유전자 알고리즘에 기반한 k-medoid 클러스터링 알고리즘에서의 최적의 k-탐색과 적용

안선영° · 윤혜성 · 이상호
이화여자대학교 컴퓨터학과
lovesy@ewhain.net · comet@ewhain.net · shlee@ewha.ac.kr

Optimal k-search and Its Application in k-medoid Clustering Algorithm based on Genetic Algorithm

Sun-Young Ahn° · Hye-Sung Yoon · Sang-Ho Lee
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

k-medoid 클러스터링 알고리즘은 고정된 클러스터 수(k)를 가지고 실험하기 때문에 데이터에 대한 사전 지식이 없으면 올바른 분석이 어렵고, 클러스터 수를 변경하면서 여러 번 반복 실험하여 실험 결과에 대한 타당성을 조사해야 하기 때문에 데이터의 크기가 커질수록 시간 비용이 증가하는 단점이 생긴다. 본 논문에서는 k-medoid 클러스터링 알고리즘 분석에 있어서 가장 어려운 문제 중 하나인 적절한 클러스터 수 k를 사회 네트워크 분석 방법 중 매개중심값을 이용하여 찾는 새로운 방법을 제안하고 이를 실제 마이크로 어레이 데이터에 적용하여 유전자 알고리즘에 기반한 k-medoid 클러스터링을 수행함으로써 좀 더 정확한 클러스터링 결과를 보인다.

1. 서 론

DNA 마이크로어레이 기술의 발달과 생명공학의 발전은 방대한 양의 바이오 데이터의 생성과 양적인 증가를 가져왔고 이런 데이터를 효과적으로 관리하고 분석, 이용하기 위한 데이터 마이닝의 필요성 또한 중요한 연구 문제가 되고 있다. 클러스터링 기법은 유전자 발현 데이터 분석을 위해 가장 널리 사용되고 있는 방법으로 데이터의 특성, 분석하고자 하는 목적에 따라 많은 종류의 서로 다른 알고리즘이 존재하므로 데이터의 특성에 따른 적절한 클러스터링 알고리즘의 선택이 무엇보다 중요하다. 분할 클러스터링의 한 종류인 k-medoid 클러스터링은 다른 알고리즘들에 비해 상대적으로 노이즈나 이상치에 영향을 덜 받기 때문에 노이즈가 많은 바이오 데이터의 특성에 잘 맞는 알고리즘으로 비교적 좋은 클러스터링 결과를 제공하는 장점을 가진다. 하지만 클러스터의 수를 고정시키고 실험을 해야 하기 때문에 데이터에 대한 사전 지식이 없으면 올바른 분석의 어려움이 있고, 클러스터의 수를 변경하면서 여러 번 반복 실험을 해야 하기 때문에 데이터의 크기가 클수록 분석을 위한 시간 비용이 증가하는 등의 단점이 있다. 본 논문에서는 k-medoid 클러스터링 수행 시 클러스터의 수를 매번 변경하여 반복 실험을 해야 하는 단점을 보완하여 데이터를 가장 이상적으로 클러스터링 할 수 있는 클러스터 수 k를 찾는 방법을 제안하고자 한다. 또한 제안된 방법을 통해 찾은 k를 가지고 유전자 알고리즘에 기반한 k-medoid 클러스터링을[1] 수행하여 데이터의 크기가 커질수록 증가하는 시간 비용을 효율적으로 감소시키고 정확하고 생물학적으로 의미 있는 클러스터링 결과를 찾고자 한다.

논문의 구성은 다음과 같다. 2장에서는 본 논문의 관련 연구로 클러스터링 알고리즘과 마이크로어레이 데이터 분석을 위한 사회 네트워크(social network) 방법 그리고 유전자알고리즘에 대해 설명하고, 3장에서는 본 논문에서 제안하는 클러스터 수 k를 찾는 방법과 유전자 알고리즘 기반의 k-medoid 알고리즘에 대해 설명한다. 그리고 3장에서 제안한 방법을 적용한 실험 결과를 4장에서 보이고, 마지막으로 5장에서는 실험적 결과와 결론 그리고 앞으로의 계획에 대해 설명한다.

