소속도 함수와 신경망을 이용한 유전자 발현 정보의 분류

염해영[0], 문영식
한양대학교 컴퓨터공학과
{hyyeom[0], ysmoon}@cse.hanyang.ac.kr

Classification of Gene Expression Data Using Membership Function and Neural Network

Hae Young Yeom[0], Young Shik Moon
Dept. of Computer Science and Engineering, Hanyang University

요 약

유전자 발현은 유전자가 mRNA와 생체의 기능을 일으키게 하는 단백질을 만들어내는 과정이다. 유전자 발현에 대한 정보는 유전자의 기능을 밝히고 유전자간의 상관 관계를 알아내는데 중요한 역할을 한다. 이러한 유전자 발현 연구를 위한 정보를 대량으로 신속하게 얻을 수 있는 도구가 DNA Chip이다. DNA Chip으로 얻은 수백~수천개의 데이터는 그 데이터만으로는 의미를 갖지 못한다. 따라서 유전자 발현 정도에 따라 수치적으로 획득된 데이터에서 의미적인 특성을 찾아내기 위해서는 클러스터링 방법이 필요하다. 본 논문에서는 수많은 유전자 데이터 중에서 주요 정보를 포함한 것으로 판단되는 유전자 데이터를 선택하여 특징값을 계산하고 신경망 학습을 이용한 클러스터링하는 알고리즘에 대해서 기술한다.

1. 서 론

인체는 다양한 세포로 구성되어 있고 각각의 세포는 단일 세포에서 성장 분화된 것이지만 동일한 게놈(생물의 생존에 필요한 최소한의 염색체)에서 만들어지는 RNA 유전자의 차이에 의해 세포의 형태와 기능이 달라진다. 이는 동일한 세포가 인체 내의 유전자 발현이 조절되어 장기를 형성하는 세포, 혹은 골격을 만드는 세포등으로 분화하기 때문이다. 이와 같은 세포 특성 때문에 유전자 발현을 조절하는 물질이 외부 환경에서 인체 내로 유입하게되면 유전자의 발현을 변화시켜 세포 특성을 변화시키고, 세포 본래의 역할에서 탈피함으로 정상적인 활동을 수행할 수 없게 되거나, 정상 세포의 활동을 억제하게되어 질병을 유발시키게 되는데 이러한 대표적인 사례를 "암"이라 한다. 이러한 유전자 발현에 대한 정보는 유전자의 기능을 밝히고 또한 유전자간의 상관관계를 알아내는데 매우 중요한 역할을 한다. 이는 병의 원인이 되는 유전자를 찾아내어 병의 정확한 진단 및 조기 진단을 가능케 하며 유전자 치료 및 신약 개발의 중요한 목표를 제공하게 된다. 이에 유전자 정보를 대량으로 신속하게 얻을 수 있는 도구인 DNA Chip이 개발되었고, 유전자 발현 연구를 위한 핵심적인 도구로 사용되고 있다. DNA Chip은 한 번의 실험으로 보다 많은 유전자의 발현 정보를 얻을 수 있으나, 실험시 발생할 수 있는 에러를 포함한 데이터에 대한 분석은 모호하고 불명확한 결과를 초래하여 데이터에 대한 패턴 분석을 어렵게한다. 본 논문에서는 이러한 DNA Chip으로 얻은 수많은 데이터들을 의미있는 정보로 조직화하고 분석하기 위한 소속도 함수를 이용한 선택적 데이터 추출 방법과 신경망을 이용한 클러스터링 방법에 대하여 기술한다.

2. DNA Chip

DNA Chip은 기존의 분자 생물학적 지식을 바탕으로 현대에 엄청난 발전을 한 기계 및 전자 공학의 기술을 접목해서 만들어졌다. 이는 기계 자동화와 전자 제어 기술들을 이용하여 적게는 수백 개부터 많게는 수십만 개의 DNA를 아주 작은 공간에 집어넣을 수 있게 만든 것으로 유전자 발현 양상, 유전자 결함 그리고 단백질 분포들을 분석해 낼 수 있는 생물학적 마이크로칩이라 할 수 있다. 이러한 데이터는 과학 기술 연구 및 임상, 진단, 검사 등의 분야에 혁신적 변화를 일으킬 것으로 주목받고 있다. 본 논문에서는 단 한번의 실험으로 빠르고 정확하게 수천 개 이상의 유전자 발현 변이를 검색 할 수 있는 DNA Chip 데이터를 이용한다. 그림1은 DNA Chip을 이용해 데이터를 획득하는 실험 과정을 보여준다.

