Non-Exhaustive, Overlapping K-means를 이용한 생물학 데이터 분석

이정주[1] 황지영[2‡]
[1]성균관대학교 생명과학과
[2]성균관대학교 컴퓨터공학과
{aih9599, jjwhang}@skku.edu

Biological data analysis using Non-Exhaustive, Overlapping K-means

Jungju Lee[1] Joyce Jiyoung Whang[2‡]
[1]Department of Biological Science, Sungkyunkwan University
[2]Department of Computer Science and Engineering, Sungkyunkwan University

요 약

K-means 알고리즘은 대표적인 클러스터링 알고리즘이다. 하나의 데이터가 반드시 단 하나의 클러스터에 속하게 되는 K-means를 변형한, overlapping cluster와 클러스터에 속하지 않는 outlier를 찾아내는 NEO-K-Means를 인간 유전자 발현 microarray data에 적용하여 분석하였다. 또한 NEO-K-Means의 변형을 통해 수행 시간을 단축시켜 대용량 데이터의 클러스터링이 용이하도록 만들었다.

1. 서 론

K-means clustering은 데이터 마이닝(data mining) 에서 대표적이며 일반적으로 사용되는 클러스터링 알고리즘이다. 각 데이터를 클러스터(cluster)로 분할하고, 클러스터 내의 데이터와 클러스터 중심과의 평균 거리를 최소화하는 방법이다. 여러 클러스터링 방법론들이 개발되어 오고 있으나 K-means는 간단하고 효과적인 기법으로서 여전히 널리 이용되고 있다[1].

기존의 K-means 알고리즘에서 하나의 데이터 포인트는 반드시 단 하나의 클러스터에 속했다. 이는 각 그룹들 간 경계가 명확하고 outlier가 존재하지 않는 data set의 경우에는 효과적이나, 그룹 간 경계가 불확실하며 outlier를 가진 data set에서는 좋은 분할 결과를 얻을 수 없다. 데이터의 cluster의 overlap조절과 outlier detect를 동시에 실행하는 알고리즘은 많지 않다[2].

Non-exhaustive, Overlapping K-Means algorithm (NEO-K-Means)은 overlapping cluster와 outlier를 함께 고려하여 데이터를 군집화 하는 알고리즘이다. Real world data의 경우 클러스터가 명확히 구분되어있지 않거나 noise가 존재하는 경우가 많기 때문에 real world data에 적용했을 경우 높은 수준의 결과물을 얻을 수 있을 것으로 예상된다[3].

생명공학의 DNA microarray 기술은 한 번에 수천 개의 유전자 발현 정보를 획득할 수 있다. 대량의 생물정보는 데이터 마이닝 기법을 이용하여 처리될 수 있으며 이를 통해 질병과 유전자의 기능 등을 분석하고 예측할 수 있다. 하나의 유전자는 다양한 functional family에 속할 수 있기 때문에 여러 클러스터에 속할 수 있다. 따라서 기존의 K-means를 적용하기에는 한계가 있다. 본 논문에서는 유전자 데이터를 클러스터링 하는 방법으로 NEO-K-Means를 이용하며, 대용량 데이터를 처리하는 데 있어 소요 시간을 단축시키기 위한 방법을 제시했다.

2. 관련연구

Microarray data는 유전자가 어떠한 조건, 또는 특정 조직 내에서 얼마나 발현되는지를 수치화한 2차원 데이터이다. Microarray는 칩의 표면에 유전자 조각들이 부착된 것이며, 형광 표지된 cDNA 검체와 상보적으로 결합하고, 그 정도에 따른 형광강도를 측정해 수치를 얻게 된다[4].

유전자 발현 패턴을 클러스터링 알고리즘을 이용해 분석하는 연구는 다양한 방법으로 꾸준히 진행되어 왔다. Eisen등의 연구는 유전자 분석에 클러스터링을 적용한 초창기의 것이다. 통계학에서 자주 사용되는 계층적 클러스터링 방법을 이용하여 인간과 효모에서 유사 발현 패턴을 보이는 유전자

‡ 교신저자 (Corresponding author)
이 논문은 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2016R1D1A1B03934766, NRF-2010-0020210). 또한, 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음(2015-0-00914).

