제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

클래스 정보에 기반한 유전자 선택

이현진*
*한국싸이버대학교 컴퓨터정보통신 학부
e-mail: hjlee@mail.kcu.ac

Gene Selection based on Class Information

Hyunjin Lee*
*Dept. of Computer , Information & Communication, Korea Cyber University

요 약

여러 분류 문제에 다층퍼셉트론이 적용되어 좋은 성능을 보이고 있다. 하지만, 암 분류를 위한 분류기로 사용되는데 있어서 문제점은 샘플데이터 수에 비해 입력으로 사용되는 유전자의 수가 너무 많기 때문에 좋은 성능을 기대하기 힘들다는 점이다. 또한 많은 입력노드로 인해 가중치 파라메터들의 수가 증가하기 때문에 학습시에 계산량의 부담을 가중시킨다. 따라서 본 논문에서는 많은 유전자중에서 암분류에 중요한 영향을 끼치는 유전자를 선택하는 방법을 제안한다. 이러한 유전자 선택을 위하여 클래스의 정보를 나타내는 척도를 분석하고 이를 기반으로하여 분류율을 향상시킬 수 있는 유전자를 선택하는 방법을 제안한다. 이렇게 선택된 유전자를 입력으로 하여 분류기를 구성하여, 제안하는 방법의 우수성을 검증한다.

1. 서 론

DNA 마이크로어레이(microarray)는 유전자의 발현 변화를 모니터링 하기 위한 cDNA 마이크로어레이 칩이 개발된 이후, 유전 정보와 그 기능을 규명하는 기능적 유전체학(Genomics)의 필수적 기술로 인식되고 있다. DNA 마이크로어레이는 다양한 질병의 메커니즘을 탐색하고 신약개발을 가속시키는 강력한 도구로 자리를 잡았다. 이러한 DNA 마이크로어레이 기술은 암의 예측과 진단분야에 활용되고 있다[1].

유전자 분석의 첨단기술로 인정되고 있는 DNA 마이크로어레이 기술은 짧은 시간에 많은양의 데이터를 만들어낼 수 있다. 하지만 이러한 많은 양의 유전자 정보가 암을 분류하는데 모두 필요한 것은 아니다. 따라서 이러한 DNA 마이크로어레이의 방대한 양의 유전자 정보중에서 암분류에 필요한 유전자 선택(Gene Selection)이 필요하다.

다층퍼셉트론 신경회로망은 여러 분류 문제에 적용되어 좋은 성능을 보이고 있다. 이러한 다층퍼셉트론은 DNA 마이크로어레이 기술로 얻은 유전자 정보를 분석하는데 사용될 수 있다. 다층퍼셉트론은 입력정보의 양이 많을수록 신경회로망의 노드들이 늘어나게 되고, 이렇게 늘어난 노드는 학습시에 많은 계산량을 요구한다[2].

따라서 본 논문에서는 다층퍼셉트론을 이용하여 암을 분류하는데 있어서 성능을 향상시키기 위한 클래스 정보에 기반한 유전자 선택 방법과 효율적인 분류를 위한 분류기 구성을 제안한다.

유전자 선택시에는 클래스내의 응집도와 다른 클래스와의 차이도를 하이브리드한 방법을 제안하다. 이를 통하여 분류를 위한 최적의 유전자들을 선택하여 신경회로망의 입력으로 사용한다. 분류기 구성에 있어서 신경회로망의 오차함수는 분류의 성능을 높일 수 있는 크로스엔트로피(Cross Entropy)함수를 사용한다. 학습은 빠른 학습과 플라토 문제를 해결한 자연기울기 강하 학습 방법을 적용한다. 학습하는데 있어서 데이터 획득시의 오차와 획득된 데이터의 수가 적음으로 저하될 수 있는 일반화 성능을 향상시키기 위하여 정규화 방법을 도입한다. 정규화 방법은 적응적으로 정규화 파라메터를 갱신할 수 있는 베이시안 정규화를 도입한다.

본 논문의 구성은 다음과 같다. 2 장 본론에서는 실험을 위한 유전자 선택 및 분류기 구성 방법을 다룬다. 2.1 절 유전자 선택 방법에서는 기존의 유전자 선택 방법과 제안하는 유전자 선택 방법을 살펴본다. 2.2 절 분류에서는 2.1 절의 방법으로 선택된 유전자들이 좋은 일반화 성능을 갖도록 분류기를 구성하는 방법

을 살펴본다. 3 장에서는 제안하는 방법의 성능을 실험을 통하여 검증한다. 3.1 절에서는 사용되는 데이터와 그 데이터를 이용한 클래스 정보에 관해 살펴본다. 3.2 절에서는 실험 결과를 살펴본다. 마지막으로 4 장에서는 결론을 내린다.

