JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 1, pp. 69-76, January 2026
https://doi.org/10.9708/jksci.2026.31.01.069

Information Theoretic Local Refinement for Genetic Algorithm based Unsupervised Feature Selection

Hyunki Lim*

*Assistant Professor, Div. of AI Computer Science and Engineering, Kyonggi University, Suwon, Korea

[Abstract]

Unsupervised feature selection (UFS) aims to identify a compact subset of features that preserves the intrinsic structure of high-dimensional data without relying on label information. However, the search space of feature subsets is combinatorially large and the evaluation criteria are often non-differentiable, making heuristic and evolutionary search approaches particularly suitable. In this paper, we propose a novel wrapper-based UFS method that integrates a genetic algorithm (GA) with an information-theoretic refinement mechanism. The proposed DEL and ADD operators adaptively remove or add features based on entropy and mutual information criteria, enabling each chromosome to evolve toward a more informative and compact subset. This hybrid strategy effectively combines GA's global exploration capability with principled local adjustments. Experimental results on multiple benchmark datasets demonstrate that the proposed method outperforms existing GA-based UFS methods in terms of structure preservation, subset compactness, and overall clustering performance.

▸ **Key words**: Genetic Algorithm, Unsupervised Feature Selection, Information Theory, Entropy, Mutual Information, Evolutionary Optimization, Local Refinement

[요 약]

비지도 특징 선택(Unsupervised Feature Selection, UFS)은 레이블 정보 없이도 고차원 데이터의 내재적 구조를 보존하는 압축된 특징 부분집합을 찾는 것을 목표로 한다. 그러나 특징 부분집합의 탐색 공간은 조합적으로 매우 크며, 평가 기준 또한 비미분적 특성을 가지기 때문에 진화적 탐색 기법이 적합하다. 본 연구에서는 유전 알고리즘(GA)에 정보이론 기반의 로컬 리파인먼트 메커니즘을 결합한 새로운 래퍼 기반 UFS 방법을 제안한다. 제안된 DEL 및 ADD 연산자는 엔트로피와 상호정보량을 기반으로 특징을 적응적으로 제거하거나 추가하여, 각 염색체가 더 정보량이 높고 중복성이 감소된 특징 집합으로 진화하도록 돕는다. 이러한 하이브리드 전략은 GA의 전역 탐색 능력과 정보이론적 판단에 기반한 지역 탐색을 결합하여 효율성을 향상시킨다. 여러 벤치마크 데이터셋을 대상으로 한 실험 결과, 제안된 방법은 기존의 GA 기반 UFS 알고리즘보다 구조 보존 성능, 부분집합의 압축도, 그리고 군집 성능 측면에서 우수한 결과를 나타냈다.

▸ **주제어**: 유전 알고리즘, 비지도 특징 선별, 정보 이론, 엔트로피, 상호정보량, 진화 최적화, 지역 정제

• First Author: Hyunki Lim, Corresponding Author: Hyunki Lim
*Hyunki Lim (hlim20@kyonggi.ac.kr), Div. of AI Computer Science and Engineering, Kyonggi University
• Received: 2025. 11. 26, Revised: 2025. 12. 10, Accepted: 2025. 12. 22.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

고차원 데이터(high-dimensional data)가 다양한 분야에서 증가함에 따라, 효율적이고 신뢰성 있는 특징 선택(feature selection)의 중요성이 지속적으로 강조되고 있다. 특히 클래스 레이블이 존재하지 않는 환경에서는 지도학습 기반의 선택(unsupervised feature selection, UFS) 기준을 활용할 수 없기 때문에, 비지도 학습에서의 특징 선택 문제는 더욱 복잡한 형태의 탐색과 판단이 요구된다. UFS의 핵심 목표는 레이블 없이도 데이터의 내재적 구조(intrinsic structure)를 보존하면서, 불필요하거나 특징을 제거하여 학습 효율성을 높이고 모델의 일반화 성능을 개선하는 데 있다 [1].

일반적으로 UFS 접근법은 필터(filter), 래퍼(wrapper), 임베디드(embedded) 방식으로 구분된다. 필터 방식은 정보이론 지표, 통계적 척도, 혹은 그래프 기반 특징을 활용하여 빠르게 특징을 평가할 수 있으나, 모델 기반의 상호작용 효과를 반영하기 어렵다는 한계가 있다. 임베디드 방식은 모델 학습 과정에서 특징 중요도를 동시에 산출할 수 있으나, 특정 모델에 종속적이라는 제약이 존재한다. 반면, 래퍼 방식은 후보 특징 집합을 직접 평가함으로써 높은 예측 성능 혹은 데이터 구조 유지도를 달성할 수 있다는 장점이 있으나, 상대적으로 계산 계산 복잡도가 높고 비지도 학습에서의 직접적인 평가가 어렵다는 단점이 있다.

