효율적인 지도 퍼지 군집화를 위한 휴리스틱 분할 진화알고리즘*

김성은° 류정우 김명원
숭실대학교 컴퓨터학부
babystep°@ssu.ac.kr , ryu0914@nate.com, mkim@comp.ssu.ac.kr

A Partitioned Evolutionary Algorithm Based on Heuristic Evolution for an Efficient Supervised Fuzzy Clustering

Sung Eun Kim° Joung Woo Ryu Myung Won Kim
Department of Computer Science, Soongsil University

요 약

최근 새로운 데이터마이닝 방법인 지도 군집화가 소개되고 있다. 지도 군집화의 목적은 동일한 클래스가 한 군집에 포함되도록 하는 것이다. 지도 군집화는 데이터에 대한 배경 지식을 획득하거나 분류 방법의 성능을 향상시키기 위한 방법으로 사용된다. 그러나 군집화 방법에서 파생된 지도 군집화 역시 군집화 개수 설정 방법에 따라 효율성이 좌우된다. 따라서 클래스 분포에 따라 최적의 지도 군집화 개수를 찾기 위해 진화알고리즘을 적용할 수 있으나, 진화알고리즘은 대용량 데이터를 처리할 경우 수행 시간이 증가되어 효율성이 감소되는 문제가 있다.

본 논문은 지도 군집화보다 강인한인 지도 퍼지 군집화를 효율적으로 생성하기 위해 진화성이 우수한 휴리스틱 분할 진화알고리즘을 제안한다. 휴리스틱 분할 진화알고리즘은 개체를 생성할 때 문제영역의 지식을 반영한 휴리스틱 연산으로 탐색 시간을 단축시키고, 개체 평가 단계에서 전체 데이터 대신 샘플링된 부분 데이터들을 이용하여 진화하는 분할 진화 방법으로 수행 시간을 단축시킴으로써 진화알고리즘의 효율성을 높인다. 또한 효율적으로 개체를 평가하기 위해 지도 퍼지 군집화 알고리즘인 지도 분할 군집화 알고리즘(SPC: supervised partitional clustering)을 제안한다. 제안한 방법은 이차원 실험 데이터에 대해서 정확성과 효율성을 분석하여 그 타당성을 확인한다.

1. 서 론

기계학습은 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 구분된다. 지도 학습은 분류 표시된 데이터(labeled data)를 이용하여 예측 모델을 생성하는 학습 방법으로 데이터마이닝에서 분류(classification)가 이에 해당된다. 반면 비지도 학습은 비분류 표시된 데이터(unlabeled data)를 분석할 때 사용되는 학습 방법으로 데이터마이닝에서 군집화(clustering)가 이에 해당한다.

최근 새로운 데이터마이닝 기법으로 세미-지도 군집화 방법(semi-supervised clustering)과 지도 군집화(supervised clustering) 방법이 소개되고 있다.

세미-지도 군집화 방법[1]은 적은 수의 분류 표시된 데이터를 이용하여 비분류 표시된 데이터를 군집화하는 방법이다. 이는 기존 군집화 방법과는 다르게 분류 표시된 데이터를 기반으로 군집화 함으로써 사용자 의도에 따라 결과를 생성할 수 있다.

지도 군집화[2][3]는 분류 표시된 데이터를 군집화 하는 방법이다. 이는 최소의 군집의 개수로 최대의 정확성을 가질 수 있도록 군집들이 형성되는 것을 목적으로 한다. 지도 군집화는 데이터를 압축하거나 데이터를 편집할 때 사용가능하고 또한 분류 방법의 성능을 향상시키기 위해 사용된다.

특히 지도 군집화에서 최적화된 지도 군집화를 찾는 것은 군집화에서 최적화된 군집화를 찾는 것과 마찬가지로 NP-complete 문제이다.

본 논문에서는 최적화된 지도 군집화를 찾기 위해 최적화 문제에 좋은 성능을 보이는 진화알고리즘을 이용한다. 그러나 진화알고리즘은 대용량 데이터를 처리할 때 수행 시간이 증가되어 효율성이 감소되는 문제가 있다. 이러한 문제를 보완하기 위해 지도 군집화를 위한 휴리스틱 분할 진화알고리즘을 제안하고 보다 강건한(robustness) 지도 군집화가 되기 위해 퍼지 이론을 이용한 지도 퍼지 군집화를 정의한다.