2. 본 론

다층퍼셉트론을 이용하여 유전자를 분류하기 위해서는 그림 1 같은 단계를 거쳐야 한다. 먼저 입력데이터로부터 유전자 선택 방법을 거쳐서 암분류를 위한 최적의 유전자들을 선택하고 이렇게 선택된 유전자를 다층퍼셉트론에 학습 시켜서 원하는 결과를 얻게된다. 입력정보가 좋지 못하면 좋은 분류성능을 기대할 수 없다. 본론에서는 적절한 유전자를 선택하는 방법과 이를 이용하여 분류하기 위한 분류기 구성에 대하여 살펴본다.

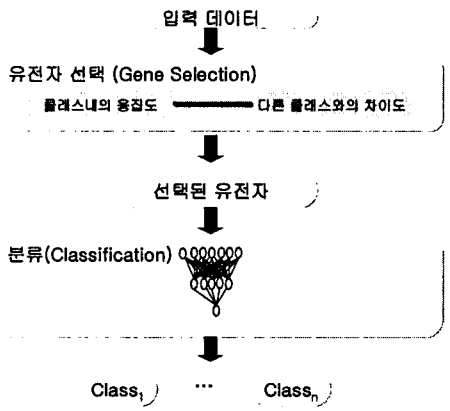

그림 1 유전자 선택 및 분류과정

2.1 유전자 선택(Gene Selection)

유전자 선택에 있어서 널리 사용되는 기준들은 통계학에서 일반적으로 많이 사용되는 변량들 간의 거리에 대한 평가 기준이나 컴퓨터 이론의 데이터의 정보량에 관련된 척도이다. 이러한 척도들로는 유클리드 거리(Euclidean distance), 피어슨 상관관계 계수(Pearson correlation coefficient), 스피어만 상관관계 계수(Spearman correlation coefficient), 코사인 상관관계 계수(Cosine coefficient), 정보이득(Information gain)등이 있다 [1][3][4].

본 논문에서는 클래스 분석을 통하여 유전자들간의 관계를 분석하여 최적의 유전자를 선택한다. 클래스 분석시 사용될 수 있는 유사도 척도로는 최단연결법, 최장 연결법, 평균 연결법, 중심 연결법등이 있다 [5][6][7]. 본 논문에서는 평균 연결법에 의하여 클래스 내의 응집도와 다른 클래스와의 차이도를 하이브리드한 척도를 이용하여 유전자를 선택한다.

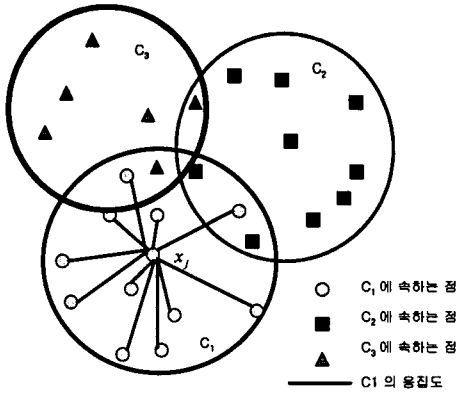

그림 2 클래스 내의 응집도

클래스 분석을 위하여 유전자에 속한 값들이 [0,1]범위에 속하도록 정규화(normalize) 시킨다. 이렇게 정규화된 값을 바탕으로 클래스 정보 분석을 수행한다.

먼저 그림 2 와 같이 어떠한 클래스 C_1에 속하는 한 점 x_j와 클래스 C_1에 속하는 나머지 점들 사이의 거리를 구한다. 이러한 거리를 클래스에 속하는 모든 점들에 대해 구하여 평균을 구하여 클래스의 응집도(cohesiveness)를 구한다. 이러한 응집도를 구하는 식은 식(1)과 같다.

$$CI = \sum_{i=1}^{n}\left(\frac{D_i}{n}\right) \tag{1}$$

$$D_i = \frac{\sum_{j}^{m_i}\sum_{k}^{m_i}\sqrt{(x_j - x_k)^2}}{m_i(m_i-1)} \tag{2}$$

식(1)은 식(2)에서 정의하는 D_i에 대한 평균으로 구성된다. 여기서 n 은 클래스의 수이다. 식(2)는 그 클래스에 속한 점들 사이의 거리에 대한 평균이다. 여기서 m_i는 클래스 i에 있는 점의 수이다.