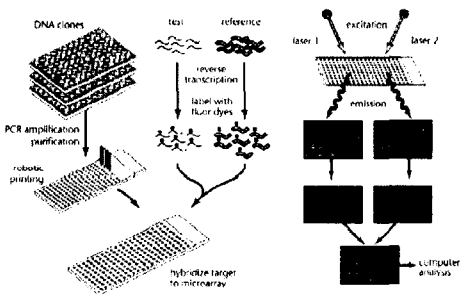

그림 1. cDNA microarray 실험과정

3. 유전자 발현 정보의 분류

DNA Chip으로 얻은 수백~수천개의 유전자 데이터를 모두 이용하여 클러스터링하는 것은 시간적으로나 공간적으로 불가능하다. 그러므로 클러스터링하기에 적절한 유전자 데이터의 선택 과정이 필요하다. 이는 분산값의 비율을 데이터를 선택하고 클러스터링 성능에 효과적인 데이터들의 분류 영향력을 향상시키기 위해 소속도 함수를 이용하여 특징값을 추출한다. 이에 획득된 특징값의 차이를 비교한 클러스터링 영향력에 따른 신경망을 이용한 가중치를 고려하여 클러스터링을 수행한다. 이러한 알고리즘의 전체 구조를 그림2에서 보여주고 있다.

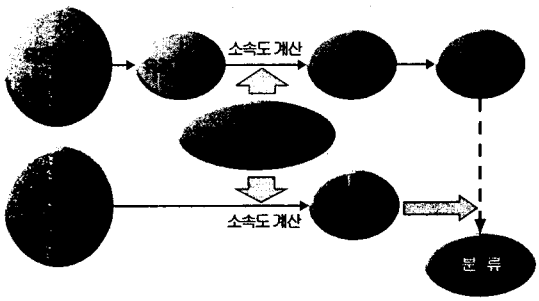

그림 2. 제안하는 알고리즘

3.1 유전자 데이터 선택

DNAChip에서 얻은 수많은 유전자 데이터를 모두 이용하여 클러스터링하는 것은 불가능한 일이다. 따라서 유전자 데이터를 이용하여 클래스를 분류하기 위해서는 클래스와의 연관성이 높은 유전자를 추출하는 과정이 필요하다. 본 논문에서는 유전자 데이터들의 분포를 고려한 선택 방법을 사용한다. 분산은 데이터의 흩어진 정도를 나타냄으로 흩어진 정도가 크면 분산이 큰 값을 갖고 클러스터링하기에 적합하지 않은 분포를 보여준다. 그러므로 식3.1과같이 분산의 비율을 계산하여 클러스터링의 성능을 높일 수 있다.

$$\frac{\text{전 체 데 이 터 의 분 산}}{\text{클 러 스 터 의 분 산}+\text{클 러 스 터 2 의 분 산}} \qquad \text{(식 3.1)}$$

그러나 식3.1을 이용하여 선택된 유전자 데이터가 클러스터링하기에 부적합하거나 클러스터링을 저해하는 데이터가 존재할 수 있으므로 이러한 데이터의 영향을 최소화하고 나머지 유전자 데이터의 클러스터링 성능을 향상 시킬 수 있는 특징값을 추출한다.

3.2 특징값 추출

3.2.1 소속도 함수 설계

소속도 함수는 선택된 데이터가 서로 다른 클래스에 속하는 정도를 수치적으로 나타냄으로, 모호한 데이터들이 나타내는 서로 유사한 데이터 분포를 데이터들의 소속 정도를 이용하여 클러스터링하기에 적합한 유전자 데이터들의 영향력을 높여주는 방법이다. 이에 그림 3은 소속도 함수 계산 과정을 도표로 보여주고 있다.

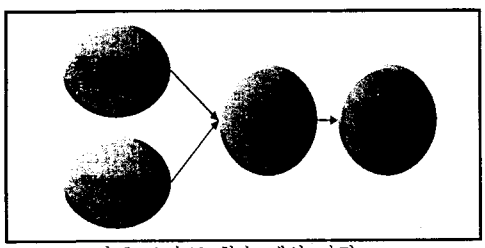

그림 3. 소속도 함수 계산 과정

[I] 정규화된 히스토그램 계산

선택된 유전자가 존재하는 구간을 N(정규화된 히스토그램 레벨)으로 나누어 데이터의 분포를 반영하는 히스토그램을 계산한다. S는 선택된 유전자에 해당하는 데이터이고, I는 두 개의 클래스, j는 선택된 유전자 데이터이고, m은 선택한 유전자 데이터의 개수이다.

$$H_j^i(k) = histogram(S_j^i) \qquad \begin{matrix} 1 \le i \le 2 \\ 1 \le j \le m \\ 1 \le k \le N \end{matrix}$$