그룹을 분석했다[5]. Hartuv는 그래프 이론에 기초해 가중치가 없는 유사도 그래프를 입력 받아 정점을 유전자로, 정점 사이의 연결선은 유전자의 유사성으로 나타냈고, 유사성이 일정 수준을 넘을 경우에 그래프가 형성되도록 구성한 알고리즘을 제안했다[6].

3. K-means와 NEO-K-Means

NEO-K-Means는 각 클러스터의 중심점과 클러스터 내 데이터 포인트 간 유클리드 거리를 최소화한다는 기본적인 면은 K-Means와 동일하다. 기존의 k-means에 cluster size 합을 α로, cluster를 갖지 않은 데이터의 수를 β값으로 표현하는 수식들을 추가하며, α와 β값은 임의로 혹은 계산을 통해 부여하며 이를 통해 겹치는 정도와 outlier 수를 정한다.

3.1 Standard K-means

n개의 데이터 X = {x_1, x_2, x_3,...,x_n}이 존재할 때 이 데이터 포인트를 k개의 클러스터 C_1, C_2, ... C_k 에 분류한다. 모든 데이터 포인트는 하나의 클러스터에 속하고, 서로 다른 클러스터에는 동일한 데이터 포인트가 존재하지 않는다. 데이터 포인트들을 가장 가까운 클러스터에 배치하고 클러스터 중심을 재계산하는 것을 반복하여 objective function을 최소화할 수 있다. K-means objective function은 다음과 같다.

$$\min_{c_j} \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - x_j \|, where\ m_j = \frac{\sum_{x_i \in C_j} x_i}{|C_j|}$$

3.2 NEO-K-Means Objective

NEO-K-Means의 핵심 아이디어는 데이터 배치 행렬인 U의 도입이다. U의 크기는 n×k 이며 데이터 x_i가 cluster j에 속할 때 u_{ij} = 1인 행렬이다. 기존 K-means결과물을 U에 적용하면 배열 내에 값이 1인 원소의 개수는 n이고, 한 행 내에서 값이 1인 원소는 하나만 존재한다. NEO-K-Means에서는 한 행 내에 1인 원소가 여러 개 있을 수 있고, 전부 0일 수도 있다. 배열 내의 모든 1의 개수는 n+αn개 이며 α는 겹쳐진 클러스터의 데이터 포인트 수를 나타낸다. 또한 어느 클러스터에도 소속되지 않아 원소가 모두 0인 행의 개수는 데이터 포인트 수를 βn을 이용해 제한한다. Non-Exhaustive Overlapping K-Means의 objective function은 아래 수식과 같다.

$$\min_{U} \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \| x_i - m_j \|^2, where\ m_j = \frac{\sum_{i=1}^{n} u_{ij} x_i}{\sum_{i=1}^{n} u_{ij}}$$

$$s.t. trace(U^T U) = (1+\alpha)n, \sum_{i=1}^{n} II(U1_i = 0 \le \beta n$$

3.3 NEO-K-Means Algorithm

NEO-K-Means의 objective function을 최적화하기 위해 알고리즘이 개발되었고, 이러한 순서를 거친다.

1)데이터를 읽어 온 후 standard k-means를 수행해 각 데이터 포인트마다 하나의 클러스터를 배당하고, 그 결과를 통해 α와 β값을 estimate 한다. βn은 outlier 수, αn은 overlapped data point의 개수이다.

2)각 클러스터의 중심을 계산하고 데이터 포인트와 각 중심과의 거리를 계산한다. 각 포인트에서 가장 가까운 클러스터를 찾은 후 n-βn만큼의 데이터를 배당한다.

3)이후 남은 포인트들 중 각 클러스터 중심과 가장 가까운 데이터 αn+βn개를 추가적으로 배당한다. 이 과정에서 overlapping cluster가 발견되고, outlier가 βn보다 줄어들 수 있다. objective function이 더 이상 감소하지 않을 때까지 2와 3 과정을 반복한다.