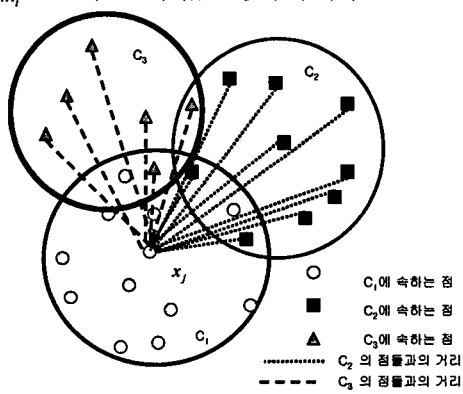

그림 3 다른 클래스와의 차이도

다른 클래스와의 차이도는 그림 3 과 같이 현재의 클래스 C_1에 속한 점 x_j와 다른 클래스의 점들과의 거리의 평균으로 구해진다.

클래스 l에 속하는 점들과 다른 클래스와의 차이도는 식(3)과 같다. 여기서 l은 현재 속한 클래스 이고 n 은 전체 클래스의 수이다. 차이도를 구할 때 $j=l$인 것은 제외된다. 식(3)의 CO_l은 식(4)의 D_l의 평균으로 구해진다. 식(4)는 현재 속한 클래스 l의 점들과 클래스 i 에 대한 점들의 거리차이의 평균값으로 구해진다. 여기서 x_j는 l클래스에 속하는 점의 값이고, y_k는 i 클래스에 속하는 점들의 값이다. m_l은 l클래스에 속하는 점들의 수이고, m_i는 i 클래스에 속하는 점의 이다.

$$CO_l = \sum_{i=1}^{n}\left(\frac{D_l}{n-1}\right), \text{단 } j \neq l \tag{3}$$

$$D_l = \sum_{j}^{m_l}\sum_{k}^{m_i}\frac{\sqrt{(x_j - y_k)^2}}{m_l m_i} \tag{4}$$

본 논문에서는 앞에서 살펴본 클래스내의 응집도와 다른 클래스와의 차이도를 하이브리드한 방법을 제안한다. 본 논문에서 사용한 척도 CH_l은 식 (5)와 같다. CH_l은 NCI에 비례하고 NCO_l에 반비례한다. NCI는 식(1)의 클래스 내의 응집도에 대해 [0,1]의 값으로 정규화한 값이다. NCO_l은 식(3)에서 구한 다른 클래스와의 차이도이며 이를 [0,1]의 값으로 정규화하였다. 정규화를 통하여 둘 사이의 값의 영향력을 균등하게 하였다. 그리고 이를 바탕으로 유전자를 선택하는데 있어서 응집도가 높은 클래스로 구성되는 것은 분류에 비례하며, 다른클래스와의 차이도는 클수록 좋기 때문에 반비례하는 관계를 식으로 구성하여 유전자를 선택하는데 적용하였다.

$$CH_l = \frac{NCI}{NCO_l} \tag{5}$$

$$NCI = \frac{CI - \min(CI)}{(\max(CI) - \min(CI))} \tag{6}$$

$$NCO_l = \frac{CO_l - \min(CO_l)}{(\max(CO_l) - \min(CO_l))} \tag{7}$$

2.3 분류(Classification)

2.2 절에서 살펴본 유전자 선택방법에 의해 선택된 유전자가 분류에 효과적인지를 검증하기 위하여 다층 퍼셉트론을 사용한다. 학습 방법으로는 자연기울기(natural gradient)강하 학습법을 적용한다. 자연기울기 강하 학습 방법은 오류 역전파(backpropagation)학습방법에 비해 좋은 성능을 보이는 것이 밝혀져 있다. 자연기울기 강하 학습 방법은 학습에 있어서 문제가 되는 플라토를 해결함으로 인해서 빠른 속도로 수렴할 수 있는 학습 방법이다[8].