이러한 문제를 해결하기 위해 래퍼 방식에서는 다양한 탐색 기반 접근이 제안되어 왔으며, 그중 유전 알고리즘(genetic algorithm, GA) 은 전역 탐색(global search)에 강점을 지녀 대규모 특징 공간에서 유용하게 활용되어 왔다 [2]. 그러나 기본 GA는 탐색의 수렴 속도가 느리거나 지역해(local optimum)에 빠지는 경우가 발생하기 쉬운데, 이는 UFS 문제의 구조적 복잡성과 결합될 때 탐색 효율을 크게 저하시킨다. 따라서 GA의 전역 탐색 성질은 유지하면서도 보다 정교한 지역 탐색(local refinement)을 결합하는 하이브리드 접근이 요구된다.

본 연구에서는 이러한 배경에서, GA 기반 래퍼 방식을 위한 새로운 해(solution) 정제 기법을 포함한 하이브리드 진화적 특징 선택 알고리즘을 제안한다. 제안하는 방법은 GA의 전역 탐색 과정에 더해, 정보이론 기반의 선택 기준을 활용한 연산자를 통해 해를 정교하게 개선한다. 이를 통해 안정적인 수렴을 유도하며, 레이블 없는 환경에서도 의미 있는 특징 조합을 발견할 수 있는 탐색 능력을 확보한다.

본 연구의 주요 기여도는 다음과 같다.

(1) 정보이론적 DEL/ADD 목적함수를 통합한 GA 기반 UFS 구조 제안: 기존 GA-wrapper 방식이 단일 평가함수에 의존하는 것과 달리, 본 연구는 특징 제거(DEL)와 추가(ADD)를 위한 두 개의 독립적 정보이론 목적함수를 설계하여 탐색 과정에서의 명확한 방향성을 주었다. 이를 통해 조기수렴을 방지하고 보다 정교한 전역 탐색이 가능하도록 GA 구조를 최적화하였다.

(2) k-cardinality 제약과 정보량 기반 정제 절차의 결합: 본 연구는 k-cardinality 경계를 정보이론적 평가지표와 직접 결합하여 고차원의 정보량 계산을 효율적으로 할 수 있는 방식을 제시하여, 불필요한 조합 공간 탐색을 억제하면서도 구조적 일관성이 높은 후보 해를 생성하도록 하였다. 이는 기존 정보이론 기반 GA 변형에서는 제시되지 않은 새로운 정제 전략이다.

(3) 다양한 벤치마크 데이터셋을 통한 성능 검증: 실험을 통해 제안 방법이 기존 GA 기반 방법 대비 우수한 구조 보존 성능 및 탐색 효율을 보여줌을 입증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 정리하고, 3장에서는 제안하는 GA 기반 래퍼와 해 정제 방식을 상세히 기술한다. 4장에서는 다양한 실험을 통해 제안 방법의 우수성을 검증하고, 5장에서 결론 및 향후 연구 방향을 제시한다.

II. Related Works

UFS는 지도학습 환경과 달리 레이블 정보를 사용할 수 없기 때문에, 데이터의 내재적 구조와 특성 간 상관관계를 기반으로 특징의 유용성을 평가해야 한다. 기존 UFS 접근법은 크게 필터, 래퍼, 임베디드 방식으로 구분된다. 이 중 필터 방식은 가장 전통적이면서도 널리 활용되는 접근 방식으로, 주로 데이터의 통계적 특성이나 구조적 정보에 기반해 특징의 중요도를 평가한다 [3]. 이러한 방법은 일반적으로 단변량(univariate) 방법과 다변량(multivariate) 방법으로 구분된다. 단변량 필터 방법은 각 특징을 개별적으로 평가하여 순위를 매기고, 이 순위를 기반으로 최종 특징 집합을 선택한다. 이러한 접근은 불필요하거나 무관한 특징을 효과적으로 제거할 수 있는 장점이 있으나, 특징 간의 상호작용이나 종속성(dependency)을 반영하지 못한다는 뚜렷한 한계를 가진다. 반면 다변량 필터 방법은 여러 특징을 동시에 고려하여 그 조합이 제공하는 정보량이나 분리 능력을 기반으로 특징의 중요도를 평가한다. 이러한 방식은 단일 특징만을 독립적으로 고려하는 단변량 방

식에 비해 중복 특징과 관련 없는 특징을 더 효과적으로 구분할 수 있다. 이에 따라 일반적으로 다변량 필터 방법이 단변량보다 학습 알고리즘의 정확도가 더 높다 [4].