휴리스틱 분할 진화알고리즘은 개체를 생성할 때 문제영역의 지식을 반영한 휴리스틱 연산으로 탐색 속도를 단축시키고, 개체 평가할 때 전체 데이터 대신 샘플링된 분할 데이터들을 이용하여 진화하는 분할 진화로 수행 시간을 단축시킴으로써 진화알고리즘의 효율성을 높인다.

2. 지도 퍼지 군집화를 위한 휴리스틱 분할 진화알고리즘

진화알고리즘은 자연현상의 자연도태와 진화의 메카니즘에 기반을 둔 확률적인 탐색 알고리즘이다. 진화알고리즘은 개체집단을 유지하며 진화한다. 개체집단에는 문제에 대한 잠재적 해(candidate solution)를 표현하고 있는 개체(individual)들로 구성된다.

본 논문에서는 최적화된 지도 퍼지 군집화를 위해 개체의 유전인자를 군집 중심으로 나타내고 개체의 길이를 군집의 개수로 나타낸 가변길이 실수 표현(real-coded representation)으로 개체를 인코딩한다.

초기 개체집단에서 각 개체의 길이는 $[|CLASS|, 2\times|CLASS|]$ 범위에서 임의로 선택하고 유전인자인 군집 중심도 각 클래스에서 임의로 선택한다. $CLASS$는 클래스 집합을 의미한다.

적합도는 잠재적 해를 표현하고 있는 개체가 최적의 해에 얼마나 근접해 있는지를 나타내는 정도이며 적합도 함수에 의해 계산된다. 따라서 적합도 함수에 최적화 되어야할 문제의 모든 조건들이 포함되어야 한다. 적합도는 개체 평가 단계에서 계산되며, 선택 단계에서 우수한 개체를 선택하기 위해 사용된다.

개체 평가 단계에서 각 개체별 군집화 알고리즘을 수행하고

* 본 연구는 한국학술진흥재단 선도연구자지원사업에 의해 수행되었습니다. (과제번호: 2004-041-D00627)

그 결과를 적합도 함수로 평가하여 최적화된 지도 퍼지 군집화인지 평가한다.

개체 평가에 적용한 퍼지 군집화 알고리즘은 본 논문에서 제안한 지도 분할 군집화(SPC) 알고리즘을 사용한다.

SPC의 수행 과정은 <그림 1>과 같이 분할 군집화(partitioning clustering) 알고리즘인 K-means, FCM과 유사하다. 차이점은 그림에서 2번째 행과 같이 p번째 데이터, $\overrightarrow{x_p}=(x_{p1},x_{p2},\ldots,x_{pd},class_p)$, $\overrightarrow{x_p}\in N, N\subset R^d$의 클래스, $class_p$와 i번째 중심, $\overrightarrow{v_i}=(v_{i1},v_{i2},\ldots,v_{id},class_i)$, $\overrightarrow{v_i}\in C, C\subset R^d$의 클래스, $class_i$가 같을 경우에만 데이터와 중심 간의 거리를 계산한다는 점이다.

```
SPC (N = (x_1, x_2, ..., x_n))
  t = 0;
1. C^t = (v_1, v_2, ..., v_c);                    // 초기 군집 중심 설정
  do
    t = t + 1;
    for 각 데이터, x_p ∈ N do
      for 각 군집 중심, v_i ∈ C^{t-1} do
2.      if (class_p = class_i)
            D_pi = |x_p - v_i|;
3.    v_i* = min_{i∈C^{t-1}}(D_pi);                // 데이터와 가장 가까운 중심 선택
4.    x_p ∈ C_i^{t-1};                             // C_i 는 군집 중심, v_i 으로 표현된 군집.
5.  C^t = (v_1, v_2, ..., v_c);                    // 군집 중심 수정
6. while (C^{t-1} != C^t);
7. 퍼지 분할 행렬 생성;
```

그림 1 SPC 알고리즘

클래스에는 반드시 하나 이상의 군집 중심이 존재해야만 한다. 그러므로 초기 군집 중심을 설정할 때 최소 군집의 중심 개수는 클래스 개수와 같다. 만약 클래스 개수보다 클 경우 각 클래스마다 군집 중심을 임의로 선택한 후 전체 데이터에서 나머지 군집 중심을 클래스에 상관없이 임의로 선택한다. 그림에서 3번째, 4번째 행은 K-means처럼 데이터가 가장 가까운 군집에만 포함된다. 군집 중심이 더 이상 변하지 않을 경우 FCM와 같은 방법으로 중심과 데이터 간의 소속정도를 계산하여 퍼지 분할 행렬을 생성한다.