[II] 평균위치를 기준으로 거리를 고려한 Weight 계산

N레벨로 정규화된 히스토그램을 기반으로 클래스 각각에 평균위치 C를 기준으로 다음 식을 적용하여 거리 관계를 계산한 Weight를 정의한다. 이는 C로부터 멀수록 데이터의 분산값이 크다는 의미다.

$$W_j^i(k) = \begin{cases} 1/\alpha\left\{\left(\dfrac{C_j^i - k}{C_j^i - 1}\right)^2 + 1\right\} & k \le C_j^i \\ 1/\alpha\left\{\left(\dfrac{C_j^i - k}{C_j^i - N}\right)^2 + 1\right\} & k > C_j^i \end{cases}$$

[III] Weight 적용

정규화된 히스토그램의 데이터를 평균위치를 고려하여 계산한 Weight로 나누면 평균위치 C로부터 멀리 떨어질수록 히스토그램은 낮은 가중치를 갖는다. 이는 선택된 유전자 데이터가 평균위치로부터 멀리 떨어진 클러스터링하기에 모호한 데이터들의 소속 정도를 낮춰주는 성능을 보여준다.

$$H_j^{i'}(k) = \frac{H_j^i(k)}{W_j^i(k)} \qquad \begin{matrix} 1 \le i \le 2 \\ 1 \le j \le m \\ 1 \le k \le N \end{matrix}$$

[IV] 누적 소속도함수 계산

누적소속도 함수는 Weight를 적용하여 얻어진 데이터들의 주변 데이터들과의 관계를 고려한 계산 결과이다.

$$F_j^i(k) = \begin{cases} H_j^{i'}(k) + H_j^{i'}(k-1) & 2 \le k \le C_j^i \\ H_j^{i'}(k) + H_j^{i'}(k+1) & C_j^i < k < N \end{cases}$$

이러한 소속도 함수를 이용하여 선택된 데이터를 적용한 특징값 벡터를 계산한다.

3.2.2 샘플의 특징 벡터

선택된 유전자 데이터를 소속도 함수 F^1과 F^2 각각에 대입한 차이를 계산한 샘플의 특징 벡터 SV를 다음 과 같이 정의한다. i, j, k는 앞서 정의한 변수들을 의미하고 선택된 유전자를 가진 조직의 개수를 $p+q$로 나타낸다. 이는 클래스 각각에 선택된 유전자 개수 m만큼의 샘플의 특징 벡터를 얻고, 특징 벡터는 서로 다른 클래스에 p개, q개의 조직을 갖음을 의미한다.

$$SV_j^i(t) = F_j^1(S_j^i(t)) - F_j^2(S_j^i(t)) \qquad \begin{matrix} 1 \le i \le 2 \\ 1 \le j \le m \\ 1 \le k \le N \\ 1 \le t \le p+q \end{matrix}$$

3.2.3 클래스의 특징 벡터

각각의 클래스에 존재하는 유전자 데이터 개수 만큼의 샘플의 특징 벡터는 p개, q개의 조직으로 구성된다. 선택된 유전자에 해당하는 샘플 특징 벡터들의 값들의 평균은 다음 식으로 구성한다. 이를 각각의 클래스에 선택한 유전자 개수를 가진

클래스의 특징 벡터(CV)라 한다.

$$CV_j^i = \begin{cases} \dfrac{\sum_{t=1}^{p} SV_j^i(t)}{p} & i=1 \\ \dfrac{\sum_{t=p+1}^{p+q} SV_j^i(t)}{q} & j=1 \end{cases}$$

이는 클래스 특징 백터를 포함하는 유전자가 선택된 유전자들에 의해 두 클래스 특징 벡터에 각각의 유전자의 차이를 계산하면 두 분포가 클러스터링하기에 적합한 분포를 가진 유전자인지 아닌지를 판별할 수 있다.

3.3 클러스터링

소속도 함수를 이용하여 특징값을 추출하여 얻은 결과가 선택된 유전자 데이터의 클러스터링 성능을 향상 시킬 수 있는 데이터 집합일 수도 있지만 클러스터링을 저해하는 구분이 모호한 선택된 유전자 데이터 집합일 수도 있다. 이에 클러스터링하기에 적절하지 않은 데이터들의 클러스터링 결과를 향상시킬 수 있는 가중치 계산을 이용하기 위해 신경망 학습을 적용한다. 본 논문에서 제안하는 신경망 알고리즘은 그림3과 같이 구성하였다. 앞서 소개한 소속도 함수를 이용하여 계산한 클래스 특징 벡터의 차이값을 가중치 벡터의 입력값으로 대입하고, 유전자 데이터의 소속도 함수의 벡터값을 대입하여 신경망 학습을 수행한다. 이는 선택된 유전자 데이터에 따른 소속 정도에 따라 가중치를 부여하므로 클러스터링이 가능하도록 한다.

