Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 4, August 2013, pp. 354-359
http://dx.doi.org/10.5391/JKIIS.2013.23.4.354

Binary Harmony Search 알고리즘을 이용한 Unsupervised Nonlinear Classifier 구현

Implementation of Unsupervised Nonlinear Classifier with Bianry Harmony Search Algorithm

이태주* · 박승민* · 고광은* · 성원기** · 심귀보*†

Tae-Ju Lee, Seung-Min Park, Kwang-Eun Ko, Won-Ki Sung, and Kwee-Bo Sim†

*중앙대학교 전자전기공학부, **강원대학교 전자정보통신공학부

† School of Electrical and Electronics Engineering, Chung-Ang University

요 약

본 논문을 통해서 우리는 최적화 알고리즘인 binary harmony search (BHS) 알고리즘을 이용하여 unsupervised nonlinear classifier를 구현하는 방안을 제시하였다. 패턴인식을 위한 기계학습이나 뇌파 신호의 분석 과정과 같이 벡터로 표현되는 특징들을 분류하는데 있어 다양한 알고리즘들이 제시되었다. 교사 학습기반의 분류 방식으로는 support vector machine과 같은 기법이 사용되어왔고, 비교사 학습 방법을 통한 분류 기법으로는 fuzzy c-mean (FCM)과 같은 알고리즘들이 사용되어 왔다. 그러나 기존에 사용해 왔던 분류 방법들은 비선형 데이터 분류에 적용하기 힘들거나 교사 학습을 적용하기 위해서 사전정보를 필요로 하는 문제점이 있다. 본 논문에서는 경험적 접근을 통해 공간상에 분포된 벡터 사이의 기하학적 거리를 최소로 만드는 벡터 집합을 선택하고 이를 하나의 클래스로 간주하는 방법을 적용한 분류법을 제시하였다. 비교 대상으로 FCM과 artificial neural network (ANN) 기반의 self-organizing map (SOM)을 제시하였다. 시뮬레이션에는 KEEL machine learing dataset을 사용하였고 그 결과, 제안된 방식이 기존 알고리즘에 비해 더 나은 우수성을 지니고 있음을 확인하였다.

키워드 : 패턴인식, 뇌파, 하모니 서치, 비교사 비선형 분류, 클러스터링

Abstract

In this paper, we suggested the method for implementation of unsupervised nonlinear classification using Binary Harmony Search (BHS) algorithm, which is known as a optimization algorithm. Various algorithms have been suggested for classification of feature vectors from the process of machine learning for pattern recognition or EEG signal analysis processing. Supervised learning based support vector machine or fuzzy c-mean (FCM) based on unsupervised learning have been used for classification in the field. However, conventional methods were hard to apply nonlinear dataset classification or required prior information for supervised learning. We solved this problems with proposed classification method using heuristic approach which took the minimal Euclidean distance between vectors, then we assumed them as same class and the others were another class. For the comparison, we used FCM, self-organizing map (SOM) based on artificial neural network (ANN). KEEL machine learning datset was used for simulation. We concluded that proposed method was superior than other algorithms.

Key Words : Pattern Recognition, EEG, Harmony Search, Unsupervised Nonlinear Classification, Clustering

1. 서 론

패턴인식을 위한 기계학습 과정이나 뇌파의 신호를 분석하는 과정에서 알 수 있듯, 추출된 특징을 분류하는 알고리즘은 신호 처리 분야에서 전체 시스템의 정확성과 신뢰성을 좌우하는 중요한 요소이다. 일반적으로 패턴 인식이나 뇌파 신호 처리는 입력 신호의 특징을 추출하여 추출된 특징을 기반으로 해당 신호가 어느 클래스에 속하는지 분류하는 과정을 거친다[1]. 이때 특징이 어느 클래스로 분류되는가에 따라 내보낼 출력을 선택하게 되는데, 사용자가 원하는 결과를 얻을 수 있도록 시스템을 제어하기 위해서는 적절한 클래스를 선택하여 특징을 분류하는 방법이 중요하다. 시스템에 어떤 분류 방법을 적용할 것인가는 입력된 신호에서 추출된 특징이 선형적, 비선형적 특성을 고려해야 하고, 교사 학습 기반의 알고리즘을 사용하는지, 비교사 학습 기반

접수일자: 2013년 3월 10일
심사(수정)일자: 2013년 4월 20일
게재확정일자 : 2013년 5월 2일
† Corresponding author
본 논문은 한국연구재단 중견연구자지원사업(No.2012-0008726)에서 지원하여 연구하였습니다. 연구비 지원에 감사드립니다.

