유전자 발현 데이터를 이용한 암의 클래스 예측을 위한 퍼지 클러스터링 알고리즘

원홍희° 유시호 조성배
연세대학교 컴퓨터과학과
{cool°, bonanza, sbcho}@sclab.yonsei.ac.kr

Fuzzy Clustering Algorithm to Predict Cancer Class Using Gene Expression Data

Hong-Hee Won° Si-Ho Yoo Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

암의 치료법은 같은 종류의 암이라 해도 그 하부 클래스에 따라 매우 다르기 때문에 암의 클래스를 예측하는 것은 그 정확한 치료를 위하여 매우 중요하다. 유전자 발현 데이터를 이용한 암의 분류에 있어 기존의 연구들은 각 데이터를 하나의 클러스터에 소속시키는 하드 분할(hard partition)에 의한 분할 방식을 사용하는 하드 클러스터링을 사용하였다. 하지만 일반적으로 유전자 발현 암 데이터와 같은 실세계의 데이터는 쉽게 나뉘어지기 힘들거나 클러스터 간의 경계가 분명하지 않기 때문에 하드 클러스터링 기법은 주어진 데이터의 성질을 손실시킬 수 있는데 반해, 퍼지 클러스터링 기법은 각 데이터가 소속 정도에 따라 여러 개의 클러스터에 속할 수 있도록 분할하기 때문에 이러한 손실을 최소화할 수 있다. 따라서 본 논문에서는 퍼지 클러스터링의 대표적인 방법인 fuzzy c-means 클러스터링을 적용하여 암의 클래스를 예측하고, 다양한 하드 클러스터링 방법과 비교함으로써 퍼지 클러스터링의 성능을 검증하였다.

1. 서 론

같은 종류의 암이라 해도 특성상 서로 다른 여러 개의 하부 클래스로 나뉠 수 있으며 그 치료법도 하부 클래스에 따라 매우 다르기 때문에 암의 클래스를 예측하는 것은 암의 정확한 치료를 위하여 매우 중요하다. 하지만, 조직병리학 등의 방법에 의존하는 임상학적 암 분류는 종종 불완전하여 오진할 수 있는 가능성이 있다. 유전자 발현 정보에 근거한 분자 수준의 암 분류는 정확하고 객관적이며 체계적인 암의 분류를 위한 방법론을 제시해준다. 하지만 유전자 발현 데이터는 일반적으로 매우 많은 양의 유전자 정보를 포함하며 모든 유전자가 암과 관련이 있는 것은 아니므로 암과 관련이 있는 중요한 유전자만을 추출하여 이를 기준으로 암을 분류하는 것이 바람직하다.

클러스터링은 주어진 전체 데이터 집합을 유사한 성질을 갖는 몇 개의 클러스터로 분할하는 것이며, 유전자 발현 데이터와 같은 대량의 데이터를 분석하는 데 용이하기 때문에 많은 연구에서 사용되고 있다. 클러스터링 알고리즘은 클러스터로 분할시키는 정도에 따라 하드(hard) 클러스터링 기법과 퍼지(fuzzy) 클러스터링 기법으로 나뉠 수 있다[1]. 하드 클러스터링 기법은 각 데이터를 하나의 클러스터에 소속시키는 하드 분할(hard partition)에 의한 분할 방식을 사용한다. 일반적으로 유전자 발현 암 데이터와 같은 실세계의 데이터는 쉽게 나뉘어지기 힘들거나 클러스터 간의 경계가 분명하지 않기 때문에 하드 클러스터링 기법은 주어진 데이터의 성질을 손실할 수 있다. 퍼지 클러스터링 기법은 각 데이터가 소속 정도에 따라 여러 개의 클러스터에 속할 수 있도록 분할한다[1]. 따라서 퍼지 클러스터링 기법은 하드 클러스터링 기법에 비하여 노이즈에 강하며, 실세계의 데이터를 분석하는 데 적합하다.

유전자 발현 데이터는 많은 노이즈를 포함할 수 있기 때문에 각 샘플의 클러스터를 deterministic하게 결정짓는 기존의 클러스터링 방법 보다 퍼지 클러스터링의 성능이 우수할 것으로 예상되며 이를 체계적으로 검증하는 연구가 필요하다. 본 논문에서는 유전자 발현 데이터 분석에 많이 사용되고 있는 대표적인 하드 클러스터링 방법인 계층적 클러스터링, *hard c-means* 클러스터링, k-means 클러스터링과의 성능 비교를 통해서 퍼지 클러스터링의 성능을 검증하고자 한다. 백혈병, 림프종, SRBCT(Small round blue cell tumor) 관련 유전자 발현 데이터에 적용한 결과 퍼지 클러스터링의 성능이 가장 우수함을 확인하였다.