임베디드 방식은 특징 선택 과정을 모델 학습 단계에 직접 통합하는 접근으로, 필터 방식과 달리 모델 구조나 목적함수에 특징 선택 항을 포함한다는 점이 특징이다. 이러한 방법들은 일반적으로 군집화 또는 희소학습(sparse learning) 모델과 결합되며, 모델의 최적화 과정에서 동시에 특징 선택이 이루어진다. Wang 등은 k-means 기반의 클러스터링 절차와 희소 회귀모델을 통합한 비지도 특징 선별 방식을 제안하였다 [5] . 이 모델은 클러스터 지표 행렬과 잠재 특징 가중치 행렬을 공동으로 최적화하며, 비선형 희소 회귀 문제를 해결한다. Guo 등은 동일한 틀에서 Frobenius norm 기반의 목적함수를 사용한 변형 모델을 제안하였다 [6]. 임베디드 방식은 모델 학습과 특징 선택을 일체화함으로써 특징 간 상호작용을 반영하고, 필터 방식보다 더 구조적인 특징 평가가 가능하다는 장점이 있다. 그러나 일반적으로 목적함수의 복잡성이 증가하고 최적화 과정이 어려워지는 경우가 많아, 계산 비용 증가나 모델 수렴 문제와 같은 한계도 존재한다.

래퍼 방식은 후보 특성 집합의 품질을 직접 평가함으로써 높은 구조 보존 능력을 보이는 것으로 알려져 있다. 그러나 특징 개수 d에 대해 탐색 공간이 $O(2^d)$으로 효율적인 탐색 전략이 필수적이다. 이에 따라 여러 진화적 탐색(evolutionary search) 기반 알고리즘이 활발히 연구되었다. Kim 등은 UFS 문제에 전역 탐색 기법을 도입하였고, 이를 위해 GA 기반의 방식을 제시하였다 [2]. 후보 특성 집합을 직접 탐색하고, 군집 품질을 평가할 수 있는 적합도 함수를 제시하였다. 그러나 기존 GA 기반 접근은 변이(mutation) 및 교차(crossover)를 통해 전역 탐색 성질은 확보했지만 수렴 속도가 느리거나 지역해(local optimum)에 빠질 위험이 존재한다. GA 외에도 개미 군집 최적화(Ant Colony Optimization, ACO) 를 활용한 UFS 연구가 제안되었다 [7]. 개미 개체의 경로 탐색 메커니즘을 기반으로 특성 선택 문제를 해결하였다. ACO 기반 방법 역시 높은 계산 복잡도와 수렴 속도 측면의 한계가 존재하며, 탐색 과정이 데이터 차원의 증가에 따라 급격히 비효율적으로 변화할 수 있다. 다른 탐색 방식으로 입자 군집 최적화(Population Swarm Optimization, PSO) 기반 방법이 제안되었다 [8]. 이 방법은 파티클을 명시적 속도 벡터 대신 확률적 분포에서 샘플링하는 bare-bones 전략을 사용하여 탐색 효율을 높였다. 하지만 특성 간 상호작용이나 조합 효과를 충분히 반영하지 못하는 한계가 남아 있다.

III. The Proposed Scheme

1. Preliminary

UFS는 레이블 정보 없이도 데이터의 내재적 구조를 가장 잘 대표하는 유용한 특징 부분집합을 찾는 것을 목표로 한다. 데이터 행렬 $X \in \mathbb{R}^{n \times d}$에서 n은 샘플 수, d는 특징 수를 나타낸다. UFS의 목적은 부분 집합 $S \subseteq \{1,2,...,d\}$ ($|S| \ll d$)을 찾아 X가 포함하는 정보의 대표성(representativeness)과 다양성(diversity)을 최대화하는 것이다. 레이블이 존재하지 않는 환경에서는 특징 부분집합의 품질을 평가하기 위해 클러스터링의 결과를 통해 얻은 클러스터간의 거리 정보 등을 활용한다. UFS의 최적화 문제는 조합 최적화 문제로 휴리스틱 탐색 기반의 기법이 필요하게 된다.

UFS 환경에서 GA는 각 염색체(chromosome)가 후보 특징 부분집합을 인코딩하며, 선택(selection), 교차(crossover), 변이(mutation)와 같은 진화 연산자를 통해 개체군(population)이 점차 더 높은 품질의 특징 집합으로 진화하도록 설계된다. 미분 가능한 목적 함수를 요구하는 경사 기반 방법과 달리, GA는 불연속 탐색 공간을 다룰 수 있으며, 확률적 재조합과 변이를 활용해 지역 최소값에 갇히는 문제를 효과적으로 방지할 수 있다 [2].

그러나 유전 알고리즘의 표준적인 연산은 특징 선택 문제에 본질적으로 적합하지 않은 한계를 갖는다. 특히 교차 연산자는 특징 간 중복성(redundancy)을 고려하지 못한 채, 개별적으로 유리해 보이는 특징을 그대로 보존하는 경향이 있다. 이와 같은 한계를 해결하고 선택된 특징 집합의 판별력을 향상시키는 동시에 부분집합의 크기를 효과적으로 제어하기 위해, 본 연구에서는 유전 알고리즘 개체군의 각 염색체에 정보이론 기반의 정제(refinement) 방식을 도입한다. 제안하는 정제 방식은 상호정보량(mutual information) 기준에 따라 특징을 적응적으로 추가하거나 제거함으로써, 진화적 탐색 과정이 보다 정보량이 높은 특징 부분집합을 향해 수렴하도록 유도한다.