학습을 통하여 궁극적으로 얻고자 하는 신경회로망은 학습데이터를 잘 분류하는 분류기를 구성하는 것이 아니라 학습되지 않은 데이터가 들어 왔을 때 분류를 잘 할 수 있는 분류기를 구성하고지 하는 것이다. 이러한 능력을 일반화 성능이라고 하며, 이러한 일반화 성능을 향상시키기 위해서는 일반화 성능을 향상시키기 위한 메커니즘이 필요하다[9]. 본 논문에서는 자연기울기 강하 학습에 베이시안 적응적 정규화를 도입하여 신경회로망 학습시에 일반화 성능을 높이는 방법을 도입한다. 적응적 정규화 항이 있는 오차 함수는 식(8)과 같이 정의 된다. $E(x,y,\theta)$는 일반적으로 사용하는 오차함수이고, $R(\theta)$는 정규화 항이고 α는 정규화의 영향력을 조정하는 파라메터이다.

$$C(x,y,\theta) = E(x,y,\theta) + \alpha R(\theta) \tag{8}$$

여기서 $E(x,y,\theta)$는 분류에 좋은 크로스엔트로피 오차함수를 적용한다. 회귀와 같은 연속적인 값에 대해서는 제곱합 오차가 적합하고, 분류를 위한 오차함수는 크로스엔트로피가 적합하다. 크로스엔트로피 오차함수는 2 개의 클래스 분류인 경우 식(9)와 같고 다중 클래스 분류인 경우는 식 (10)과 같다.

$$E(x,y,\theta) = -y\log f(x,\theta) - (1-y)\log(1-f(x,\theta)) \tag{9}$$

$$E(x,y,\theta) = -\sum_{i=1}^{l} y_i \log f_i(x,\theta) \tag{10}$$

정규화 항은 식 (11)과 같이 가중치 파라메타 θ의 식으로 표현된다.

$$R(\theta) = \|\theta\|^2 \tag{11}$$

이런 정규화 항이 있는 경우 자연기울기 학습은 식(12)와 같이 주어진다.

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta_t \tilde{\nabla} C(x,y,\theta_t) = \theta_t - \eta_t G^{-1}\nabla C(x,y,\theta_t) \\ &= \theta_t - \eta_t G^{-1}(\nabla E(x,y,\theta_t) + \alpha\nabla R(\theta_t))\end{aligned} \tag{12}$$

η은 학습률 이고, 리마니안 메트릭 텐서는 식(13)과 같이 계산된다.

$$\begin{aligned}G(\theta) &= \iint \frac{\partial \log p}{\partial\theta}\left(\frac{\partial \log p}{\partial\theta}\right)^T p(y|x,\theta)q(x)dydx \\ &= E_x\left[E_{y|x,\theta}\left[\frac{\partial \log p(y|x,\theta)}{\partial\theta}\left(\frac{\partial \log p(y|x,\theta)}{\partial\theta}\right)^T\right]\right]\end{aligned} \tag{13}$$

식(8)의 정규화항 파라메터 α에 의해 정규화의 성능차이가 발생하며 이를 자동적으로 적응시키는 방법은 식(14)와 같다[10](자세한 유도는 [2] 참조).

$$\alpha = \frac{n}{2\,NR(\theta)} \tag{14}$$

여기서 n은 가중치의 수이고, N은 데이터의 수이다.

3. 실험 결과 및 분석

3.1. 실험 데이터

실험데이터는 백혈병데이터를 사용하였다. 이 데이터는 백혈병의 2 가지 종류인 급성 골수성 백혈병

(Acute Myeloid Leukemia)과 급성 림프성 백혈병(Acute Lymphoblastic Leukemia) 환자를 분류하는 문제이다. 백혈병 데이터는 72 개의 샘플데이터로 구성되며 이중 38 개는 학습데이터로 34 개는 테스트데이터로 사용하였다. 각각의 샘플은 7129 개의 유전자 발현정보를 가지고 있다[11].

그림 4 와 같이 학습데이터에 대해 다음과 같은 분석을 통하여 유전자를 선택한다. 1 개의 유전자에 대해 AML, ALL 2 개의클래스로 나뉘어 지고 이들 클래스 간에 클래스 내의 응집도, 다른 클래스와의 차이도, 이들간의 하이브리드 방법을 분석한다. 이러한 클래스 분석을 7129 개의 유전자에 대해 수행하고 이들의 값을 비교하여 유전자를 선택한다.

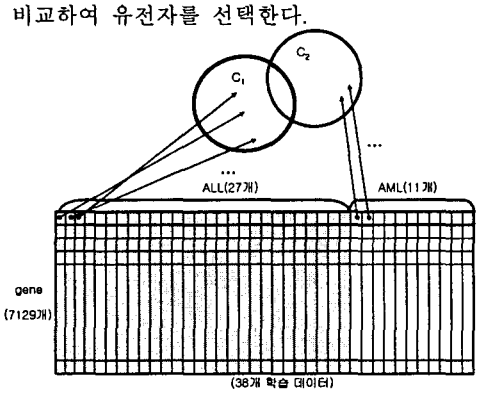

그림 4 백혈병 데이터의 클래스 분석

3.2. 실험 결과

실험은 클래스내 응집도, 다른 클래스와의 차이도, 클래스내 응집도와 다른 클래스와의 차이도의 하이브리드 방법을 비교하였다. 표 1 에서 CI 는 클래스내 응집도 CO 는 다른 클래스와의 차이도, CH 는 클래스내의 응집도와 다른 클래스와의 차이도의 하이브리드 방법이다.