2. 관련 연구

본 장에서는 관련 연구로서 클러스터링 알고리즘, 마이크로어레이 데이터 분석을 위한 사회 네트워크방법, 유전자알고리즘에 대하여 설명한다.

2.1 클러스터링 알고리즘

클러스터링 알고리즘은 유전자 발현 데이터 분석을 위해 가장 널리 사용되고 있는 방법으로 기능을 알지 못하는 유전자의 기능 분석과 유전자 상호 관련성 분석 등에 중요한 의의를 가진다. 클러스터링 기법은 크게 계층적 클러스터링 방법과 분할 클러스터링 방법으로 나뉜다. 분할 클러스터링 방법이란 임의의 클러스터 내부의 객체들이 다른 클러스터 내부의 객체들 보다 유사한 발현 패턴을 보이도록 객체들을 분할하는 방법으로 클러스터의 무게 중심점을 대표 값으로 분할해 나가는 k-means 방법과 클러스터 내에 중심과 가장 가까운 객체를 대표점으로 하는 k-medoid 방법이 있다[2]. 그러나 k-means 알고리즘은 매우 간단하고 효율적이지만 클러스터 중심을 기반으로 하는 알고리즘이기 때문에 노이즈나 이상치에 영향을 많이 받는다. 이는 노이즈에 특히 민감한 바이오 데이터에 적용할 때에 잘못된 클러스터링 결과를 가져올 수 있다. 따라서 본 논문에서는 k-means 보다는 상대적으로 이런 극단적인 값에 강하고 비교적 좋은 클러스터링 결과를 제공하는 k-medoid 방법을 적용하였다.

2.2 마이크로어레이 데이터 분석을 위한 사회 네트워크 분석

과거 생명 공학 분야의 연구 방식은 가설을 수립하고 접근하는 방식으로 실험자가 어떤 가정을 하고 그것을 증명하는 매우 제한적인 방식으로 진행되었다. 그러나 실험 기술이 발달하고 대량의 데이터 생산이 가능해 짐에 따라서 대량의 데이터 속에서 유용한 지식을 찾기 위해 다수의 유용한 가설들을 생성해 내는 데이터 위주의 연구 방식으로 변화하고 있다. 유전자가 발현되기 위해서는 그 유전자 자체만이 아닌 다른 여러 가지 환경적인 요소가 많이 작용을 하게 되는데 그 요소들 중 하나가 다른 유전자들과의 관계이

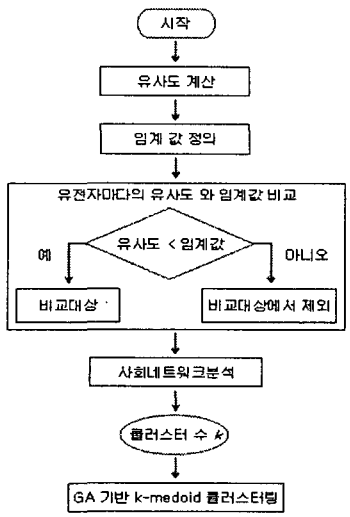
다. 즉, 수많은 유전자들과 다른 물질들 사이의 복잡한 상호작용에 의한 유전자들의 기능을 개개의 유전자의 활동으로 보는 것이 아니라 하나하나의 유전자의 활동이 다른 유전자에게 영향을 미치는 것으로 보는 전반적인 유전체 관점에서의 분석이 필요하다. 따라서 본 논문에서는 이러한 관점으로 마이크로어레이 데이터를 분석을 위해 사회 네트워크 방법을[3] 적용하였다. 노드들 하나하나를 분석하고자 하는 유전자로 보고 그들 사이의 에지를 유전자들 사이의 관계나 사건으로 보는 것으로 가장 활동적인 유전자는 많은 다른 유전자들과도 연결 관계를 갖는다는 것을 의미한다고 가정한다. 또한 그 중에서도 매개 중심(betweenness centrality) 방법을 적용하였는데 이는 다른 노드들과의 사이에서 'bridge' 역할인 노드를 찾는 방법으로 전체 데이터 네트워크가 이 노드에 의해서 밀접하게 연관되어 있다는 것을 말한다.