3.4 속도 향상

이러한 과정은 standard k-means에 추가적인 계산과 메모리를 필요로 하게 되므로 시간과 자원이 더 많이 필요하게 된다. 따라서 대량의 데이터를 처리하기 위해 필요로 하는 자원과 소요시간을 줄였다.

먼저 matrix U의 경우 n×k의 크기로 k가 커질수록 sparse해진다. 따라서 (n+αn)×2 크기의 matrix를 이용해 기존의 U에서 0이 아닌 원소들만 저장함으로써 기억공간의 사용을 줄였다.

Objective function이 최소값으로 수렴하는 중 그 감소 폭은 점점 줄어드는 것을 발견할 수 있었는데(그림1), 이는 iteration이 종료되기 일정 시점 전부터는 data point의 소속 클러스터가 크게 변화하지 않았음을 의미한다. 클러스터 중심과 가까워 초기에 클러스터에 배당되는 n-βn개의 data point들의 소속 변화를 관찰한 결과 일정 시점에서부터는 동일하게 소속됨을 볼 수 있었다. 그러므로 이러한 시점이 발견된 경우 n-βn개 배당은 추가적 계산 없이 이전의 iteration과 동일하도록 만들었다.

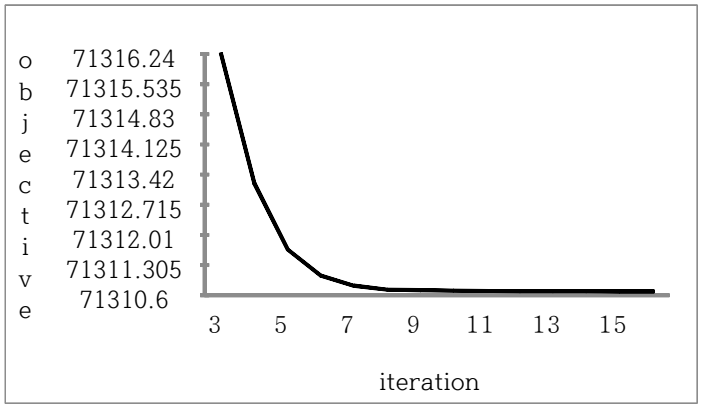

그림 1 k=3, α = 0.1 β = 0.005 일 때 iteration 3부터의 objective

또한 αn+βn개를 배당하는 과정에서 클러스터 중심과 데이터 포인트 사이 거리의 최소값을 n×k×(αn+βn)번 찾는 iteration은 k = 100의 경우 약 7초로 각 과정 중 가장 많은 시간이 소모되었다. 따라서 이 부분의 iteration을 줄이기 위해 배당되지 않은 데이터들의 거리 값들을 추출한 후 quick sort를 수행해 최소값을 가진 데이터αn+βn 개를 찾아내는 방법으로 대체해 실행 속도를 향상시켰다.

4. 결과

4.1 이용한 데이터

미국 보건성 산하의 국립의학도서관 운영 분야 중 The National Center for Biotechnology Information (NCBI)의 유전체 서열 데이터베이스에서 제공하는 Large-scale analysis of the human transcriptome 데이터를 이용했다. 단백질을 만드는 22283개 인간 유전자의 156개 조직에서의 발현 정도를 기록한 데이터이다.
(https://www.ncbi.nlm.nih.gov/geo/)

4.2 속도 향상

먼저 n-βn개의 데이터 배당 생략의 경우, k = 10일 때 마지막 3번의 배당이 생략되었고, NEO-K-Means 수행 시간은 202.603초에서 150.295초로 감소되었다. 하지만 속도 향상 폭이 크지 않았고, k-means를 수행할 때 초기 중심지를 랜덤하게 설정했기 때문에 k-means의 결과물이 달라짐에 따라 향상 정도도 차이가 있었다. 따라서 αn+βn데이터 할당의 반복을 줄이는 방법과 병행하여 표1과 같은 결과를 얻었다.