2. 관련 연구

Eisen의 연구는 유전자 분석에 있어 클러스터링 방법을 적용한 시초라고 여겨질 수 있다[2]. 그들은 유사한 발현 패턴을 보이는 유전자 그룹을 체계적으로 찾아내기 위하여 클러스터링 분석을 적용하였으며, 계층적 클러스터링 방법을 사용하여 효모와 사람의 유전자 데이터로부터 유사한 기능을 갖는 유전자 클러스터를 밝혔다.

계층적 클러스터링은 bottom-up 방식의 agglomerative 계층적 클러스터링 알고리즘과 top-down 방식의 divisive 계층적 클러스터링 알고리즘으로 나뉜다. 계층적 클러스터링 방법으로는 단일연결(single

linkage: SL), 완전연결(complete linkage: CL), 평균연결(average linkage: AL), 워드 기법(Ward's method) 등이 있다. 단일 연결 알고리즘은 두 클러스터 간의 가장 가까운 개체의 거리를 클러스터 간의 거리로 정의하고, 완전 연결 알고리즘은 두 클러스터 간의 가장 먼 개체의 거리를 클러스터 간의 거리로 정의하며, 평균 연결 알고리즘은 두 클러스터 내의 모든 개체 사이 거리의 평균을 클러스터 간의 거리로 정의한다. 워드 기법은 두 클러스터를 merge하였을 때, 제곱 오류의 합(Error sum of squares)이 가장 작은 두 개의 클러스터를 merge한다.

Tavazoie 등은 분할 클러스터링 방법의 하나인 k-means (KM) 클러스터링 방법을 사용하여 유전자 발현 데이터를 분석하였다[3]. 발현된 패턴에 따라 그룹화된 많은 클러스터들이 기능적으로 유사한 유전자들을 많이 포함하고 있음을 보였으며, 같은 클러스터 내의 유전자들의 서열을 분석한 결과 up-stream region에서 공통적으로 나타나는 새로운 cis-regulatory motif를 발견하였다. 클러스터의 밀집성(tightness)이 중요한 서열 motif의 존재와 연관되어 있음을 밝혔다.

분할 클러스터링은 클러스터 간에 중복이 없으며, 각 개체를 가장 가까운 클러스터에 할당하는 과정을 반복하여 가장 적합한 클러스터를 구성한다. Hard c-means (HCM) 알고리즘과 k-means (KM) 알고리즘, ISODATA 알고리즘이 분할 클러스터링의 대표적인 예이다.

3. 퍼지 클러스터링

Fuzzy c-means(FCM) 알고리즘은 Bezdek에 의해 제안된 것으로, 가장 널리 이용되는 퍼지 클러스터링 방법이다. Fuzzy c-means 알고리즘은 퍼지 이론을 적용한 목적 함수의 반복 최적화에 기반을 둔 방식으로 각 데이터가 특정 클러스터에 속하는 소속 정도(membership)를 이용하여 데이터에 대한 보다 정확한 정보를 제공한다. 주어진 데이터 집합이 $X = \{x_1, x_2,..., x_n\}$이고 퍼지 클러스터링의 중심 벡터가 $V = \{v_1, v_2,..., v_c\}$일 때, 목적함수는 각 데이터 x_j와 각 클러스터 중심 v_i와의 거리와 클러스터 소속 정도 값으로 정의된다.

$$J_m(X,U,V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (\mu_{ij})^m d^2(x_j, v_i) \quad (1)$$

여기서 u_{ij}는 x_j와 j번째 클러스터에 대한 소속 정도를 나타내며 $(c \times n)$의 소속 행렬 $U = [u_{ij}]$의 원소이다. $d^2(,)$는 유클리디안 거리(Euclidean distance)의 제곱이며, 매개변수 m은 각 데이터의 소속 정도에 대한 퍼지 값을 나타내며 1 보다 큰 값을 사용한다.

Fuzzy c-means 알고리즘의 수행절차는 다음과 같다.

1) 클러스터의 수 c와 퍼지 계수 m의 값을 정한다.
2) 다음의 조건을 만족하도록 x_j의 소속 정도인 u_{ij}를 초기화한다.

$$\sum_{i=1}^{c} \mu_{i,j} = 1, \ 1 \le j \le n \quad (2)$$

3) 각 클러스터의 중심 v_i를 계산한다. $(i=1,2,...,c)$

$$v_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m x_j}{\sum_{j=1}^{n} \mu_{ij}^m} \quad (3)$$

4) 소속 행렬 U를 계산한다.

$$\mu_{ij} = \frac{\left(\frac{1}{d^2(x_j, v_i)}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{c} \left(\frac{1}{d^2(x_j, v_k)}\right)^{\frac{1}{m-1}}} \quad (4)$$

5) 다음의 종료 조건이 만족될 때까지 3)과 4)를 반복한다. l은 반복 단계를 의미한다.

