다양한 클러스터 결과에 의해 진화적 접근법을 사용하는 이종 클러스터링 앙상블 기법

윤혜성[O]* 안선영* 이상호* 조성범** 김주한**
이화여자대학교 컴퓨터학과*, 서울대학교 의과대학 생명의료정보학연구실**
comet@ewhain.net, lovesy@ewhain.net, shlee@ewha.ac.kr, csb@medigate.net, juhan@snu.ac.kr

Heterogeneous Clustering Ensemble Method using Evolutionary Approach with Different Cluster Results

Hye-Sung Yoon[O]* Sun-Young Ahn* Sang-Ho Lee* Sung-Bum Cho** Ju Han Kim**
Dept. of Computer Science and Engineering, Ewha Womans University*, Seoul National University Biomedical Informatics(SNUBI), Seoul National University College of Medicine**

요 약

데이터마이닝 기법의 클러스터링 알고리즘은 생물정보학에서 데이터 셋의 사전 정보를 고려하지 않고 중요한 유전적, 생물학적 상호작용을 찾기 위하여 적용되고 있다. 그러나 다양한 형식의 수많은 알고리즘들은 바이오데이터의 다양한 특성들과 실험의 가정 때문에 다른 클러스터링 결과들을 만들 수 있다. 본 논문에서는 바이오 데이터 셋의 특성에도 적합하면서 양질의 클러스터링 결과를 만들기 위한 새로운 방법을 제안한다. 이 방법은 여러 가지 클러스터링 알고리즘의 결과들을 유전자 알고리즘의 기본 개념인 진화적 환경에서 가장 적합한 형질을 선택하는 문제와 결합하였다. 그리고 실제 데이터 셋을 이용하여 우리의 제안하는 방법을 증명하고 실험 결과로 최적의 클러스터 결과를 보인다.

1. 서 론

데이터 마이닝 알고리즘들은 대용량의 데이터베이스에서 흥미 있는 패턴들을 발견하기위하여 광범위하게 이용되어지고 있는 접근방법이다[1]. 그 중 클러스터링 분석 방법은 데이터 셋에 있는 원소(element)들을 유사성에 따라서 그룹핑하는 것으로 데이터 셋에 대한 클래스 레이블 정보를 요구하지 않는다. 그리고 대부분의 바이오 데이터들이 본래 사전 지식이 많지 않기 때문에 바이오데이터 셋에서 더 나은 유전적 이해와 생물학적 정보를 얻기 위하여 클러스터링 알고리즘을 적용하는 경향이 있다. 그러나 바이오 데이터를 분석할 때에 클러스터링 알고리즘들은 각자 서로 다른 특징들을 가지고 적용되어 진다[2]. 이것은 최고의 알고리즘을 선택하는 것이 어려운 문제로서, 서로 다른 형식의 클러스터링 방법이 데이터 셋에 적용되면 그들의 다양한 특징들 때문에 일관되지 않는 결과를 만들 수 있다는 것을 말한다. 따라서 최근에 클러스터링 앙상블 기법 즉, 바이오 데이터의 특징은 잃지 않으면서 다양한 파티션들을 결합하는 문제가 소개되었다[3]. 이것은 클러스터링 결과들의 가용성과 신뢰성을 향상시키기에 유용하다는 것을 보이고 하나의 승리(winning) 파티션을 선택하는 것 보다 다른 클러스터링 알고리즘들의 결과 클러스터들을 결합하는 것이 더 나은 결과를 도출한다는 것을 보이고자 한 방법이다. 따라서 본 논문에서는 다양한 클러스터링 알고리즘으로부터의 서로 다른 결과 클러스터들을 유전자 알고리즘에 기반하여 결합하는 이종 클러스터링 앙상블 프레임워크를 새롭게 보인다.

논문의 구성은 다음과 같다. 2장에서 제안하는 방법을 위한 기존 연구와 3장에서는 제안하는 알고리즘을 실제 데이터에 적용한 방법을 설명한다. 그리고 제안한 방법을 적용한 실험 결과와 결론을 각각 4장과 5장에서 설명한다.