k	2	4	8	16
기존	8.159	15.867	104.316	146.375
변환	2.539	6.558	20.068	62.986
k	32	64	128	256
기존	399.306	926.685	1389.662	2887.606
변환	110.261	275.106	703.557	773.401

표 1 k에 따른 소모 시간(단위 : 초). α = 0.1, β = 0.005

4.3 결과 데이터의 분석

k = 256의 경우 각 클러스터의 크기는 그림 2와 같은 분포를 보인다. 450보다 큰 클러스터는 없으며, 주로 50개에서 100개 구간에 분포한다.

유전자에서 생성된 것들의 기능을 나타내는 gene ontology function과 클러스터링 결과를 비교해보았다. 많은 유전자에서 나타난 40개의 function을 랜덤하게 선정하고, 각각의 클러스터에 (해당 기능을 가진 유전자의 수)/(클러스터의 유전자 개수)×100으로 function의 비율(FR)을 표현했다. 서로 겹쳐져 있는 7, 117, 121, 176번 클러스터, 61, 67, 86, 95클러스터, 그리고 86, 117, 160, 163, 175, 176 클러스터 각각의 FR의 표준편차는 평균적으로 0.77, 0,76, 0.50이었다. 반면, 다른 클러스터와 겹쳐지지 않은 14, 20, 110, 208클러스터의 경우 FR의 표준편차는 이전의 그룹들에 비해 높았으며 평균은 1.97이었다. 즉 overlapped cluster에는 유전자의 기능에 대해 유사한 분포를 보일 것으로 추정할 수 있다.

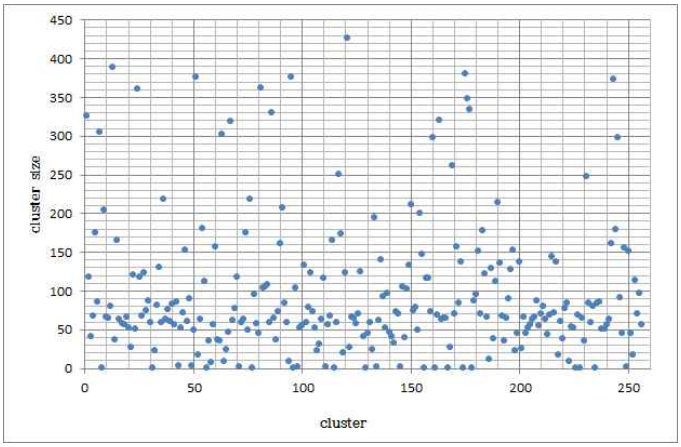

그림 2 k = 256일 때 각 클러스터의 사이즈

5.결론 및 향후 과제

본 논문에서는 기존 알고리즘의 속도 측면의 단점을 보완하기 위한 방법을 제시하였으며 생물학 데이터에 적용한 후 분석해보았다. 클러스터링 결과의 분석을 통해 클러스터와 기능 간의 연관관계를 밝히고, 나아가 기능이 밝혀지지 않은 유전자에 대해 규명할 수 있을 것이다.

참고문헌

[1] 이신원, “K-Means 클러스터링에서 초기 중심 선정 방법 비교”, 인터넷정보학회논문지 13(6) p.1-8, 2012.

[2] Guojun Gan et al. “k-means clustering with outlier removal“, Pattern Recognition Letters 90, p.8-14, 2017.

[3] Joyce Jiyong Whang et al. “Non-exhaustive, overlapping k -means”, SIAM International Conference on Data Mining (SDM), 2015.

[4] 김선주, “Microarray를 이용한 유전자 발현 연구” 대한임상미생물학회지 4권 2호, p.82-86, 2001.

[5] Michael B. Eisen et al. "Cluster analysis and display of genome-wide expression patterns", proc. Natl. Acad. Sci. USA, Vol. 95, p. 14863-14868, 1998.

[6] Erez Hartuv et al. "An algorithm for clustering cDMA fingerprints", Genomics 66, p.249-256, 2000.