$$|\{J_m^{(l)} - J_m^{(l-1)}\}| \le \varepsilon \quad (5)$$

4. 실험 및 결과

4.1 실험 데이터

실험 데이터로 사용한 백혈병 데이터는 38개의 샘플 데이터로 구성되어 있으며, 백혈병의 두 가지 종류인 급성 골수성 백혈병(acute myeloid leukemia: AML) 환자 11명과 급성 림프성 백혈병(acute lymphoblastic leukemia: ALL) 환자 27명으로부터 얻어진 데이터이다. 림프종 데이터는 45개의 샘플 데이터로 구성되어 있으며, 이 중에서 22개가 GC B-like 림프종 샘플이고, 23개가 Activated B-like 림프종 샘플이다. SRBCT 데이터는 63개의 샘플로 구성되어 있으며, NB (neuroblastoma), RMS (rhabdomyosarcoma), NHL (non-Hodgkin lymphoma), EWS (Ewing family of tumors)의 네 가지 클래스로 이루어져 있다.

4.2 실험 결과

표 1은 백혈병 데이터를 두 개의 클러스터(c=2)로 퍼지 클러스터링한 결과이다. AML 클래스와 ALL클래스로 클러스터링한 결과, histological diagnosis에 의해 기존에 알려진 각 샘플의 클래스가 퍼지 클러스터링에 의해 구해진 클래스와 모두 일치하는 것을 볼 수 있다. 대부분의 샘플이 멤버쉽 값이 0.99 이상으로 높은 소속 정도로 클러스터에 속한 데 반해, AML_13 샘플의 경우 상대적으로 낮은(0.586805) 멤버쉽 값으로 AML에 속하였다. 이 샘플은 클러스터의 수를 네 개로 하였을 때, ALL_B_cell 클래스에 속하였다. 클러스터의 수를 네 개로 하여 클러스터링한 결과 cluster 1과 cluster 2를 ALL_B_cell로, cluster 3을 ALL_T_cell, cluster 4를 AML로 정하였을 때, 세 개의 샘플(ALL_9723_T_cell, ALL_17638_T_cell, AML_13)을 제외한 모든 클래스가 원래의 클래스와 일치하였다.

림프종 데이터에 대하여 클러스터 수를 2로 하여 퍼지 클러스터링한 결과, 각 샘플에 대하여 원래 알려진 histological diagnosis와 실험의 결과인 fuzzy diagnosis

결과로 나누어 분석해본 결과, 하나의 샘플(DLCL-0020)만 빼고는 다 원래 알려진 클래스로 판별되었다. 멤버쉽 값을 보면, 대체로 확연히 구분이 갈 정도로 두 클래스에 대하여 많은 차이를 보인다.

표 1. 백혈병 데이터의 멤버쉽 값(c=2)

Sample label	Cluster1 membership	Cluter2 membership	Fuzzy diagnosis	Histological diagnosis
ALL_19769_B-cell	**0.999973**	0.000027	ALL	ALL_B_cell
ALL_23953_B-cell	**0.999975**	0.000025	ALL	ALL_B_cell
ALL_28373_B-cell	**0.999857**	0.000143	ALL	ALL_B_cell
ALL_9335_B-cell	**0.999526**	0.000474	ALL	ALL_B_cell
ALL_9692_B-cell	**0.999897**	0.000103	ALL	ALL_B_cell
ALL_14749_B-cell	**0.824578**	0.175422	ALL	ALL_B_cell
ALL_17281_B-cell	**0.999917**	0.000083	ALL	ALL_B_cell
ALL_19183_B-cell	**0.999884**	0.000116	ALL	ALL_B_cell
ALL_20414_B-cell	**0.999990**	0.000010	ALL	ALL_B_cell
ALL_21302_B-cell	**0.997542**	0.002458	ALL	ALL_B_cell
ALL_549_B-cell	**0.999187**	0.000813	ALL	ALL_B_cell
ALL_17929_B-cell	**0.999940**	0.000060	ALL	ALL_B_cell
ALL_20185_B-cell	**0.994075**	0.005925	ALL	ALL_B_cell
ALL_11103_B-cell	**0.999322**	0.000678	ALL	ALL_B_cell
ALL_18239_B-cell	**0.998214**	0.001786	ALL	ALL_B_cell
ALL_5982_B-cell	**0.999979**	0.000021	ALL	ALL_B_cell
ALL_7092_B-cell	**0.992895**	0.007105	ALL	ALL_B_cell
ALL_R11_B-cell	**0.999944**	0.000056	ALL	ALL_B_cell
ALL_R23_B-cell	**0.999813**	0.000187	ALL	ALL_B_cell
ALL_16415_T-cell	**0.996841**	0.003159	ALL	ALL_T_cell
ALL_19881_T-cell	**0.999529**	0.000471	ALL	ALL_T_cell
ALL_9186_T-cell	**0.999927**	0.000073	ALL	ALL_T_cell
ALL_9723_T-cell	**0.999377**	0.000623	ALL	ALL_T_cell
ALL_17269_T-cell	**0.999775**	0.000225	ALL	ALL_T_cell
ALL_14402_T-cell	**0.999971**	0.000029	ALL	ALL_T_cell
ALL_17638_T-cell	**0.999914**	0.000086	ALL	ALL_T_cell
ALL_22474_T-cell	**0.999896**	0.000104	ALL	ALL_T_cell
AML_12	0.000216	**0.999784**	AML	AML
AML_13	0.413195	***0.586805***	AML	AML
AML_14	0.000292	**0.999708**	AML	AML
AML_16	0.000078	**0.999922**	AML	AML
AML_20	0.004410	**0.995590**	AML	AML
AML_1	0.003644	**0.996356**	AML	AML
AML_2	0.001089	**0.998911**	AML	AML
AML_3	0.000157	**0.999843**	AML	AML
AML_5	0.002019	**0.997981**	AML	AML
AML_6	0.000301	**0.999699**	AML	AML
AML_7	0.000212	**0.999788**	AML	AML