의 알고리즘을 사용할 것인지 결정하여야 한다[2].

현실 세계에서 얻을 수 있는 신호들에서 추출된 특징 벡터의 경우, 대부분의 신호가 비선형적인 특성을 갖고 있음을 알 수 있다. 선형 신호를 기반으로 한 선형 시스템의 경우에는 시스템을 행렬로 표현하여 간단하게 풀 수 있다[3]. 이것과 반대로 비선형 특성을 지닌 시스템은 데이터의 분석에 고려해야 할 요소들이 더 존재한다. 기계는 사람이 아니므로 직선이 아닌 곡선을 이용해 분류를 수행하는 것은 어려움이 있어, 이 문제점을 해결하기 위해 다양한 수학적 방식을 고안하였다. 그러나 수학적 방식의 사용은 시스템이 복잡해지고 어려운 계산을 수행해야 하는 단점이 있다. 특히 비선형 분류기법으로 널리 사용되는 커널 트릭의 경우, 본래의 차원이 아닌 더 높은 차원에서 계산을 수행하여야 하기 때문에, 차원 증가에 따른 데이터 및 계산량이 증가하여 복잡도가 커지는 어려움을 갖고 있다[4-5]. 뿐만 아니라, 어떠한 함수를 커널로 선택하는가에 따라 분류의 성능이 좌우되기도 해, 좋은 함수를 찾아야 하는 단점이 있다. 커널 트릭 외에도, 선형 분류를 작게 나누어 사용하는 piecewise linear 분류를 적용하거나, k nearest neighbor (k-NN)처럼 특징 사이의 거리를 이용한 방법 등이 제시되었다[6-7].

또 분류 방법은 교사 학습 기반의 분류법과 클러스터링으로도 불리는 비교사 학습 기반의 분류법으로 나뉘는데, 교사 학습 기반의 분류법은 사전에 주어진 정보를 바탕으로 입력 신호와 출력 신호 사이의 관계를 찾는 것이 특징이다. 학습을 위한 신호를 바탕으로 연결 강도를 조절하는 artificial neural network (ANN) 기반의 알고리즘이나 최외각 벡터를 지지벡터로 활용하는 support vector machine (SVM) 등이 대표적인 교사학습 방식의 분류 기법이다. 비교사 학습 방식의 알고리즘으로는 K-means algorithm (KMA)과 같은 기법이 대표적으로, 유클리드 거리 기반의 방식을 비롯한 다양한 방법이 사용된다. 시스템이 운용되는 환경에 따라 두 방식 중 적절한 방식을 선택하여야 한다. 교사 학습 방식을 사용하게 되면 사전 정보가 필요하게 되어 시스템을 구성하는데 이를 고려하여 설계를 해야 한다. 교사 학습 방식은 사전 정보를 활용할 수 있어, 아무런 정보도 없이 분류를 해야 하는 비교사 학습 방식보다 높은 분류율을 얻을 수 있는 장점이 있다. 그러나 특징을 학습하기 위해서 발생하는 추가적인 자원이 존재하고, 계산량이 늘어나게 되어 시스템의 부하가 증가한다. 또 입력 받은 데이터가 부정확하거나 누락된 부분이 있을 경우, 이에 대한 대처 능력이 미흡한 점 등이 문제점으로 지적된다[8]. 두 알고리즘의 장단점을 적절히 살리고 기존 알고리즘의 단점을 보완하기 위한 수많은 알고리즘들이 제시되어 왔다[9-10].

본 논문에서는 기존에 제시되었던 분류 및 클러스터링 알고리즘들이 지니고 있는 문제점을 피하고 더 나은 분류 정확도를 얻기 위해서 meta-heuristic한 접근 방식을 통한 분류 과정을 수행하였다. 차원이 증가하는 등의 복잡한 수학 계산이 발생할 수 있는 상황을 경험적인 분류 방법을 통해 피하고 알고리즘의 반복적인 시도를 이용하여 사전 정보 없이 분류가 가능하도록 하였다. 이때 사용한 알고리즘은 최적화 알고리즘인 harmony search 알고리즘을 사용했다.

본 논문의 2 장에서는 제안된 방법의 이론적 배경과 비교 대상이 되는 알고리즘들의 이론을 다루었고, 3 장에서는 제안된 알고리즘에 대한 설명을 제시하였다. 4장과 5장에서는 시뮬레이션을 통해 검증하는 과정을 가졌다.