확률 변수 X의 불확실성 정도는 샤논(Shannon)의 엔트로피(entropy)를 통해 정량화할 수 있으며, 이는 다음과 같이 정의된다.

$$H(X) = -\sum P(X) \log P(X) \qquad (1)$$

여기서 $P(X)$는 변수 X의 확률질량함수(probability mass function)를 의미한다. 엔트로피는 X의 결과를 기술하는 데 필요한 기대 정보량(expected amount of information)을 나타내므로, 정보이론에서 불확실성을 표현하는 가장 기본적인 척도로 사용된다.

2. Chromosome Refinement

특징 집합을 $F=\{f_1,f_2,...,f_d\}$라고 하자. 정보이론적 관점에서 비지도 특징 선택의 목표는 전체 특징 집합 F와 선택된 부분집합 S 사이의 정보 손실을 최소화하는 것이다. 즉, $H(F)$와 $H(S)$의 차이를 최소화하는 것이 목적이다. 여기서 $H(F)$는 상수이며, 항상 $H(S) \le H(F)$이므로 목적 함수는 다음과 같이 표현될 수 있다.

$$\arg\max_{S} H(S) \tag{2}$$

유전 알고리즘의 관점에서 부분집합 S는 이진 염색체(binary chromosome) 형태로 인코딩되며, 각 비트는 해당 특징이 선택되었는지를 나타낸다. 주어진 염색체를 정제하기 위해서는 특징을 제거하거나 추가하는 방식이 사용될 수 있다. 정보이론적 관점에서 보면, 특징을 제거할 때 감소하는 엔트로피는 최소화해야 하며, 반대로 특징을 추가할 때 증가하는 엔트로피는 최대화해야 한다. 이는 염색체가 가능한 한 많은 정보를 유지하면서도 불필요한 특징을 억제하여 진화하도록 적응적으로 갱신한다.

이 때 하나의 염색체를 정제하기 위해 두가지 방법을 고려할 수 있다. 하나는 이미 선택된 특징들 중에서 일부를 제거하여 염색체를 정제하는 것이고, 다른 하나는 선택되지 않은 특징들 중에서 일부를 추가하여 정제하는 것이다. 하나의 염색체가 선택한 특징들을 S라 하자. S에서 특징 f^-를 제거하기 위한 목적 함수를 다음과 같이 나타낼 수 있다.

$$J_{DEL} = \arg\max_{f^- \in S} H(S \backslash f^-) \tag{3}$$

$H(S\backslash f^-)$는 고차원의 결합 엔트로피(joint entropy)이기 때문에 정확한 추정이 어렵다. 이 문제를 완화하기 위해 $H(S\backslash f^-)$를 저차원의 결합 엔트로피 항들로 표현하고자 한다. 이를 위해 k-카디널리티(cardinality) 엔트로피를 정의한다 [9].

$$U_k(X) = \sum_{Y \in X_k'} H(Y), \tag{4}$$

여기서 X'은 X의 부분 집합이며 X_k'는 다음과 같이 정의된다.

$$X_k' = \{e \mid e \in X', |e| = k\} \tag{5}$$

정의 식 (4)를 기반으로 Han의 부등식을 나타낼 수 있다 [10].

$$H(X) \le \frac{1}{n-1} U_{n-1}(X), \tag{6}$$

여기서 n은 X의 변수의 개수이다. 이 부등식을 기반으로 다음 새로운 부등식을 얻을 수 있다 [9].

$$U_k(S) \le \left(\frac{n-k+1}{k-1}\right) U_{k-1}(S) \tag{7}$$

이 부등식은 $U_k(S)$의 상한 값이 (k-1)-카디널리티 엔트로피로 결정될 수 있음을 보여준다. 이 부등식을 재귀적으로 응용하면 고차원의 결합 엔트로피를 k-카디널리티를 이용하여 추정할 수 있는 부등식을 얻을 수 있다 [11].

$$H(X) \le \left(\prod_{i=1}^{b} \frac{i}{n-i}\right) U_k(X), \tag{8}$$

여기서 $b = \min(n-k, k-1)$. 이 부등식은 k가 클수록 상한 값에 더 가깝게 근사한다는 것을 나타내기 때문에 $H(X)$를 더 잘 추정할 수 있음을 나타낸다. k가 1일 때는 단일 변수의 엔트로피가 된다. 우리는 다중 변수의 엔트로피를 고려하기 위한 최소의 값으로 k를 2로 설정했다.