데이터를 군집화 시키고 군집화된 결과가 최적의 지도 퍼지 군집화인지 평가하기 위해 <식 1>와 같이 두 개의 항목의 가중치 합으로 적합도 함수를 정의한다.

$$Fit(C)=wHo(C)+(1-w)Pe(C) \qquad \text{식 1}$$

가중치 w, ($0\le w\le 1$)는 0에 가까울수록 군집 개수에 따른 패널티(penalty)항목인 $Pe(C)$에 가중치를 상대적으로 크게 함으로 적은 개수의 군집으로 군집화가 이루어진다. 반대로 1에 가까우면 동질성(homogeneousness)항목인 $Ho(C)$에 상대적으로 가중치가 커져 군집이 많아진다.

동질성은 지도 퍼지 군집화의 클래스 순도를 측정하는 항목으로 <식 2>와 같이 생성된 군집들의 평균 클래스 순도로 정의한다. 이때 i번째 군집에 대한 클래스 순도는 <식 3>과 같다.

$$Ho(C)=\frac{\sum_{i=1}^{|C|}Pu(c_i)}{|C|},\ c_i\in C \qquad \text{식 2}$$

$$Pu(c_i)=\frac{\sum_{k\in N_i^*}\mu_{ik}^{k_i^*}}{\sum_{k\in N}\mu_{ik}},\ k_i^*=\arg\max_{k\in CLASS}\left(\sum_{k\in N}\mu_{ik}^{k}\right) \qquad \text{식 3}$$

c_i는 i번째 군집을 의미하고 N_k는 k번째 클래스에 포함된 데이터 집합을 의미한다. μ_{ip}^{k}는 k번째 클래스에 포함된 p번째 데이터가 i번째 군집에 포함될 소속정도를 의미한다. k_i^*는 i번째 군집에서 대표 클래스를 말한다.

동질성만 고려하면 많은 군집으로 군집화가 이루어진다. 최악의 경우에는 데이터 개수만큼 군집의 개수가 생성될 수 있다. 이는 과잉학습 결과와 같으며 잡음에 의해 쉽게 왜곡될 수 있고 데이터 분석에 있어 아무런 의미가 없다. 따라서 생성되는 군집의 개수에 따라 지도 군집화의 일반화를 제어할 수 있는 패널티 항목을 <식 4>와 같이 정의한다.

$$Pe(C)=\begin{cases}\frac{|N|-|C|}{|N|}, & |C|\ge|CLASS|\\ 0, & |C|<|CLASS|\end{cases} \qquad \text{식 4}$$

패널티 항목은 군집의 개수가 증가할수록 0에 가깝게 되어 적합도에 크게 반영되지 못한다. 만약 군집의 개수가 클래스의 개수보다 작을 경우 적합도를 0으로 정의한다. 제안한 방법에서는 적합도가 클수록 최적의 해에 가까운 개체로 평가한다.

선택 단계에서는 룰렛휠(roulette wheel)방법과 엘리트 방법(elitist method)을 같이 사용한다.

재생산 단계에서 제안한 방법은 유전인자의 순서나 위치에 따라 개체가 구분되지 않고 개체의 길이에 따라 구분됨으로 클래스 정보를 이용하여 개체의 길이를 다양화 할 수 있는 클래스별 한 점 교배연산을 정의한다. 돌연변이 연산으로는 가우시안 돌연변이 연산[4]을 적용한다. 이 때 가우시안 함수에서 표준편차 σ는 <식 5>와 같이 세대가 증가함에 따라 값이 작아지도록 정의함으로써 세대에 따라 개체집단의 다양성을 제어한다.

$$\sigma_i^{g+1}=\frac{G-g}{G}\cdot\sigma_i^{g} \qquad \text{식 5}$$

G, g는 각각 전체 세대 수와 현재 세대 수를 나타내고, σ_i는 i번째 군집의 표준편차를 의미한다.

2.1 휴리스틱 연산

재생산 단계에서 교배와 돌연변이 연산은 특별한 정보 없이 확률을 기반으로 발생됨으로 개체집단의 다양성을 유지하지만 진화 수렴 속도가 저하되는 단점이 있다. 본 논문에서는 이러한 단점을 보완하기 위해 휴리스틱 연산을 정의한다. 휴리스틱 연산은 자식 개체가 생성될 때 문제영역의 지식을 이용하여 보다 우수한 해를 보다 초기에 생성시킬 수 있는 진화연산이다.