2. 이론적 배경

2.1 Binary Harmony Search

Harmony search (HS) 알고리즘은 재즈의 연주자가 악기의 화음을 맞추어 조화로운 음을 이끌어내는 것에 착안한 알고리즘으로, 주어진 해 공간에서 임의로 선택된 해들을 조합하는 과정을 반복해 최적 해 조합을 찾는 방법이다[11]. HS 알고리즘은 이전 시도에서 선택했던 값 및 목적 함수의 값들을 저장하기 위한 harmony memory (HM)를 갖는다. 메모리의 값을 업데이트하는 방법에는 크게 세 가지가 존재한다. n개의 입력 $X=\{x_1, x_2, \dots, x_n\}$을 갖는 목적 함수를 $f(X)$라 할 때, x_k의 값을 기존에 메모리에서 가져오는 방식과 x_k가 가질 수 있는 범위 내에서 무작위로 선택하는 방식, 마지막으로 pitch adjustment를 수행하여 값을 정하는 방식을 통해 메모리에 존재하는 값을 업데이트하고 모든 반복이 끝난 후, HM 내의 목적함수가 가장 좋은 결과를 내는 X가 HS의 결과이자 목적함수의 최적해가 된다.

HS가 x_k의 값을 업데이트 할 때, HM에서 참조하여 선택할 확률을 harmony memory consideration rate (HMCR)이라고 한다. 이 값을 높게 설정하면 메모리를 자주 참조하게 되어 빠르게 수렴하게 된다. 조기 해 수렴 및 지역 해에 빠지는 것을 막기 위해 무작위로 값을 업데이트하는 방법 외에도 악기를 조율하여 음색을 맞추는 행동과 유사한 pitch adjustment (PA)가 존재한다. PA 과정을 통해 기존에 구했던 해 근처를 탐색하게 되는데, 이 과정의 수행 비율을 pitch adjustment rate (PAR)라 부른다. HS에 사용되는 파라미터는 HM의 크기, HMCR 그리고 PAR만 존재하여 간단하게 사용 할 수 있는 장점이 있다.

Binary harmony search (BHS)는 HM에 저장되는 x_k의 해 범위가 binary로, 0과 1 뿐인 HS를 말한다. BHS를 사용하는 것을 통해 시스템으로 입력받은 신호의 특정 채널이나 공간상의 특징 벡터 중에서 어떤 데이터를 같은 클래스로 선택했는지 직관적으로 표현 할 수 있다. 또한, 처리해야 할 값이 두 가지 뿐이므로 빠른 속도로 최적 값에 수렴을 할 수 있다. BHS 알고리즘의 기본적인 동작 원리는 일반적인 HS와 동일하다. 단, PA가 발생 할 경우, 0과 1 이외의 값은 존재하지 않으므로 원래 HM에 저장된 x_k의 값이 1이었다면 0으로, 0이었다면 1로 바뀌게 되는 차이점이 있다.

2.2 Fuzzy C-Mean

FCM은 1974년 Dunn이 제안한 방법으로, k-means 알고리즘을 fuzzy 기법을 이용해 최적화 하는 방안이다[12]. 이후 Bezdek이 1981년에 정리하여 개선하였다[13]. FCM의 목적함수는 least-squared error인 다음 수식을 최소화하는 것이다.

$$J_m(U,V)=\sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik})^m\|y_k-v_i\|_A^2 \qquad (1)$$

여기서 y_k는 주어진 k 번째 데이터를 의미하는 벡터이고, v_i는 i 번째 클러스터의 중심을 나타내는 벡터이다. U는 y의 집합인 Y의 퍼지 c분할이고, U의 멤버십 함수 u_{ik}는 y_k가 i 번째 클러스터에 소속되는 정도를 나타내고,

m은 1보다 큰 실수로 가중치이다. $\| \bullet \|_A$는 A-norm 을 의미한다. 중심 벡터와 u_{ik}의 갱신은 다음 수식을 이용하여 수행한다.

$$u_{ik} = \sum_{j=1}^{c} \left(\frac{\|y_k - v_i\|}{\|y_k - v_j\|} \right)^{-\frac{2}{m-1}} \quad (2)$$

$$v_i = \frac{\sum_{k=1}^{N} \left\{ (u_{ik})^m \bullet y_k \right\}}{\sum_{k=1}^{N} (u_{ik})^m} \quad (3)$$

최초 U를 임의로 초기화하여 중심 벡터를 계산한 후, 이를 이용해 U를 갱신한다. 이러한 과정을 k번째와 $k+1$ 번째 u_{ij}에 대해 다음 식을 만족할 때까지 반복한다.

$$\max_{i,j} \left\{ |u_{ij}^{(k+1)} - u_{ij}^{(k)}| \right\} < \varepsilon \quad (4)$$

이 방법을 이용하여 클러스터의 중심 및 거리가 최소가 되는 클러스터의 멤버들을 찾을 수 있고 분류가 가능하다.