2. 관련 연구

본 장에서는 관련 연구로서 바이오데이터 분석을 위한 클러스터링 방법과 클러스터링 알고리즘 적용결과를 이용하는 클러스터링 앙상블 방법에 대하여 설명한다.

2.1 바이오데이터 클러스터링

기술의 발전으로 지놈들에서 동시에 수천개의 유전자들의 발현을 모니터하는 것이 가능하게 되었고, 그 기술력으로 대량의 정보를 효과적으로 분석하고 해석할 수 있게 되었다. 따라서 대량의 정보를 다루는 것이 중요한 문제 중에 하나가 되었다. 클러스터링 분석 방법은 바이오데이터의 분석을 위하여 하나의 유용한 탐험적 기법이다. 그러나 하나의 방법이 바이오데이터 분석에 최선의 선택적 방법이라고 대표할 수 없다. 왜냐하면, 대부분의 제안 되어지는 클러스터링 알고리즘들은 상당히 휴리스틱하게 제안되어 지고 있고, 정확한 수의 클러스터를 결정하고 최선의 클러스터링 알고리즘을 선택하는 것이 아직까지 풀기에는 어려운 문제이기 때문이다. 이것은 접근방법의 어떠한 것도 충분하지가 않다는 것과 다양한 기법들의 적용은 탐구하고자하는 데이터들의 여러 가지 다른 면들이 인정되어야 한다는 것을 명시한다. 따라서 우리는 다양한 클러스터링 알고리즘의 단점들은 버리고, 그것들의 결과 클러스터들을 효과적으로 결합하기위한 새로운 방법을 제안한다.

2.2 클러스터링 앙상블

양질의 클러스터링 결과를 만드는 것은 실험 데이터의 잡음과 서로 다른 형식의 클러스터링 알고리즘 사이의 불일치 때문에 매우 어려운 문제이다. 따라서 최근 연구는 클러스터링 알고리즘의 장점들을 조합하여 더 나은 결과를 이끌어내는 클러스터링 앙상블 기법이 적용되어지고 있다. 그러나 같은 알고리즘일지라도 다양한 반복과 무작위 시작으로 다른 결과들을 초래하기 때문에, 어떠한 클러스터링 결과가 최적이라고 말하기 어렵다. 클러스터링 앙상블의 주요 이슈중에 하나가 어떻게 서로 다른 클러스터링 결과들을 결합하느냐이다. 이전 연구들은 클러스터링 알고리즘으로부터의 클러스터링 결과를 같은 수로 조절하였다[4]. 그러나 같은 수의 클러스터링 결과들을 곧장 결합하는 것은 의미 있는 결과를 만들 수 없다. 따라서 다른 클러스터링 알고리즘의 다양한 결과인 다른 개수의 클러스터링 결과들을 결합하기위한 새로운 메카니즘이 더 나은 클러스터링 결과들을 얻기 위해서 필요하다. 따라서 본 논문에서는 여러 클러스터링 알고리즘의 효과적 결합은 최종 결과 클러스터들의 질을 향상시키기에 중요한 방법이라고 가정한다. 그리고 바이오데이터 셋의 더 나은 클러스터링 결과를 만들기 위해서 서로 다른 결과 클러스터들을 모아서 최적의 정보를 찾는 것에 초점

을 맞추고, 최적의 클러스터 결과를 보인다.

3. 알고리즘

본 장에서는 적용한 GA의 기본 개념에 대하여 설명하고, 제안하는 전체적인 프레임워크 순서대로 설명한다.