지도 퍼지 군집화의 최적화를 위해 휴리스틱 연산으로 분할 연산과 합병 연산을 정의한다.

분할 연산은 <식 3>에서 계산된 클래스 순도 $Pu(c_i)$가 임계값 ($0\le\theta_s\le 1$)보다 작은 모든 군집에 적용되어 두 개의 군집으로 분할시키는 연산이다. 이 때 분할된 두 개의 군집 중심은 대표 클래스와 차기 클래스에 각각 포함된 데이터에서 임의로 선택한다.

합병 연산은 가장 가깝고, 대표 클래스가 같은 두 군집을 하나의 군집으로 합병시키는 연산이다. 합병 연산은 분할 연산이 적용된 유전인자를 적용 대상에서 제외시킨다. 합병된 군집의 중심은 두 중심 사이의 중앙값으로 설정한다.

2.2 분할 진화

진화알고리즘을 이용하여 대용량 데이터로 모델을 최적화 시킬 경우 가장 수행 시간이 오래 걸리는 단계가 개체 평가 단계이다. 그 이유는 평가할 때마다 전체 데이터로 모델을 생성시킨 후 평가하기 때문이다.

분할 진화는 개체 평가할 때 전체 데이터를 이용하는 대신 사전에 분할한 샘플 데이터들을 이용함으로써 평가 시간을 단축시키는 진화이다. 분할 진화할 때 개체마다 분할된 샘플 데이터를 임의로 선택하여 평가한다. 이 때 세대마다 가장 우수한 개체는 평가에서 제외된다.

샘플링 데이터는 사전에 전체 데이터로부터 클래스 별로 계통 추출(systematic sampling) 방법[5]으로 생성한다. 이 때 비복원으로 데이터를 선택한다.

분할 진화는 진화 수행 속도가 전체 데이터를 분할한 샘플링 개수만큼 빨라질 수 있다. 그러나 분할 개수를 크게 하면 생성되는 모델의 정확성이 떨어진다.

3.실험

본 논문에서는 휴리스틱 분할 진화알고리즘을 이용한 효율적인 지도 퍼지 군집화 방법의 가능성을 확인하기 위해 인위적인 이차원 실험 데이터를 사용하여 정확성 측면과 효율성 측면에 대해 실험하였다.

본 실험에서 제안한 방법의 변수는 <표 1>과 같이 설정하였다. 휴리스틱 분할 진화알고리즘의 종료 조건을 최대 세대수를 초가 하거나, 일정 세대(*MAXCOUNT*)동안 엘리트 개체들의 적합도 차이의 합이 ε보다 작을 경우로 정의하였다.

표 1 변수 설정 값

변수명	설정값	변수명	설정값
개체크기	10	θ_s	0.6
교배 확률 (P_c)	0.6	최대 세대수(G)	500
돌연변이 확률 (P_m)	0.1	ε	0.01
w	0.2	*MAXCOUNT*	10

3.1 정확성 실험

<그림 2>은 인위적인 실험 데이터에서 휴리스틱 분할 진화알고리즘을 이용하여 찾아낸 지도 군집화 중심의 개수와 위치를 보이고 있다. 실험 데이터는 두 개의 클래스로 구성되어 있다. 실험 결과 클래스 분포를 반영한 적합한 중심의 개수와 위치를 선정함으로써 정확성 측면에서 그 가능성을 확인할 수 있다.

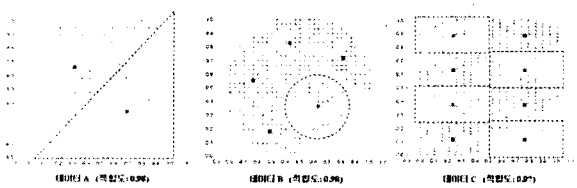

그림 2 생성된 지도 군집화 중심

3.2 효율성 실험

본 실험은 휴리스틱 분할 진화알고리즘의 효율성을 확인하기 위한 실험이다. <그림 3>은 제안한 휴리스틱 연산이 진화 수렴 속도를 향상시킬 수 있는지를 확인한 실험 결과로 데이터 C에 대한 진화 과정이다. 데이터 A와 B의 경우 탐색 공간이 크지 않기 때문에 확인할 수 없었다. 탐색 공간은 차원 수와 군집 수가 될 수 있으며 데이터 A와 데이터B는 최적의 군집의 개수와 초기 개체 집단의 군집의 개수가 거의 차이가 없기 때문에 탐색 공간이 크기 않다고 말할 수 있다.