2.3 Self-Organizing Map

ANN 기반으로 한 다른 분류 알고리즘과 달리 self-organizing map (SOM)은 비교사 학습 기반의 분류 기법이다. 코호넨이 제안하여 코호넨 네트워크로도 불리며, 입력층과 경쟁층의 두 계층으로 이루어진 뉴런들이 승자 독점의 원리에 의해 학습하고 결과를 출력한다[14]. 입력 벡터와 계층 사이의 연결 강도 벡터를 비교했을 때 가장 가까운 뉴런이 경쟁에서 승리하고 해당 뉴런과 그 주변의 뉴런만 연결강도를 수정 할 수 있다. 이때 뉴런의 연결강도를 조절하는 식은 다음과 같다.

$$m_c(t+1) = m_c(t) + \alpha(t)[x(t) - m_c(t)] \quad (5)$$

$$m_i(t+1) = m_i(t) \quad \text{for } i \neq c \quad (6)$$

여기서 m_c는 유클리드 기하에서 입력 벡터 x와 가장 가까운 연결 강도 벡터이고, α는 0과 1사이의 값을 갖는 학습 상수이다. 입력 벡터와 연결 강도 벡터 사이의 거리는 유클리드 거리를 구하는 함수 d를 포함한 다음 수식을 통해 구할 수 있다.

$$d(x, m_c) = \min_j \{ d(X, m_j) \} \quad (7)$$

SOM 알고리즘의 작동 순서는 연결 강도를 초기화한 상태에서 식 (7)을 이용해 입력 벡터에 대해 최소 거리가 되는 출력 뉴런을 선택하고 식 (5)와 (6)을 반복하여 학습을 하게 된다. 정해진 반복 횟수에 도달하게 되면 알고리즘은 종료되고 이 때 사용된 반복 횟수에 따라서도 다른 결과를 얻을 수도 있어 적절한 반복횟수를 설정하여 학습을 수행하여야 한다.

3. BHS 알고리즘 기반 분류 기법

특징 벡터들이 공간상에 N 개 존재하고 이들이 둘 이상의 클래스로 나누어진다고 할 때, 같은 클래스에 속하는 좌표상의 점은 기하학적 거리가 작을 것이라고 예측할 수 있다. 제안하는 BHS 기반의 분류 기법은 주어진 N 개의 벡터 중에서 임의로 n개의 벡터를 선택해 이들 사이의 거리가 최소가 된다면 하나의 클래스로 간주해 분류하는 방법이다. HS 알고리즘은 heuristic 기반의 알고리즘이어서 어느 벡터를 선택해야 할지, 확인해야 할 많은 경우의 수 중에서 최적의 해에 빠르게 접근해 나갈 수 있다. 단 항상 정확한 해를 보장해주지 않고 시도할 때 마다 값이 다르게 나올 수 있으나 이를 고려하더라도 최적의 해와 근사한 값을 얻을 수 있다는 점에 의해 본 알고리즘을 선택하게 되었다. 전체 거리의 합산이 최소가 되는 벡터를 탐색하기 위해서 BHS를 이용해 최적의 그룹을 찾는 것이 본 알고리즘의 목표이다. 거리를 사용하는 방식은 다른 클러스터링 기법에서도 볼 수 있으나 본 알고리즘에서 사용한 거리 방식은 Minkowski 거리의 지수가 1인 맨해튼 거리를 사용하였다. 전체 거리의 합산 D는 다음과 같다.

$$D(x) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|x_i - x_j\|_1 \quad (8)$$

여기서 x_k는 선택된 n개에 속하는 특징 벡터를 의미한다. 거리를 합산할 때 각 좌표 사이의 거리가 두 번씩 들어가게 되므로, 2로 나누어 준다. 거리 합산 D를 이용해 이것이 최소가 되는 벡터를 찾기 위한 BHS의 목적 함수 다음 수식을 최소로 하는 x를 찾는 것으로 표현이 가능하다.