3.1 유전자 알고리즘

여러 클러스터링 알고리즘들의 결과 클러스터들에 초점을 맞추어서 고유한 지식을 발견하는 과정을 수행하기 위하여 GA를 이용하는 것을 설명한다. 실제 유전자 알고리즘은 더 나은 탐색을 위해서 이전에 얻어진 정보들을 효과적으로 활용하는 선택적 방법을 제공한다[5]. GA는 염색체(chromosome) 또는 연쇄(chain)라고 말하는 것을 다루며, 복사(copy)와 부연산(subchain) 교환과 같은 연쇄를 처리하는 것을 실행한다. 다음과 같은 유전적 연산들이 본 논문에서 적용되었다.: 재생산(Reproduction), 교차(Crossover)

3.1.1 재생산 단계

데이터 셋에 여러 가지 클러스터링 알고리즘들을 적용하고 모든 클러스터링 결과들에서 두개의 클러스터가 하나의 쌍이 되도록 서브 셋을 구성하였다. 이 재생산 단계는 최적화 함수에 따라서 복사되어지는데, 적합도 함수 값보다 큰 한쌍의 염색체는 자식을 만들 때 더 큰 확률이 주어지며, 각 염색체가 생존할 것인지를 결정한다. 본 논문에서는 교차 연산을 위한 한쌍을 선택하는 적합도 함수로, 결과 클러스터들의 원소들이 가장 많이 겹쳐서 가장 많이 겹치는 수를 가지는 한 쌍을 선택한다. 이렇게 각 클러스터마다 대표 값을 더하여 모집단의 모든 쌍의 대표 값들을 비교하여 교차 연산을 수행할 하나의 서브 셋을 선택한다.

3.1.2 교차연산 단계

교차 연산의 목적은 두 부모로부터 자식을 만드는 데에 가능한 많이 의미 있는 부모 정보를 상속하고자 하는 것이다. 그림 1은 본 논문에서 제안하는 교차 방법을 설명한다. 예를 들어, 그림 1에서, 모집단에서 A와 K가 두 부모로 선택되었다. 각 부모는 전체 원소들을 가지고 첫 번째 부모가 A_1, A_2, A_3 3가지 클러스터링결과와 두 번째 부모가 K_1, K_2, K_3, K_4, K_5인 5가지 클러스터링 결과를 가진다. 먼저, 우리는 첫 번째 부모로부터 두개의 클러스터들 A_2, A_3 보다 두 번째 부모의 클러스터들 사이에서 가장 중복이 많이 되는 원소들을 가진 클러스터 A_1을 선택한다. 그런 다음, A_1은 두 번째 부모의 클러스터를 가운데 A_1 (원소들 7, 27, 39, 58, 63, 65, 71 and 84)과 가장 많은 수의 중복이 되는 원소들을 가지는 K_5와 교차 연산을 적용한다. 그리고 예를 들어, 교차 연산을 적용 할 때, A_1에 있는 원소들 중에서 K_5와는 중복이 되지 않는 원소들 (63, 71 and 84) 가운데 두 번째 부모의 다른 클러스터 K_2에 63 원소가 나타나고, K_4 에 84 원소가 나타나면, 모든 원소가 단지 하나의 클러스터에만 속하게 하기위해서 제거한다. 그리고 A_1에 남아있는 나머지 다른 원소들 (그림 1에서는 71)은 K_5에 교체할 때 그대로 가져간다. 마지막으로, 첫 번째 자식으로 표현되어지는 새로운 결과 클러스터는 K_1, K_2, K_3, K_4, 변경된 A_1이다. 이러한 교차 연산이 한번 더 같은 방법으로 두 번째 부모로부터 두 번째 자식을 만들기 위해서 하나의 클러스터를 선택함으로서 반복된다.

이러한 과정은 여러 가지 클러스터링 알고리즘을 적용한 다른 클러스터링 결과인 클러스터들의 원소들이 변하는 것으로, 중복이 많이 되면서 의미 있는 정보를 가지는 원소들이 자식으로 전달되어 최적화된 최종 클러스터링 결과를 얻게 될 것이다.