목적 함수 식 (4)를 상한 부등식을 이용하여 표현하면 다음과 같다.

$$\begin{aligned} J_{DEL} &\approx \arg\max_{f^-} \frac{1}{|S|-2} U_2(S \backslash f^-) \\ &= \arg\max_{f^-} U_2(S \backslash f^-) \\ &= \arg\max_{f^-} \sum_{f_i, f_j \in S \backslash f^-} H(f_i, f_j) \end{aligned} \tag{9}$$

S에서 특징 f^+를 추가하기 위한 목적 함수를 다음과 같이 나타낼 수 있다.

$$J_{ADD} = \arg\max_{f^+ \in F \backslash S} H(S, f^+) \tag{10}$$

$H(S,f^+)$ 또한 고차원의 결합 엔트로피이기 때문에 정확한 추정이 어렵고, 결합 엔트로피의 상한 부등식 (8)을 활용하여 다음과 같이 목적 함수를 표현할 수 있다.

$$J_{ADD} \approx \arg\max_{f^+} \frac{1}{|S|} (U_2(S) + U_2(S \times f^+)) \tag{11}$$

여기서 $\times$는 두 집합 사이의 곱집합을 의미하고, $U_2(S)$는 상수이므로 다음과 같이 정리할 수 있다.

$$\begin{aligned} J_{ADD} &\approx \arg\max_{f^+} U_2(S \times f^+) \\ &= \arg\max_{f^+} \sum_{f \in S} H(f, f^+) \end{aligned} \tag{12}$$

3. The Proposed Method

염색체에 활용할 수 있는 DEL, ADD 정제 방식을 활용하여 비지도 특징 선별을 위한 유전 알고리즘을 설계할 수 있다.

Algorithm 1. Chromosome Refinement

Input: C: set of chromosomes, p: number of DEL operations, q: number of ADD operations
procedure Local Refinement(C, p, q)
  for i=1 **to** |C| **do**
    for j=1 **to** p **do**
      apply DEL to C_i using ()
    end
    for j=1 **to** q **do**
      apply ADD to C_i using ()
    end
  end
end

Algorithm 1은 염색체 집합에 대해 DEL, ADD를 적용하는 알고리즘을 나타낸다. 입력으로 염색체 집합 C에 대해 DEL, ADD 정제 방식을 각각 p, q번씩 적용한다.

Algorithm 2. Proposed Algorithm

Input: P: population of chromosomes, p: number of DEL operations, q: number of ADD operations, h: number of chromosomes for refinement, T: maximum number of generation
procedure Proposed Algorithm(P, p, q, h, T)
  initialize $P(0)$
  evaluate $P(0)$
  for t=1 **to** T **do**
    select top-h chromosomes C in $P(t-1)$
    create $N_1(t)$ from C using Algorithm 1
    create $N_2(t)$ from $P(t-1)$ using genetic operators
    evaluate $N_1(t)$ and $N_2(t)$
    add $N_1(t)$ and $N_2(t)$ to $P(t-1)$
    select $P(t)$ from $P(t-1)$
  end
end

Algorithm 2은 비지도 특징 선별을 위해 제안하는 알고리즘으로 염색체 정제 알고리즘 Algorithm 1을 활용한 진화 알고리즘이다. 초기 염색체 집합 P는 세대가 반복됨에 따라 교차, 변이, 염색체 정제로 새로운 염색체들이 생성되고, 평가를 통해 정해진 크기의 염색체를 선정하여 세대수를 유지한다. $P(t)$는 t번째 세대를 나타내며 세대 생성의 반복은 정해진 횟수 T만큼 반복된다.

IV. Experiments

제안하는 비지도학습 특징 선별 알고리즘의 성능을 판단하기 위해 실험을 진행한다. 실험 비교를 위해 여섯 개의 데이터를 사용하였다. COIL20은 20개의 사물에 대해 서로 다른 회전 각도에서 촬영한 1,440장의 32×32 회색조 이미지로 구성된 물체 이미지 데이터셋이다 [12]. Lung 데이터는 203개의 샘플과 3,312개의 유전자로 구성된 폐암 관련 마이크로어레이 유전자 발현 데이터셋이다 [13]. Lymphoma (Lymph) 데이터는 96개의 샘플과 4,026개의 유전자를 포함하며 9개의 림프종·정상 세포 유형으로 구성된 고차원 유전자 발현 데이터셋이다 [14]. UMIST 데이터는 20명의 피험자에 대해 다양한 포즈에서 촬영한 575장의 92×112 얼굴 이미지로 구성된 멀티뷰 얼굴 데이터셋이다 [15]. USPS 데이터는 16×16 크기의 회색조 손글씨 숫자 이미지 9,298장(0-9)으로 구성된 손글씨 숫자 인식용 표준 데이터셋이다 [16]. YaleB 데이터는 38명의 피험자를 다양한 조명 조건에서 촬영한 2,414장의 얼굴 이미지로 구성된 조명 변화 얼굴 데이터셋이다 [17]. 이미지 데이터의 경우 일반적으로 특징 선별이 의미가 없지만, 실험에 사용한 UMIST, USPS, YaleB같은 경우 아주 이미지의 크기가 작기 때문에 픽셀 단위에서의 특징 선별이 의미를 가질 수 있다. 예를 들어 YaleB 데이터는 32×32의 크기로 얼굴 이미지로 픽셀 단위에서 눈, 코, 귀 등의 위치가 나타날 수 있고 이는 얼굴 인식의 중요한 특징으로 볼 수 있다. 데이터의 명세는 Table 1에 정리되었다.

