns," School of Information Technology, Deakin University, 221 Burwood Hwy, Burwood 3125, Australia, 2005.

[4] G Beliakov and M King, "Density based fuzzy c-means clustering of non-convex patterns," *Eur. J. Operational Res.*, vol. 173, no. 3, pp. 717-728, Sept. 2006.

[5] Joseph O'rourke, "*Computational Geometry in C*," Cambridge University Press, pp. 175-177, 1994.

[6] J. Yang. Y Ma, X Zhang, and S Li, "A method for initializing the K-means clustering algorithm using delaunay triangulation," *DEStech Trans. Eng. and Technol. Res.*, Jun. 2017.

[7] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. ACM SIGMOD 3rd Int. Conf. Knowledge Discovery and Data Mining*, pp. 226-231, AAAI Press, 1996.

[8] N. Pal, K.Pal, and J. Bezdek, "A possibilistic fuzzy c-means algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 517-530, 2005.

[9] R. Krishnapuram, H. Frigui, and O. Nasroui, "Fuzzy and possibilistic shell clustering algorithm and their application to boundary detection and surface approximation," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 1, pp. 29-60, 1995.

[10] W. Yang, J. Hu, and S. Wang "A delaunay triangle-based fuzzy extractor for fingerprint authentication," *2012 IEEE 11th Int. Conf. Trust, Secur. and Privacy in Comput. and Commun.*, pp. 66-70, Liverpool, United Kingdom, Jun. 2012.

[11] S. Ghosh and S. K. Dubey, "Comparative analysis of K-Means and fuzzy C-Means algorithms," *IJACSA*, vol. 4, no. 4, pp. 35-39, 2013.

[12] X. Yang and W. Cui1, “A novel spatial clustering algorithm based on delaunay triangulation,” *ICEODPA*, Wuhan, China, Dec. 2008.

이 광 규 (Kwang-Kyu Lee)

1985년 2월 : 동국대학교 수학과 학사
1991년 2월 : 동국대학교 이학석사 응용수학
2002년 8월 : 충북대학교 전자계산과 이학박사
1996년~현재 : 신한대학교 IT융합공학부 교수
<관심분야> 데이터마이닝, 퍼지논리, 정보보안

우 정 현 (Jung-Hyun Woo)

1990년 2월 : 인하대학교 응용물리학과 학사
2008년 8월 : 인천대학교 공학석사 IT정책
2018년 2월 : 인천대학교 컴퓨터공학과 공학박사
2010년~2016년 : 앤텍코리아(주)
2013년~2015년 : 신한대학교 컴퓨터정보학과 겸임조교수
2017년~현재 : 유한대학교 출강
<관심분야> 데이터베이스, 데이터마이닝, 네트워크, 정보보안