$$\min_{x,n,t} \frac{1 + D(x)}{n^t} \quad (9)$$

수식에 사용된 t는 선택된 벡터의 개수에 따른 차이를 보정해주는 값이다. 분자항의 1은 n의 값이 0이 되었을 때 전체 목적함수의 결과가 0/0이 되지 않도록 하기위해 사용되었으며, 이 경우에 값은 ∞가 되어 최소값이 될 수 없으므로 빈 클래스가 되는 경우는 자동적으로 메모리에서 저장되지 않는다. 또, n의 값이 1일 때, D의 값은 0이 되어 전체 목적함수의 값은 1이 된다. 따라서 t의 값은 최적의 해 그룹을 찾았을 때, 목적 함수의 값을 1보다 작게 만들어 주는 값을 가져야 한다.

$$\log_n \{1 + D(X)\} < t \quad (10)$$

결론적으로 t의 값은 X에 따라 결정 된다고 볼 수 있다. 이를 pseudo code로 표현하면 다음과 같다.

```
HS_CLSFYR(Input signal)
{
  initialize HMCR, PAR, itr, HM
  for i = 1 : itr
  {
    if rand()<HMCR? new_member = HM(rand)
    else if rand()<PAR? new_member =~HM(rand)
    else new_member = rand([x])
    find D, n, t
```

```
    if f(mem_worst) > f(new_member)
    {
      update HM
    }
  }
  result=best(HM)
}
```

제안한 방법의 입력으로는 특징 벡터의 좌표가 주어진다. 최초 동작에는 각종 파라미터를 설정하고 HM을 랜덤한 값으로 초기화한다. 설정된 반복 횟수만큼 탐색과정을 반복하여 어떤 이진 조합이 가장 최적의 값을 갖는지 찾게 되는데, HS의 업데이트 방법을 이용하여 이를 수행하게 된다. rand()는 normal 분포를 갖고 0과 1 사이의 값을 생성하는 함수를 의미한다. 즉, 파라미터의 확률에 따라 값을 업데이트 하는데, HM(rand)는 메모리 내에 존재하는 무작위의 값을 하나 취한다는 뜻을 의미한다. 제안한 방법은 BHS를 사용하므로, PA가 발생하면 1을 0으로, 0을 1로 바꾸기 위해 논리기호인 NOT을 사용해 PA를 구현하였다. HMCR과 PAR 두 확률을 만족하지 못할 경우 x의 범위 [x]내에서 무작위 값을 선택하게 된다. 이렇게 선택된 값들을 이용해 D, t, n 및 목적함수의 값을 정한다. 새로운 조합의 목적함수가 메모리에 존재하는 목적함수 중 가장 나쁜 값보다 좋은 결과를 지니고 있을 때, 메모리 내 가장 나쁜 값을 새로운 조합으로 메모리를 업데이트 한다. 이 과정을 반복해 설정 된 횟수에 도달하면 메모리안의 가장 좋은 값을 출력으로 내보내고 알고리즘은 종료된다.

4. 시뮬레이션 방법

본 논문에서는 제안한 알고리즘의 분류 성능을 확인하기 위해서 KEEL-dataset repository에서 제공하는 기계 학습 분류 데이터 셋을 사용하였다[15]. 그 중에서 특히 분류기의 성능 분석에 많이 사용되는 standard dataset의 2-클래스 류 데이터와 unsupervised dataset의 데이터를 사용하였다. 시뮬레이션을 수행하기 위한 프로그램은 Mathworks사의 MATLAB 8.0을 사용하였으며 FCM 및 SOM 알고리즘은 프로그램에 기본적으로 내장되어 있는 함수를 사용하여 구현하였다. FCM의 설정 값은 기본 설정 값을 사용하였으며, 이때의 값은 U 행렬에 사용된 지수의 값이 2.0, 최대 반복 횟수가 100 회, 오차 범위 10^{-5}이었다. SOM은 MATALB에 내장된 ANN tool과 연동되어 동작하고, 2개의 클래스로 맵을 나누도록 설정하였다. 시도 횟수는 100회, 초기 이웃의 개수는 3개로 지정하였으며, 거리의 측정 방식은 한 layer의 뉴런 사이에 거리를 측정하는 link distance 함수를 사용하여 구하도록 했다. 마지막으로 HS 알고리즘은 최대 반복 횟수 5000 회, HM의 크기를 10개로 하여 시뮬레이션을 수행하였으며, HMCR은 95 %, PAR은 5%로 설정하였다.