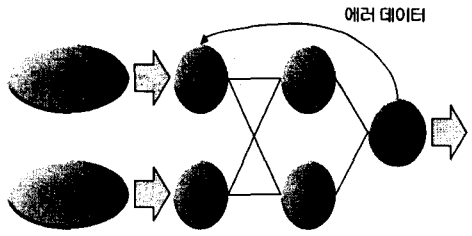

그림 3. 제안한 알고리즘의 신경망 구조

4. 실험 결과

특징 선택 방법으로 Euclidean, Cosine, MI(상호정보), SN(신호대 잡음비)를 이용한 결과 데이터와 본 논문에서 제안한 알고리즘을 이용하여 얻은 데이터와의 성능을 비교해 본다. 이는 특정 알고리즘을 이용하고 신경망을 적용한 데이터 값이므로 본 논문에서의 데이터 선택과 신경망 적용이 같아 적합한 비교 분석이 가능하다. 실험적으로 결정된 바에의해 제안한 알고리즘에서 가장 좋은 성능을 보였던 소속도 함수 레벨이 5이고 선택된 유전자 데이터가 10개 일 때의 인식율과 성능비교를 해보면 다른 알고리즘을 적용한 데이터보다 우수한 성능을 보인다.

표 1. 성능 비교(%)

적용 알고리즘	인식율
Euclidean	91.2%
Cosine	94.1%
MI(상호정보)	64.7%
SN(신호대 잡음비)	67.6%
제안하는 알고리즘	**97.2%**

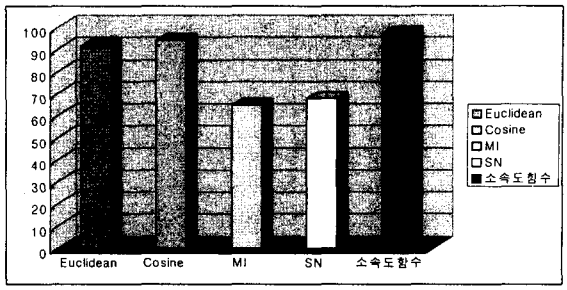

그림 4. 성능 비교 결과

5. 결론 및 향후과제

본 논문에서는 소속도 함수와 신경망을 이용한 유전자 발현 정보를 분류하는 기법을 제안하였다. DNA Chip에 의해 얻어진 수많은 데이터를 분석하기 의해 의미있는 적절한 유전자 데이터를 선택하고 소속도 함수를 사용하여 특징값을 구하고, 신경망에 적용하여 클러스터링하였다. 본 논문에서 제안하는 알고리즘의 결과는 기존의 연구와 비교하여 3~30% 향상된 인식율을 얻을 수 있었다. 그러나 유전자 발현 데이터들 간의 의미있는 상관 관계를 찾는 일은 클러스터링으로만 적용하기에는 어려운 일이므로 앞으로 어 많은 유전자 데이터를 통한 실험과 기존의 연구에 대한 비교 검증이 필요하며 더 효과적인 클러스터링을 위한 학습 방법과 유전자 발현 데이터 분석에 대한 연구가 계속되어야 할 것이다.

참고문헌

[1] 권영준, 류중원, 조성배, "신경망 분류기를 이용한 암 관련 유전자 발현정보의 분류", *한국정보과학회 춘계학술발표논문집(B)*, vol.28, pp.295-297, 2001.

[2] 원홍희, 조성배, "암 분류를 위한 기계학습 분류기의 성능 평가", *한국정보처리학회*, vol.9, 2002.

[3] U.Alon, N.Barkai, D.A.Notterman, K.Gish, S.Ybarra, D.Mack, A.J.Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proceedings of National Academy of Sciences of the USA*, vol.96, pp.6745-6750, June 1999.

[4] Eisen MB, Spellman PT, Brown PO et al. "Cluster analysis and display of genome-wide expression patterns", *Proceedings of National Academy of Sciences of the USA*, vol.95, pp.14863-14868, 1998.

[5] T.R.Golub, D.K.Slonim, P.Tamayo, C.Huard, M.Gaasenbeek, J.P.Mesirov, H.Coller, M.L.Loh, J.R.Downing, M.A.Caligiuri, C.D.Bloomfield, E.S.Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *SCIENCE*, vol.286, October 1999.