Table 1. Used Datasets in Experiments

Data	# of Samples	# of Features	# of Classes
COIL20	1,440	1,024	20
Lung	203	3,312	5
Lymph	96	4,026	9
UMIST	575	644	20
USPS	9,298	256	10
YaleB	2,414	1,024	38

제안하는 정제 방식을 GA와 비교하였다. 제안하는 방법과 GA는 50개의 염색체로 세대를 구성하고, 200개의 새로운 염색체를 생성하는 과정을 거쳤다. 세대의 염색체 수는 너무 적을 경우 충분한 다양성을 얻기 어렵고 너무 많을 경우 탐색의 효율성이 떨어지게 되어 적당한 개수가 중요한데, 실험적으로 50개로 설정하였다. 이 때 사용된 염색체 평가 방식(fitness)은 선택된 특징으로 k-means 클러스터링을 통해 얻은 결과를 바탕으로 평가한다. 이 방식은 클러스터 중심간의 거리, 샘플과 해당 샘플이 속하지 않은 다른 클러스터 중심간의 거리, 선택된 특징의 개수를 기반으로 결정된다 [2]. 이 값은 낮을수록 좋은 클러스터링 결과를 의미한다. 선택된 특징의 우수성을 평가하기 위해 k-means의 결과를 이용하여 클러스터링 정확도

(clustering accuracy, CLACC), 정규화된 상호정보량(normalized mutual information, NMI)를 사용했다[18]. 사용되는 알고리즘들의 무작위성 때문에 모든 실험은 30번 반복 실험하였고, 평균 값을 기입하였다.

Table 2. Comparison of Clustering Accuracy

Data	GA	Proposed
COIL20	0.5772 (±0.0464)	**0.6001 (±0.0617)**
Lung	0.6865 (±0.0939)	**0.6979 (±0.1085)**
Lymph	0.5122 (±0.0709)	**0.5163 (±0.0912)**
UMIST	0.4113 (±0.0258)	**0.4181 (±0.0282)**
USPS	0.5673 (±0.0421)	**0.5846 (±0.0529)**
YaleB	0.0905 (±0.0050)	**0.0923 (±0.0056)**

Table 2와 Table 3은 제안한 방법과 기존 GA 기반 방법을 다양한 벤치마크 데이터셋에서 비교한 결과를 보여준다. 데이터 별로 우수한 성능을 보인 결과는 진하게 표시하였다. 괄호 안에는 표준편차를 나타낸다. CLACC 결과(Table 2)를 보면, 모든 데이터셋에서 제안한 방법이 GA 대비 더 우수한 정확도를 보였다. 특히 COIL20, USPS와 같은 중간 규모의 이미지 데이터셋에서 비교적 큰 향상을 확인할 수 있다. Lymph, UMIST, YaleB의 경우에도 제안방법은 GA와 유사하거나 소폭 개선된 성능을 유지하였다. YaleB 의 경우 클래스의 개수가 많기 때문에 군집 구조가 명확히 식별되기 어려워 클러스터링 정확도가 상대적으로 낮게 나타나는 경향이 있었다. NMI 비교(Table 3)에서도 전반적으로 제안방법이 높은 값을 보이며, 선택된 특징 집합이 더 일관된 클러스터 구조를 형성함을 확인할 수 있다. 특히 Lymph, YaleB 데이터셋에서는 GA 대비 NMI가 확연히 향상되었는데, 이는 제안된 평가 방식이 클러스터 간 분리도와 특징 선택 개수를 균형있게 반영한 결과로 해석된다. 특히 YaleB의 NMI의 경우 t-검정 결과 유의수준 5%에서 귀무가설을 기각하는 것으로 나타나 통계적 유의미함을 보였다. 신뢰도 이러한 결과는 제안한 방법이 기존 GA 기반 접근법보다 더 안정적이며, 특징 선택 과정에서 클러스터 품질을 효과적으로 반영함으로써 클러스터링 성능을 개선함을 보여준다.