시뮬레이션에 사용된 데이터는 3종류이며 각각의 명칭은 'banana', 'weather', 그리고 'haberman'이었다. 'banana' 데이터는 2차원 데이터로, 총 5200개의 데이터를 갖고 있으나, 다른 데이터와의 개수를 비슷하게 만들기 위하여 1번부터 순서대로 10번째 데이터를 사용하여 총 520개의 데이터를 분류하였다. 'weather' 데이터는 총 4개의 차원으로 되어 있으며, 값이 문자열로 되어 있어 적당한 숫자를 할당해 공간상에 표현하였다. 데이터셋에 포함된 데이터의 개수는 14개이다. 'haberman' 데이터는 306개의 데이터를 갖고 있고 3개의 차원을 갖고 있다. 사용된 dataset들은 누락된 값이 없었으며, 클래스는 2개였다. 각 데이터셋을 구성하고 있는 모든 차원을 사용하여 분류를 수행하였고, 시뮬레이션은 한 데이터 당 50번 반복하여 진행되었다.

5. 시뮬레이션 결과 및 고찰

시뮬레이션 결과는 그림 1과 같다. 'banana'의 경우, HS는 약 55%, FCM 및 SOM은 약 51%의 평균 분류율을 보였다. 세 알고리즘 모두 비교적 저조한 분류율을 보였는데 이는 2개의 차원만 존재하는 데이터인 'banana' dataset의 특성상, 낮은 차원에 집중된 데이터가 거리를 기반으로 하는 알고리즘의 정확도에 영향을 미쳐 낮은 분류율을 보인 것으로 생각 할 수 있다. 실제로 세 알고리즘 FCM, SOM

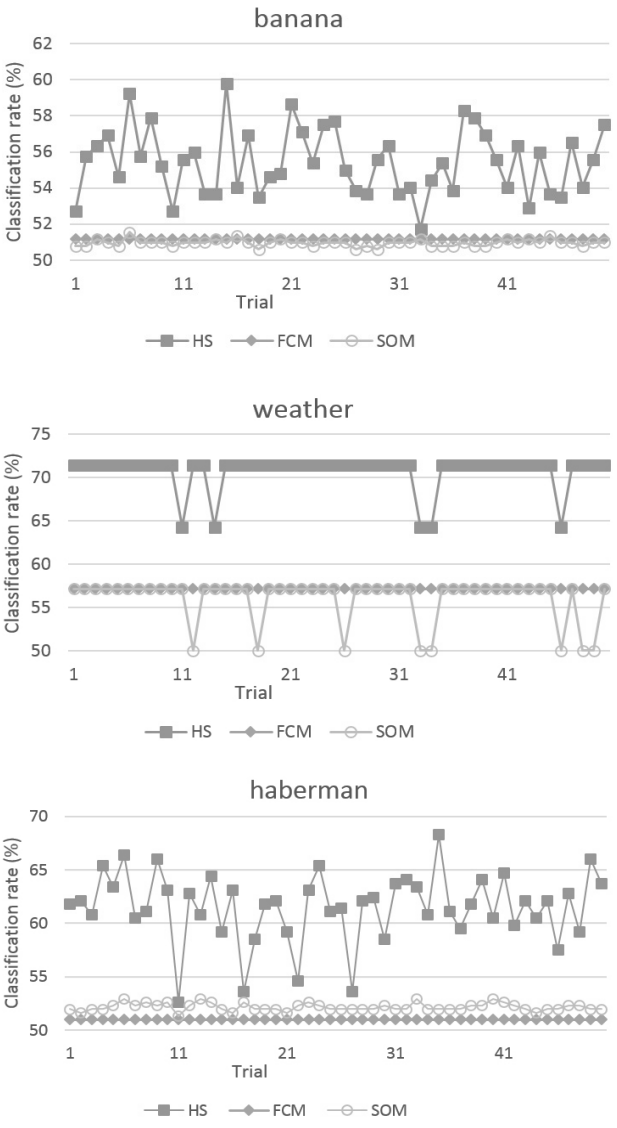

그림 1. 세 가지 데이터에 대한 분류 결과들
Fig 1. Classification results for three dataset

표 1. 각 dataset에 대한 평균 분류율
Table 1. Average classification rate for each dataset

	HS [%]	FCM [%]	SOM [%]
banana	55.4423	51.1538	50.9423
weather	71.4286	57.1429	56.0000
haberman	61.5163	50.9804	52.1569