백혈병 데이터의 ALL 클래스는 다시 ALL_B_cell 클래스와 ALL_T_cell 클래스로 나뉘기 때문에 이러한 클러스터를 찾기 위해서 클러스터의 수를 4로 하여 실험한 결과, 38개의 샘플 중 잘못 분류된 세 개의 샘플을 그림 1에 정리하였다. 이 샘플들은 모두 FCM에 의해 ALL_B_cell 클래스로 클러스터링 되었으며 나머지 35개의 샘플은 올바르게 분류되었다.

ALL_9723_T_cell : ALL_T_cell 클래스
ALL_17638_T_cell : ALL_T_cell 클래스
AML_13 : AML 클래스

그림 1. c=4일 때 백혈병 데이터의 잘못 분류된 샘플

표 2는 하드 클러스터링과 FCM을 이용하여 클러스터링한 후 그 결과를 histological diagnosis에 의해 기존에 알려진 각 샘플의 클래스와 비교하여 구한 암 분류 인식률이다. 모든 데이터에 대해 FCM이 가장 우수한 성능을 보였으며, 백혈병 데이터와 SRBCT 데이터에서는 모든 샘플에 대해 올바른 분류를 하였고 림프종 데이터에서는 45개의 샘플 중에 한 개의 샘플(DLCL-0020)만 틀리게 분류하였다. 림프종 데이터의 이 샘플은 k-means 클러스터링을 제외한 다른 클러스터링 방법에서도 모두 틀리게 분류되었다.

하드 클러스터링 방법 중에는 hard c-means 클러스터링과 평균 연결 클러스터링이 백혈병 데이터와 림프종 데이터에 대해 가장 좋은 성능을 보였고, 완전 연결 클러스터링이 SRBCT에 대하여 가장 좋은 성능을 보였다. 단일 연결 클러스터링은 모든 데이터에 대해 가장 저조한 성능을 보였다. 즉 이진 클래스의 경우는 hard c-means 클러스터링과 평균 연결이 우수한 성능을 보이고, 다중 클래스의 경우는 완전 연결 클러스터링이 우수한 성능을 보였다. 이는 문제 영역에 맞는 거리 척도와 알고리즘이 각기 다를 수 있음을 의미한다. 이에 반해 FCM은 문제 영역에 의존하지 않는 우수한 성능을 보임을 확인할 수 있다.

표 2. 클러스터링 방법의 암 분류 인식률 비교(%)

	계층적 클러스터링				KM	HCM	FCM
	SL	CL	AL	Ward			
백혈병 (c=2)	73.7	81.6	97.4	97.4	97.4	97.4	**100.0**
림프종 (c=2)	51.1	91.1	91.1	82.2	75.6	93.3	**97.8**
SRBCT (c=4)	34.9	**100.0**	87.3	77.8	52.4	77.8	**100.0**

5. 결 론

본 논문에서는 퍼지 클러스터링의 대표적인 알고리즘인 fuzzy c-means 클러스터링 방법을 사용하여 암의 하부 클래스를 예측하였다. 백혈병, 림프종, SRBCT의 세 가지 유전자 발현 암 데이터에 적용한 결과 퍼지 클러스터링 방법이 기존의 하드 클러스터링 방법에 비하여 우수한 성능을 보임을 확인할 수 있었다.

감사의 글

본 논문은 한국전자통신연구원의 지원에 의하여 이루어진 것임.

참고 문헌

[1] F. Höppner, F. Klawonn, R. Kruse and T. Runkler, *Fuzzy Cluster Analysis*, Wiley, 2000.

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

[3] S. Tavazoie, *et al.*, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, pp. 281-285, 1999.