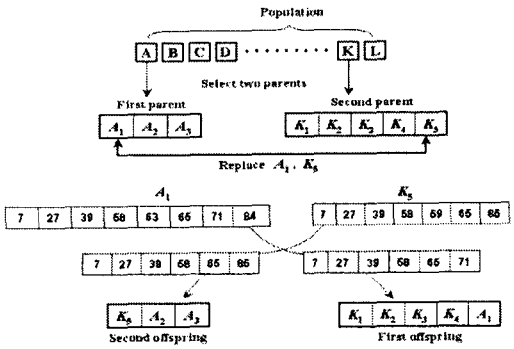

그림 1. 결과 클러스터들을 교환하기 위한 교차연산

3.2 이종 클러스터링 앙상블

제안하는 유전자 알고리즘 연산방법에 기반한 이종 클러스터링 앙상블 접근법의 기본 아이디어는 다음과 같다. 먼저 여러 타입의 클러스터링 알고리즘들이 실험 데이터에 적용된다. 그리고 서로 다른 클러스터링 결과들에 대해서 3.1절의 과정을 반복하여 최적의 클러스터링 결과 셋으로 만든다. 다른 수의 클러스터링 결과들을 결합하기위한 제안하는 유전자 알고리즘에

Algorithm: Heterogeneous Clustering Ensemble (HCE)

Input :

(1) The data set of N data points $D = X_1, X_2, \ldots X_N$
(2) A set of clustering algorithms, K
- i : the number of clustering algorithms available for analysis

(3) The different cluster results, C_j
- K_i generates different clustering results, C_j, from the data set D

(4) The cluster result is $S = \{Sk_1c_j, Sk_2c_j, \ldots, Sk_ic_j\}$
- Sk_ic_j are different numbers of cluster results, C_j, of the i^{th} algorithm

Output :

The optimal cluster results on the dataset D

1. Run clustering algorithms K_i on D
2. Construct a paired non-empty subset, $SM^{(g)}$, with only two clusters, from the cluster results, S
3. Iterate n until convergence (permute the cluster results of the data for every iteration) :
   3.1 Compute the fitness function $F(t)$ and select two parents/subsets from $SM^{(g)}$
   3.2 Crossover two parents
   - compare between the first parent cluster results and the second parent cluster results
   - replace the clusters of the second parent based upon the first parent which has the largest number of highly-overlapped elements
   - repeat again clusters replacement with the first parent based upon the second parent

   3.3 Replace $SM^{(g)}$ parent subsets by newly-created offspring

기반한 이종 클러스터링 앙상블 알고리즘은 아래와 같다.

본 논문의 실험에서는 환자들 사이에의 연관성을 찾고자 하였다. 따라서 알고리즘의 인풋은 환자들을 기준으로 적용하였다. 그 결과 아웃풋은 유사한 환자들의 클러스터링 결과를 보인다. 알고리즘 1단계에서는 인풋 데이터에 여러 가지 다른 클러스터링 알고리즘을 적용한다. 그 결과 다른 클러스터링 알고리즘들의 클러스터링 결과들을 S라 하였을 때, S에서 2개의 원소들을 하나의 쌍으로하는 $SM(g)$를 구성한다. 3단계부터가 이종 클러스터링 앙상블 알고리즘에 유전자 알고리즘을 적용하는 단계인데, $SM(g)$을 모집단으로 보고 적합도를 평가하기 위하여 가장 많이 중복이 되는 원소를 가지는 한쌍을 교차 연산을 실행하기 위한 부모로 선택한다. 그리고 4번째 단계에서 이러한 교차 연산 과정을 수행한 후에 만들어진 자식들로 모집단의 두 부모를 새로 생성된 자식들로 교체하고 최적의 클러스터 서브셋인

$SM(g)$를 만들기 위하여 이러한 과정을 계속 반복한다.

4. 적용

본 장에서는 논문에서 적용된 실험 데이터와 실험 결과에 대하여 설명한다.