그리고 제안한 방법 모두 기하학적 거리를 이용한 알고리즘이기 때문에 어느 정도 분류 가능한 여백이 없다면 분류율이 저조함을 보이고, 결과를 통해서도 이러한 내용을 확인할 수 있었다. 거리 기반의 분류기가 동작하기 힘든 상황에서도 제안한 방식은 다른 알고리즘보다 4% 정도 더 나은 성능을 평균적으로 보이는 것을 확인 할 수 있었다. 'weather' 데이터의 결과는 HS가 독보적으로 높은 결과를 보이고 있음을 확인할 수 있다. HS의 평균 분류율은 약 71%의 값을 보였고, FCM은 57%, SOM은 56%의 분류율을 보였다. 이 같은 결과를 통해서 제안한 분류법이 약 15% 더 높은 분류율을 보이며 좋은 결과를 얻을 수 있음을 확인할 수 있었다. 이와 같은 경향은 다음 dataset인 'haberman'에서도 확인 할 수 있었는데, HS가 약 61%, 나머지 두 알고리즘이 각각 50%, 52%로 약 10%가량 제안한 알고리즘이 더 높은 분류율을 보였다. 주어진 세 가지 데이터 셋의 결과에서 HS를 이용한 알고리즘이 다른 두 알고리즘보다 더 나은 결과를 보여주는 것을 확인 할 수 있었다.

반면 각 실험의 개별적인 결과를 통해서 기존에 많이 사용되었던 heuristic 기반의 클러스터링 기법인 FCM과 SOM에 비해 제안된 알고리즘은 분류율의 편차가 비교적 큰 것을 알 수 있는데, 다른 두 알고리즘에 비해 항상 같은 결과를 얻기 힘들고, 심할 때에는 다른 알고리즘 수준으로 성능이 떨어지는 경우도 드물게 있어 HS의 수렴성에 한계를 확인할 수 있었다. 그러나 반대의 경우 역시 존재하여 평균을 훨씬 상회하는 결과를 얻을 수도 있기에 수렴성 개선을 한다면 분산의 분포를 줄이고 전체적인 성능을 개선할 수 있으리라 기대한다.

6. 결론 및 향후 연구

본 논문을 통해 heuristic 기반의 새로운 분류 알고리즘을 제시하고 그 성능을 기존 알고리즘과 비교하여 우수성을 발견하였다. 기존에 비교사 클러스터링 기법으로 주로 사용되었던 FCM 알고리즘과 SOM 알고리즘에 비교했을 때 평균적으로 최소 5%에서 최대 15% 가량의 성능 향상이 존재하였고, 모든 시도에서 기존 알고리즘보다 우수한 성능을 보이는 것을 확인 할 수 있었다. 사용한 dataset의 분류 결과를 통해서 우리는 제안한 알고리즘이 기존 방법으로 분류하기 힘든 데이터들에 대해 분류가 가능하다는 점을 확인하였다. 따라서 제안된 알고리즘을 통한 비선형 비교사 분류를 수행할 수 있으리라는 결론을 내릴 수 있었다.

다만 제안된 방법의 경우 다른 기존 알고리즘과 다르게 분류율 결과의 분포가 고르지 못하고 최대 최소의 간극이 큰 양상을 보이고 있어 시스템의 정밀도에 영향을 미치리라 생각된다. 이러한 특성은 heuristic 알고리즘의 특성에서 비롯된 것으로, 항상 최적해를 보장해 주지 못하기 때문이다. 그럼에도 불구하고 다른 두 알고리즘보다 더 나은 결과를 보이고 있음은 제안된 방법이 적절한 분류 기법임을 의미한다. 향후 연구에서는 HS 알고리즘의 수렴성을 개선하여 정밀한 결과를 얻을 수 있도록 하여야 하겠고, 속도 및 메모리의 활용을 최적화 할 수 있도록 개선하고자 한다.

References

[1] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms", *IEEE Transactions on Neural Networks,* vol. 16, no. 3, 2005.

[2] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Elsevier, San Diego, CA, 2009.

[3] C. Chen, *Linear System Theory and Design,* Oxford University Press, New york, NY, 1999.