2.3 유전자 알고리즘

유전자 알고리즘은[4] 자연세계의 진화 과정을 컴퓨터상에서 시뮬레이션 함으로써 복잡한 실세계의 문제를 해결하고자 하는 계산 모델인 진화 알고리즘의 한 분야이다. 유전자 알고리즘은 최종 세대의 결과 값이 곧 최적 해를 나타내기 때문에 결과에 대한 별도의 해석이 필요 없고 결과 값의 적용이 쉬우며 다양한 데이터 형태에 적용가능하다는 장점을 가진다. 또한 단독으로도 사용 가능 하지만 다른 여러 알고리즘의 최적화 문제를 위해 하나 이상의 다른 알고리즘들과 혼합되어 사용되기도 한다. 사이즈가 큰 데이터 셋에서의 k-medoid 클러스터링 문제는 NP-hard 문제이다. 이런 NP-hard 문제의 최적화를 다루기 위해 유전자 알고리즘을 함께 사용하는 방법이 제안되었고, 이때 고려되야 할 계산 비용을 감소시키기 위해서 본 논문에서는 유전자 알고리즘의 여러 연산자를 적용하여 효율적이고 정확한 클러스터링 결과를 보인다.

3. 알고리즘

본 장에서는 논문에서 제안하는 클러스터 수 k를 구하는 방법과 유전자 알고리즘 기반의 k-medoid 클러스터링 알고리즘에 대하여 설명한다. 전체적인 순서는 그림 1과 같다.



[그림 1] k값의 탐색과 유전자 알고리즘 기반의 k-medoid 클러스터링

3.1 클러스터 수 k의 탐색과 적용

■ 임계값 정의

유전자들 간의 유사성이 높을수록 하나의 클러스터에 속할 확률이 높으며 낮은 유사성을 가진 유전자 일수록 서로 다른 클러스터에 속할 확률이 높다. 따라서 유전자들이 같은 클러스터에 속하는지 다른 클러스터에 속하는지를 예측하기 위해 판단의 기준이 될 임계값을 정의한다. 하나의 유전자와 다른 유전자들 간의 유사성을 계산하기 위해 유클리디안 거리 공식을 이용하여 유전자들 간의 유사성을 계산하고 거리 값의 평균을 구한다. 계산된 평균값을 이용하여 임계값을 정의하게 되는데, 이때 너무 작은 평균값과 너무 큰 평균값에 의해 임계값이 한쪽으로 치우치는 것을 막기 위해 중간 값들의 평균을 임계값으로 한다. 이렇게 구해진 임계값과 유전자들 간의 유사성과 비교하여 임계값보다 작은 유사성을 가지는 유전자는 한 클러스터에 속할 확률이 크므로 비교 대상에 넣고, 큰 유사성을 가지는 유전자들은 다른 클러스터일 확률이 높으므로 비교 대상에서 제외시키는 방법으로 비교 대상일 될 유전자를 일차적으로 가려낸다.

■ 사회 네트워크 방법을 통한 클러스터 수 예측

바이오데이터는 데이터들의 상호 작용이 의존적인 경우가 많기 때문에 어떤 상호 작용의 매개가 되는 유전자 하나를 제거하게 되면 처음과 다른 결과를 보일 수 있다. 이런 바이오 데이터의 특징을 이용하여 임계값을 가지고 추출된 유전자들에 대해 사회 네트워크 분석법 가운데 매개 중심 방법을 적용한다.

(1) 우선 임계값을 기준으로 비교 대상으로 추출된 유전자에 대하여 사회 네트워크의 매개 중심 방법을 적용한다. 실험에서는 유전자마다 매개 중심 값을 계산하여 가장 높은 값을 가지는 유전자를 찾고 이 유전자를 추출한다.

(2) 앞의 (1)에서 가장 높은 매개 중심 값을 가지는 유전자들을 제외한 나머지 유전자들을 가지고 다시 매개 중심 값을 계산한다. 그리고 가장 높은 매개 중심 값을 가지는 유전자를 또 추출한다.

(3) 이러한 방법을 계속적으로 반복하면 매개 중심 값이 급격하게 줄어드는 시점을 발견할 수가 있는데 이를 특정 유전자가 더 이상 유전자들 간의 매개 중심 역할을 하지 못한다고 판단하여 그 시점까지 추출된 유전자들을 가지고 클러스터 수 k를 예측할 수 있다.