Table 3. Comparison of Normalized Mutual Information

Data	GA	Proposed
COIL20	0.8843 (±0.0365)	**0.8903 (±0.0342)**
Lung	**0.6107 (±0.0869)**	0.6091 (±0.0829)
Lymph	0.6998 (±0.0649)	**0.7226 (±0.0961)**
UMIST	0.8181 (±0.0293)	**0.8193 (±0.0327)**
USPS	0.7919 (±0.0456)	**0.7968 (±0.0486)**
YaleB	0.7241 (±0.0067)	**0.7290 (±0.0070)**

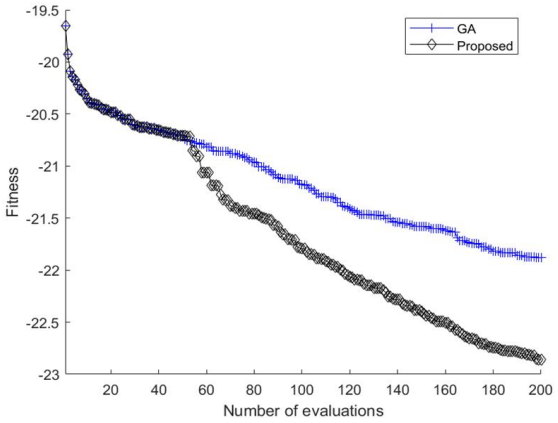


Fig. 1. Comparison of convergence behavior between GA and the proposed method on the COIL20 dataset in terms of fitness evaluations.

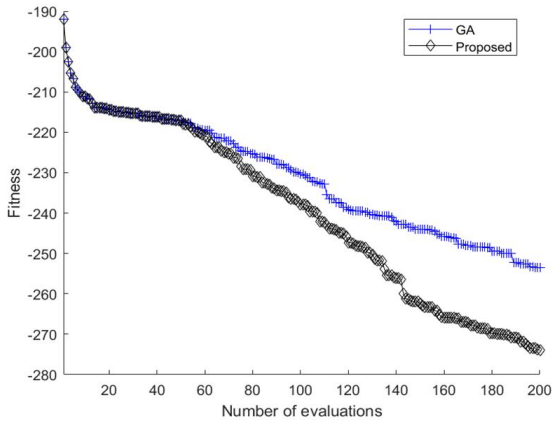


Fig. 2. Comparison of convergence behavior between GA and the proposed method on the USPS dataset in terms of fitness evaluations.

Fig. 1과 Fig. 2는 각각 COIL20 데이터와 USPS 데이터에 대해 제안하는 방법과 GA의 fitness 수렴도를 보여준다. 이를 위해 초기 염색체 집합과 k-평균 알고리즘의 초기점을 통일하였다. 따라서 초기 세대 50개의 염색체의 결과는 동일하고, 이후에 수렴도가 달라진다. 초기 50개는

무작위로 결정되었기 때문에 이후 탐색 과정에서 GA와 제안하는 방법이 모두 수렴도가 좋아지는 것을 볼 수 있다. 하지만 평가 횟수가 증가할수록 제안된 방법이 더 빠르고 안정적으로 낮은 fitness 값을 달성하는 것을 확인할 수 있다. 특히 COIL20(Fig. 1)의 경우 제안 방법은 초기 세대 이후 GA보다 지속적으로 더 낮은 fitness를 기록하며 더 우수한 탐색 효율을 보인다. USPS 데이터셋(Fig. 2)에서도 유사한 경향을 확인할 수 있는데, GA는 중간 단계에서 수렴 속도가 완만해지는 반면 제안된 방법은 꾸준히 fitness 값을 감소시키며 평가 후반부에서 GA 대비 뚜렷한 성능 우위를 확보한다. 이러한 결과는 제안된 방법이 기존 GA 기반 접근법보다 전반적으로 더 안정적이고 빠른 수렴 특성을 가진다는 것을 보여주며, 특징 선택 문제에서도 효율적인 탐색 능력을 제공함을 확인해준다.

Table 4. Comparison of Execution Time (seconds)

Data	GA	Proposed
COIL20	32.0	41.9
Lung	11.3	58.8
Lymph	5.6	71.2
UMIST	9.2	10.9
USPS	162.4	138.6
YaleB	97.1	110.3

Table 4는 GA와 제안하는 방법의 측정된 학습 시간을 나타낸다. 실험 환경은 인텔 i7(12세대), 64GB, MATLAB 2023b으로 구성되었다. 초기 세대수를 100, 최대 생성 세대수는 3000으로 설정하였다. 제안하는 방법은 세대 생성 이전에 특징 간의 엔트로피를 미리 계산하여 평가 과정에 활용하는데, 이 연산은 특히 특징 수가 매우 많은 데이터셋에서 추가적인 비용을 요구한다. 하지만 Lung, Lymph을 제외하면 제안하는 방법과 GA 사이의 시간 소모가 비슷함을 알 수 있다. 반면 USPS 데이터셋에서는 제안된 방법이 GA보다 더 빠르게 동작하였다. USPS는 상대적으로 샘플 수가 매우 많은 데이터셋으로, fitness 평가를 위해 수행되는 k-평균 클러스터링의 반복 실행이 전체 계산 비용을 크게 증가시킨다. 이 과정에서 상대적으로 제안하는 방법의 특징 개수가 적게 선택되어 전체 실행 시간이 GA보다 짧아지는 결과를 보인다. 특징 수가 매우 큰 데이터셋에서는 제안된 방법의 엔트로피 사전 계산 비용이 증가하지만, 샘플 수가 많은 데이터셋에서는 오히려 제안된 방법이 상대적으로 효율적인 실행 시간을 보일 수 있다.