4.1 실험 데이터

본 논문에서는 제안하는 방법을 적용하기 위한 데이터 셋으로 CAMDA(Critical Assessment of Techniques for Microarray Data Analysis) 2006 conference 데이터 셋 (http://www.camda.duke.edu/camda06/datasets)을 사용하였다. 이 데이터 셋은 chronic fatigue syndrome(CFS)에 대하여, microarray, proteomics, SNP, clinical data로 구성되어 있다. 본 실험에서는 데이터의 두 가지 타입, microarry와 clinical data를 각각 제안하는 방법의 적용과 입증을 위하여 이용하였다. 첫 번째 마이크로어레이 데이터는 20,160 유전자에 대하여 177명의 DNA를 single-channel로 실험한 데이터 셋이고, 두 번째 데이터는 227명의 환자를 실험 기준에 의하여 CFS에 대하여 3가지 서브그룹, least, middle, worst로 분류하여 놓은 데이터 셋이다. 우리는 제안하는 방법의 평가를 위하여, 마이크로어레이 데이터와 clinical data사이에 동일한 참여자이면서 CFS에 대하여 이미 분류된 3가지 그룹 중에 하나로 분류되어 있는 118명의 환자 데이터 셋을 적용하였다. 우리의 분석 접근 방법은 CFS 병의 통찰력을 이끌어 내기 위함이 아니다. 이 연구는 우리가 CFS 같이 여러 가지 발병원인을 가지고 있고 어떠한 분석 방법을 이용해야 하는지 결정하기 어려울 때 적용할 수 있는 방법을 보이고자 한다. 따라서 제안하는 방법은 하나의 알고리즘을 적용하는 것 보다 다양한 형식의 클러스터링 알고리즘들을 적용하고 다양한 각도에서 데이터 셋을 분석한다.

4.2 실험 적용과 결과

본 논문에서는, 데이터마이닝 기법을 제공하는 AVADIS 분석 도구 (http://avadis.strandgenomics.com)를 이용하여 3가지 다른 클러스터링 알고리즘들의 여러 가지 파라미터를 변형하여 여러 가지 클러스터링 결과들을 만드는데 적용하였다. 그리고 이렇게 만들어진 결과와 우리의 제안하는 방법을 비교하였다.

우리는 데이터 실험 분석과 제안 알고리즘의 효과를 평가하기 위해서, 제안하는 알고리즘에서는 같은 인풋의 마이크로어레이 데이터 파일을 가지고 유전자 알고리즘 적용 단계에서 10,000번의 교차 연산을 반복 수행한 결과 가장 적합도가 높은 클러스터를 가지는 클러스터링 결과가 4개일 때임을 발견하였다. 따라서, 3개의 다른 클러스터링 알고리즘을 이용하여 4개의 클러스링 결과를 만든 것과 제안하는 방법을 적용하여 만들어진 4개의 클러스터들의 결과를 비교하였다. 유전자 알고리즘에 기반한 이종 클러스터링 앙상블을 방법을 적용하여 3가지 클러스터링 알고리즘의 클러스터링 결과와 제안하는 방법의 최종 클러스터링 결과들과 비교한 것은 표 1과 같다. KM, HC, PCA, HCE는 각각 k-means, hierarchical 클러스터링, PCA에 기반한 클러스터링, 제안하는 이종 클러스터링 앙상블 방법을 말한다. 그리고 테이블에서 보는 것과 같이 clinical 데이터에서 정의된 클러스터 그룹을 알지 못할 때, 기존의 클러스터링 알고리즘을 마이크로어레이 데이터 셋에 적용한 클러스터 결과들은 같은 클러스터로 묶인 원소들의 집합이 실제로 분류된 클러스터 그룹의 원소들의 분류와 많이 일치하지 않는 경향을 보인다. 그러나 우리가 제안하는 방법인 HCE 방법이 clinical 데이터에서 분류된 클러스터 그룹과 거의 일치하는 것을 발견할 수 있다. 여기서 L/M과 M/W의 의미는 least와 middle, middle과 worst인 환자들(샘플)이 거의 같은 비율로 같은 클러스터로 클러스터링 되어 있음을 말한다.