유전자 선택은 각각의 방법으로 50 개의 유전자를 선택하여 2.3 절에 설명한 분류기로 분류하였다. 신경회로망의 구성은 입력노드 50 개 은닉노드 10 개 출력노드 1 개이다. 학습률은 0.07 로 하였으며, 정규화 파라메터는 주기적으로 갱신하였다. 학습회수는 최대 100 회로 하였다.

표 1 암분류 실험 결과

	CI	CO	CH
학습데이터	100%	94.74%	100%
테스트데이터	79.42%	76.47%	97.06%

표 1 에서 보는 바와 같이 암 데이터를 분류하는데 있어서 집단내의 응집도(CI)가 다른 클래스와의 차이도(CO)보다 분류에 있어서 좋은 성능을 보였다. 하지만 제안하는 집단내의 응집도와 다른 클래스와의 차이도를 하이브리드한 방법으로 선택된 유전자들이 분류에 가장 좋은 성능을 보였다.

4. 결 론

본논문에서는 암 분류를 위한 유전자 선택방법들을 비교해 보고, 이를 바탕으로 분류기를 구성하여 인식하는 방법을 제안하였다.

본 연구를 통하여 유전자 선택에 있어서 클래스의 정보가 유전자 분석에 영향을 미치는 것을 확인하였다. 또한 유전자 선택시 클래스 분석을 위한 척도가 분류 향상을 위한 유전자 선택에 중요한 영향을 미치는 것을 실험을 통하여 확인하였다.

향후 연구과제는 다음과 같다. 다양한 데이터에 대한 실험을 통하여 제안하는 방법의 일반성을 검증하는 것이 필요하다. 또한 클래스 분석을 통하여 유전자 선택을 위한 중요도를 구할 수 있었다. 이러한 유전자를 분류하는 것 뿐만 아니라 이들간의 관계성 규명이 필요하다. 이를 위하여 신경회로망의 구조를 최적화시키는 프루닝 방법을 도입하여 신경회로망을 단순화시키고, 이러한 신경회로망을 바탕으로 관계성을 분석하는 연구가 필요하다.

참고 문헌

[1] 여상수, 김성권, "DNA 마이크로어레이 데이터 클러스터링 알고리즘의 연구 동향, " 한국정보과학회 컴퓨터이론연구회지, vol. 12, no.1, pp.2-11, 2001.

[2] C. M. Bishop,"Neural Networks for Pattern Recognition," Oxford University Press, 1995.

[3] Kyeong Eun Lee, Naijun Sha, Edward R. Dougherty, Marina Vannucci, Bani K. Mallick, "Gene selection: a Bayesian variable selection approach," Bioinformatics, vol.19, no.1, pp 90-97, 2003.

[4] 원홍희, 조성배,"암 분류를 위한 음의 상관관계 특징을 이용한 앙상블 분류기," 정보과학회 논문지: 소프트웨어 및 응용, vol. 30, no, 11.12, pp.757-759, 2003.

[5] A.K.Jain, M.N.Murty, P.J.Flynn,"Data clustering:a review," the ACM Comput.Surv. 31,3,pp.264-323, Sep.1999.

[6] 안병주, "데이터 마이닝을 위한 계층적 대표값 군집화 기법," 연세대학교 석사학위 논문, 2001.

[7] Richard O.Duda, Peter Eart, David G.Stork, "Pattern Classification, Second Edition,"WILEY-INTERSCIENCE, 2001.

[8] 박혜영, 아마리 슌이치, 이일병, "다층 퍼셉트론 학습의 플라토 문제에 대한 정보기하 이론적 접근," 정보과학회논문지, vol. 26, no.4, pp 546-556, 1999.

[9] H. Park, "Practical Consideration on Generalization Property of Natural Gradient Learning, LNCS, 2084, 402-409, 2001.

[10] Hyunjin Lee, Hyeyoung Park, Yillbyung Lee, "Network Optimization through Learning and Pruning in Neuromanifold," LNCS, 2417, 2002.

[11] http://www.genome.wi.mit.edu/MPR