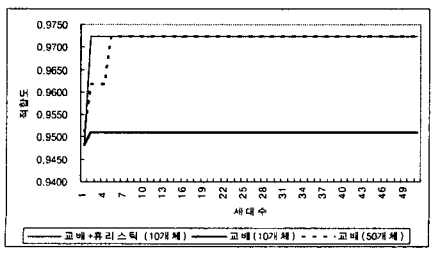

그림 3 휴리스틱 연산에 의한 진화 과정

<표 2>는 분할 진화 방법의 가능성을 확인하기 위해 실험 데이터를 분할한 개수와 분할에 포함된 평균 클래스별 데이터 개수를 나타내고 있다.

표 2 분할 개수에 따른 클래스별 평균 개수

분할 개수		1	2	4	8
데이터A	클래스1	200.0	100.0	50.0	25.0
	클래스2	200.0	100.0	50.0	25.0
데이터B	클래스1	532.0	266.0	133.0	67.0
	클래스2	109.0	54.5	27.3	13.6
데이터C	클래스1	400.0	200.0	100.0	50.0
	클래스2	400.0	200.0	100.0	50.0

이와 같은 데이터들을 사용하여 분할 진화한 결과는 <그림 4>와 같다. <그림 4>에서 (a)는 분할 데이터 집합의 개수에 따른 수행 시간을 보여주고 있으며 (·)의 숫자는 진화한 세대수를 나타낸다. 반면 (b)은 평균 오차를 나타내고 있다. 오차는 분할 진화하여 생성된 군집의 중심과 전체 데이터를 이용하여 생성된 군집의 중심 간의 차이를 말한다. 따라서 평균 오차가 작을수록 분할 진화에 의해 생성된 결과가 전체 데이터의 특성을 잘 반영하고 있다고 볼 수 있다. 데이터 A의 경우 8개로 분할하였을 때 4개의 군집 개수가 생성되었다.

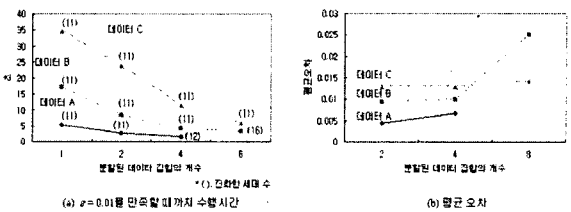

그림 4 분할 진화 방법에 의한 수행 시간과 평균 오차

4. 향후연구

본 논문에서는 제안한 휴리스틱 분할 진화알고리즘의 타당성을 이차원 실험 데이터로만 확인함으로써 그 가능성을 보였다. 향후 연구로는 다차원 대용량 데이터에 제안한 알고리즘을 적용하여 그 타당성을 확인하고 또한 의사결정트리와 같은 데이터마이닝 모델을 최적화하는데 응용하여 그 타당성을 확인할 계획이다.

5. 참고문헌

[1] Sugato Basu, "Semi-supervised Clustering with Limited Background Knowledge", Proc. of the Ninth AAAI/SIGART Doctoral Consortium, pp.979-980, 2004.

[2] Christoph F. Eick, Nidal Zeidat, and Zhenghong Zhao, "Supervised Clustering - Algorithms and Benefits", Proceedings of the 16th IEEE Int. Conf. on Tools with Artificial Intelligence, pp.774-776, 2004.

[3] Christoph F. Eick, Nidal Zeidat, Ricardo Vilalta, "Using Representative -Based Clustering for Nearest Neighbor Dataset Editing", Fourth IEEE Int. Conf. on Data Mining, pp.375-378, 2004.

[4] Johannes A. Roubus, Magne Setnes, Janos Abonyi,"Learning fuzzy classification rules from labeled data", Information Sciences Informatics and Computer Science: An International Journal archive Volume 150, Issue 1-2, pp.77-93, 2003.

[5] Richard L. Scheaffer, William mendenhall III, R. Lyman Ott, " Elementary Survey Sampling", 5th edition, Duxbury Press, 1996.