3.2 유전자 알고리즘 기반의 k-medoid 클러스터링

유전자 알고리즘을 기반으로 하는 k-medoid 클러스터링 방법(GA-KMC)은 다음과 같이 구성한다.

1. P개의 개체들로 이루어진 개체군을 랜덤하게 생성한다.
2. 초기개체군에서 각 개체들의 유클리디안 거리 계산으로 fitness function 값을 정한다.
3. 종료 조건을 만족할 때까지 다음의 과정 반복한다.
   (1) 토너먼트 선택법으로 재생산에 참여할 P/2개의 부모 개체 쌍(pair)들을 구한다.
   (2) Mix subset recombination 교배와 flip mutation을 수행하여 새로운 자식 개체를 생성한다.
   (3) k-medoid 클러스터링을 수행한다.
   (4) 자식 개체들의 fitness 값을 구한다.
   (5) 이전 세대의 최고 멤버와 자식들 중 최고 멤버로부터 크기 P의 새로운 개체군을 생성한다.
4. 최고의 fitness를 가지는 최종 개체군 멤버를 구한다.

4. 실험 결과

본 논문에서는 매개 중심 값 계산을 위해 UNISET 6.0 프로그램을 이용하였으며, 유전자 알고리즘 기반의 k-medoid 클

러스터링 구현은 java 1.5 SDK를 사용하여 windows server 2003 환경에서 실행하였다. 실험에 사용된 데이터 셋은 Serum 데이터[5]와 Subyeast 데이터[6]이다. 먼저 클러스터링 수가 명확하게 알려진 앞의 두개의 데이터와 본 논문에서 제안하는 방법을 같은 데이터 셋에 적용하여 얼마나 정확하게 클러스터 수 k를 찾아내는지에 대하여 비교 하였다. 아래 표 1에서 보는 것처럼 매개 중심 값이 급격히 떨어지는 시점에서 예측된 클러스터 수 k가 기존의 논문에서 알려진 k와 정확하게 10개와 30개로 일치하는 것을 발견 할 수 있었다.

[표 1] 알려진 k와 예측된 클러스터 수 k 비교

데이터	클러스터 수 k	
	알려진 k	매개중심값으로 계산된 k
serum 데이터	10	10
subyeast 데이터	30	30

다음으로 우리는 기존에 제안된 유전자 알고리즘 기반의 k-medoid 알고리즘들과의 성능을 비교하였다. 먼저 표 2는 세 개의 유전자 알고리즘 기반 k-medoid 클러스터링 알고리즘인 Genetic Clustering Algorithm(GCA)[7], Random Assorting Recombination Clustering Algorithm(RARw-CA)[8], Hybrid K-medoid Algorithm(HKA)과 논문에서 사용한 유전자 알고리즘 연산자에 대해 설명한 것이다. 연산자의 선택엔 크게 달라진 점이 없지만 HKA와 본 논문에서 제안하는 GA-KMC는 k-medoid 클러스터링 수행 시 해가 지역적 최적화에 빠지는 것을 방지하기 위해 인접 이웃(nearest neighbor)의 개념을 함께 사용하였다. 논문에서 제안하는 GA-KMC방법은 기존 방법에서 연산자와 연산 확률만을 변화시켜 실험을 하였다[9].

[표 2] 유전자 알고리즘에 기반한 k-medoid 알고리즘들의 유전자 알고리즘 연산자

연산자	알고리즘			
	GCA	RARw-CA	HKA	**GA-KMC**
선택 연산자	룰렛	룰렛	2-fold 토너먼트	**토너먼트**
교배연산자	Mix Subset Recombination	Random Assorted Recombination	Mix Subset Recombination	**Mix Subset Recombination**
돌연변이	Flip bit Mutation	Flip bit Mutation	Flip Mutation	**Flip Mutation**

표 3은 클러스터링 알고리즘 적용 후 각 알고리즘의 클러스터링 성능과 결과를 비교, 분석한 표이다. 클러스터링 결과에 대한 비교는 클러스터링 후의 유클리디안 거리 값의 합이 얼마나 작은가에 따라 비교하였고, 수행시간은 본 논문에서는 미리 k 값을 계산하고 클러스터링을 하기 때문에 모든 기존 알고리즘 또한 k값을 미리 알고 수행되었다는 가정 하에 계산되었다. 표에서 나타나는 것처럼 GA-KMC를 사용하면 기존 알고리즘에 비해 수행시간이 많이 향상되고 유전자들 간의 거리 값의 합 역시 줄어들어 기존 방법에 비해 조밀한 클러스터링 결과를 보인다는 것을 알 수 있었다.