V. Conclusions

본 연구에서는 비지도 학습 환경에서 효율적이고 정보 보존 능력이 높은 특징 부분집합을 찾기 위해, GA 기반의 래퍼 구조에 정보이론 기반 염색체 정제 연산자를 결합한 새로운 진화적 특징 선택 방법을 제안하였다. 제안된 방법은 전역 탐색 능력을 유지하면서도, 상호정보량 기준을 활용한 특징의 정교한 추가·제거 과정을 통해 빠른 수렴과 높은 구조 보존 성능을 달성하였다. 다양한 데이터셋을 대상으로 한 실험 결과는 기존 GA 기반 UFS 알고리즘보다 우수한 탐색 품질과 더 작은 특징 집합을 출력함을 보여주었다. 이를 통해 제안 방법이 비지도 특징 선택 문제에 효과적인 대안이 될 수 있음을 확인하였다.

향후 연구로 제안한 염색체 정제 연산자를 다른 진화적 메타휴리스틱과 결합하여 더 강건한 탐색 구조를 구현할 수 있을 것이다. 또한 대규모 고차원 데이터에서의 효율성을 높이기 위해 엔트로피 및 상호정보량 계산의 근사 기법을 도입하는 것도 중요한 확장 방향이다. 마지막으로, UFS 결과를 이상탐지 등과 같은 직접적인 응용과 연계하여 보다 실용적인 응용 성능을 검증하는 후속 연구가 필요하다.

REFERENCES

[1] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," Artificial Intelligence Review, Vol. 53, No. 2, pp. 907-948, 2020. DOI: 10.1007/s10462-019-09682-y

[2] Y. S. Kim, W. N. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.

[3] S. Alelyani, J. Tang, and H. Liu, "Feature selection for clustering: A review," Data Clustering, pp. 29-60, 2018.

[4] S. Tabakhi, et al., "Gene selection for microarray data classification using a novel ant colony optimization," Neurocomputing, Vol. 168, pp. 1024-1036, 2015. DOI: 10.1016/j.neucom.2015.05.022

[5] S. Wang, J. Tang, and H. Liu, "Embedded unsupervised feature selection," Proc. of the AAAI Conference on Artificial Intelligence, Vol. 29, No. 1, 2015.

[6] J. Guo and W. Zhu, "Dependence guided unsupervised feature selection," Proc. of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.

[7] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization,"

Engineering Applications of Artificial Intelligence, Vol. 32, pp. 112-123, 2014. DOI: 10.1016/j.engappai.2014.03.007

[8] Y. Zhang, et al., "A filter-based bare-bone particle swarm optimization algorithm for unsupervised feature selection," Applied Intelligence, Vol. 49, No. 8, pp. 2889-2898, 2019. DOI: 10.1007/s10489-019-01420-9

[9] J. Lee and D.-W. Kim, "Mutual information-based multi-label feature selection using interaction information," Expert Systems with Applications, Vol. 42, No. 4, pp. 2013-2025, 2015. DOI: 10.1016/j.eswa.2014.09.063

[10] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," Information and Control, Vol. 36, pp. 133-156, 1978.

[11] W. Seo, D.-W. Kim, and J. Lee, "Generalized information-theoretic criterion for multi-label feature selection," IEEE Access, Vol. 7, pp. 122854-122863, 2019. DOI: 10.1109/ACCESS.2019.2927400

[12] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-100)," Technical Report CUCS-006-96, Columbia University, 1996.

[13] A. Bhattacharjee, et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," Proc. of the National Academy of Sciences, Vol. 98, No. 24, pp. 13790-13795, 2001.

[14] A. A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, Vol. 403, No. 6769, pp. 503-511, 2000. DOI: 10.1038/35000501

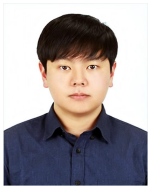
[15] D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," Face Recognition: From Theory to Applications, Springer, pp. 446-456, 1998.

[16] J. J. Hull, "A database for handwritten text recognition research," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No. 5, pp. 550-554, 2002. DOI: 10.1109/34.291440

[17] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 6, pp. 643-660, 2002. DOI: 10.1109/34.927464

[18] H. Lim and D.-W. Kim, "Pairwise dependence-based unsupervised feature selection," Pattern Recognition, Vol. 111, pp. 107663, 2021. DOI: 10.1016/j.patcog.2020.107663

Authors



Hyunki Lim is currently an assistant professor in the Division of AI Computer Science and Engineering, Kyonggi Univ. (KGU) in Suwon, Korea. Prior to coming to KGU, he did his postdoc at KIST, and Ph.D., M.S. and B.S. at Chung-Ang Univ., Korea. His research interest includes advanced machine algorithms and related optimization methods with innovative applications such as music emotion recognition and smart factory.