[4] G. Yuan, C. Ho, and C. Lin, "Recent Advances of Large-Scale Linear Classification," *Proceedings of the IEEE*, vol. 100, no. 9, pp.2584-2603, 2012.

[5] S. Ding, H. Zhu, W. Jia, C. Su, "A survey on feature extraction for pattern recognition," *Artificial Intelligence Review,* vol. 37, issue 3, pp. 169-180, 2012.

[6] X. Huang, S. Mehrkanoon, J. A. K. Suykens, "Support vector machines with piecewise linear feature mapping," *Neurocomputing,* vol. 117, pp. 118-127, 2013.

[7] T. Cover, P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.

[8] E. Carrizosa, D. R. Morales, "Supervised classification and mathematical optimization," *Computers and Operations Reserch*, vol. 40, issue 1, pp. 150-165, 2013.

[9] W. Linfang, W. Ning, W. Juanjuan, and G. yue, "A Method of Nonlinear Manifold Classification using ISOMAP," *Proceedings of First International Conference on Innovative Computing, Information and Control,* pp. 22-25, 2006.

[10] U. Maulik, S. Bandyopadhyay, "Genetic algorithm-based clustering technique," *Pattern Recognition,* vol. 33, issue 9, pp. 1455-1465, 2000.

[11] Z. W. Geem and K. B. Sim, "Parameter-setting-free harmony search algorithm," *Applied Mathematics and Computation,* Vol. 217, Issue 8, pp. 3881-3889, 2010.

[12] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32-57, 1974.

[13] J. C. Bezdek, R. Ehrlich, W. Full, "FCM: The Fuzzy c-Means Clustering Algorithm," *Computers and Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[14] T. Kohonen, "The Self-Organizing Map," *Proceedings of the IEEE,* vol 78, issue 9, pp. 1464-1480, 1990.

[15] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac,

S. Garcia, L. Sanchez, F. Herrera, "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithm and Experimental Analysis Framework," *Journal of Multiple-Valued Logic and Soft Computing,* vol. 17, no. 2-3, pp. 255-287, 2011.

저 자 소 개

이태주(Tae-Ju Lee)
2013년 : 중앙대학교 전자전기공학부 공학사
2013년 ~ 현재 : 중앙대학교 대학원 전자전기공학부 석박사통합과정.

관심분야 : Brain-Computer Interface, Intention Recognition, Neuro-Robotics, Soft Computing 등.
Phone : +82-2-820-5319
E-mail : bindaman@cau.ac.kr

박승민(Seung-Min Park)
2010년 : 중앙대학교 전자전기공학부 공학사
2010년 ~ 현재 : 중앙대학교 대학원 전자전기공학부 석박사통합과정

관심분야 : Brain-Computer Interface, Intention Recognition, Soft Computing 등.
Phone : +82-2-820-5319
E-mail : sminpark@cau.ac.kr

고광은(Kwang-Eun Ko)
2007년 : 중앙대학교 전자전기공학부 공학사
2007년 ~ 현재 : 중앙대학교 대학원 전자전기공학부 석박사통합과정

관심분야 : Multi-Agent Robotic System (MARS), Machine Learning, Context Awareness, Emotion Recognition System 등.
Phone : +82-2-820-5319
E-mail : kkeun@cau.ac.kr

성원기(Won-Ki Sung)
1986년 : 중앙대학교 대학원 전자공학과 공학박사
1986년~현재 : 강원대학교 전자정보통신공학부 교수

관심분야 : 감성인식, 지능시스템, 지능형 홈 네트워크, 지능형 감시 시스템 등.
Phone : +82-33-570-6354
E-mail : sungwk@kangwon.ac.kr

심귀보(Kwee-Bo Sim)
1990년 : The University of Tokyo 전자공학과 공학박사
1991년~현재 : 중앙대학교 전자전기공학부 교수
2006년~2007년 : 한국지능시스템학회 회장

관심분야 : 인공생명, 뇌-컴퓨터 인터페이스, 의도인식, 감성인식, 유비쿼터스 지능형로봇, 지능시스템, 컴퓨테이셔널 인텔리전스, 지능형 홈 및 홈 네트워크, 유비쿼터스 컴퓨팅 및 센서 네트워크, 소프트 컴퓨팅(신경망, 퍼지, 진화연산), 다개체 및 자율분산로봇시스템, 인공면역시스템, 지능형 감시시스템 등.
Phone : +82-2-820-5319
E-mail : kbsim@cau.ac.kr
Homepage URL : http://alife.cau.ac.kr