[표 3] 기존 알고리즘과의 클러스터링 결과 분석

알고리즘	Serum 데이터	
	수행시간(초)	거리값
GCA	69.5	872,313
RARw-CA	66.3	876,632
KHA	17.2	861,076
GA-KMC	12	860,543

사회 네트워크의 매개중심방법을 이용한 클러스터 수의 예측은 이미 알려진 클러스터 수와 비교적 정확하게 일치하는 것을 관찰 할 수 있었고, 유전자 알고리즘 기반의 k-medoid 클러스터링 역시 유전자 알고리즘 연산자와 연산이 일어난 확률을 변화시킴으로써 좀 더 나은 클러스터링 결과 값과 시간 비용이 줄어드는 걸 확인 할 수가 있었다.

5. 결론 및 향후 연구 과제

본 논문에서는 클러스터링 알고리즘을 수행하는 데 있어서 가장 어려운 문제였던 클러스터 수 k를 결정하는데 기존의 반복적인 실험과 사용자의 경험 위주로 했던 방법들의 효율성을 높이고자 사회 네트워크의 매개중심값을 적용하였다. 유전자를 개개의 노드로 보고 이들 사이의 연결정도를 파악하여 유전자와 유전자를 연결하는데 있어 가장 많은 연결 정도를 보이는 노드가 하나의 클러스터에서 또한 중심이 될 수 있다는 가정을 하였고 본 논문에서는 이 가정이 맞다는 것을 증명하였다. 또한 제안하는 방법을 통해 얻은 k값을 가지고 실제 클러스터링을 한 결과 유전자 알고리즘의 연산자와 연산 확률을 적절히 조절함으로써 기존 방법들보다 더 좋은 클러스터링 결과를 얻었고, 계산 비용도 기존 방법에 비해 줄어든 것을 확인할 수가 있었다. 그러나 계속적으로 데이터의 크기와 양이 커지고 있는 요즘 데이터의 크기가 커질수록 이런 방법은 매번 매개 중심값을 계산해 줘야 하기 때문에 계산 비용적인 면에서 효율적이지 못 할 수도 있다는 문제를 가지고 있다. 따라서 향후 계획은 대용량 데이터에 적용할 때에 이런 계산 비용을 줄 일 수 있는 방법을 연구하고 여러 데이터에 적용해 봄으로써 좀 더 정확한 클러스터링 결과를 얻을 수 있는 연구를 계속 진행하고자 한다.

6. 참고 문헌

[1] Weiguo Sheng, Xiaohui Liu, A Hybrid Algorithm for k-medoid Clustering of Large Data Sets. *IEEE Congress on Evolutionary Computation(CEC-2004)*, pp. 77-82, 2004

[2] 박우창, 데이터마이닝(개념 및 기법), *자유아카데미*, 2003

[3] 김용학, 사회연결망 분석, *박영사*, 2003

[4] http://alife.cau.ac.kr/info/EAs/GAs.html

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci*, vol. 95, no. 25, pp. 14863-14868, 1998

[6] S. Tavazoie, D. Hughes, J. M.J. Cambell, R.J. Cho, and G.M. Church: Systematic determination of genetic network architecture, *Nature Genetics*, 22, pp.281-285, 1999

[7] C.B. Lucsdiud, A.D Dane and G. Kateman: On k-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasibility and comparison. *Analytical Chimica Acta*, 282, pp.647-669, 1993

[8] V.Estivill-Castro and A.T. Murray: Spatial Clustering for Data Mining with Genetic Algorithms, *International ICSC Symposium on Engineering of Intelligent Systems*, 1998

[9] 공성곤, 유전자 알고리즘 입문, *진영사*, 1999