표 1. 세 개의 클러스터링 알고리즘의 클러스터 결과와 제안하는 방법의 결과 비교

Microarray data set for CFS		Clustering results	
Method	Cluster set #	Algorithms	True clusters
KM	Cluster1	M	W
	Cluster2	M	W
	Cluster3	L	W
	Cluster4	L	L
HC	Cluster1	L	M
	Cluster2	L	L
	Cluster3	M	W
	Cluster4	L	W
PCA	Cluster1	M	M
	Cluster2	L	W
	Cluster3	M	W
	Cluster4	M	M
HCE	Cluster1	L	L/M
	Cluster2	M	M
	Cluster3	M	M/W
	Cluster4	L	L

제안하는 알고리즘에서는 클러스터링 결과가 4개일 때가 적합도가 가장 높게 나와 비교하였다. 그러나 나머지 3개와 5개일 때의 클러스터링 결과와 본 논문에서 제안하는 알고리즘의 클러스터링 결과의 비교에 있어서도 적합도가 가장 높지는 않으나 더 나은 실험 결과의 정확도를 보인다.

5. 결론 및 향후 연구

본 논문에서는 바이오데이터와 클러스터링 알고리즘들의 특성들을 고려하여 다른 개수의 클러스터 결과들을 결합하여 최적의 결과 클러스터를 보이고자 하였다. 그리고 최종적으로 최적의 클러스터를 만들기 위해서 GA의 교차 연산 방법을 새롭게 디자인하여 적용하는 HCE 접근법을 제안하였다. 그 결과 제안하는 방법이 여러 가지 클러스터링 알고리즘들을 결합하여 생성된 결과 클러스터들을 이해하고 적용하기에 유용함을 발견할 수 있었다. 그리고 실제로 마이크로어레이 데이터를 가지고 실험하여 이 방법이 효과적으로 해답을 찾을 수 있고, 많이 중복되어지는 원소들을 가지는 클러스터들을 한쌍으로 교차 연산을 수행함으로서 클러스터의 타당성을 더욱 높일 수 있었다.

제안하는 방법은 클러스터링 개수를 처음 적용 단계에서 맞추거나 실험전에 데이터 셋의 원소들을 전처리 작업을 통하여 제거하지 않아도 되기 때문에 다른 클러스터링 알고리즘의 클러스터링 결과들보다 더 믿을 수 있는 결과를 이끌어 낼 수 있다. 또한 여러 가지 클러스터링 알고리즘의 결과를 결합한다는 것은 바이오데이터의 특성을 고려했을 때 혹은 클러스터링 알고리즘의 결과 클러스터를 해석하는 데 있어서 클러스터링 알고리즘의 문제들을 극복할 수 있을 것으로 보인다.

6. 참고 문헌

[1] Alexander, T., Behrouz, M-B., Anil, K. J., William, F. P.: Adaptive clustering ensembles. Proceedings of the International conference on Pattern Recognition, 1 (2004) 272-275

[2] Whistler, T., Unger, E. R., Nisenbaum, R., Vernon, S. D.: Integration of gene expression, clinical, and epidemiologic data to characterize Chronic Fatigue Syndrome. Journal of Translational Medicine, 1 (2003)

[3] Jouve, P. E., Nicoloyannis, N.: A new method for combining partitions, applications for distributed clustering. Proceedings ofthe International Workshop on Parallel and Distributed Machine Learning and Data Mining, (2003)

[4] Xiaohua, H.: Integration of cluster ensemble and text summarization for gene. Proceedings of the 4th IEEE Symposium on Bioinformatics and Bioengineering, (2004) 251-258

[5] Banerjee, A., Krumpelman, C., Basu, S., Mooney, R., Ghosh, J.: Model-based overlapping clustering. Proceedings of the International Conference on Knowledge Discovery and Data Mining, (2005) 532-537