유전자 알고리즘과 신경망을 이용한 DNA Chip 유전자 선택 방법 연구

*이호일 ○ *최요한 ○ *윤경오 *김명선 *강연수 **박현석
*마크로젠 BI연구소 ** 이화여자대학교 컴퓨터학과
headil@macrogen.com skyhani@macrogen.com yoonko@macrogen.com
keystore@macrogen.com kys1113@macrogen.com hspark@macrogen.com

DNA Chip Gene Selection Method Research using Genetic Algorithm and Neural Network

*Ho Il Lee○, *Yo Han Choi○, *Kyong Oh Yoon, Myoung Sun Kim,*Youn Soo Kang, **Hyun Seok Park
*Bioinformatics Institute, Macrogen,
**Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 유전자 칩의 발전으로 다양하고 방대한 양의 유전자 정보를 이용한 정확하고 신뢰성 높은 분류, 군집 및 질병을 예측하는 분석 기법이 증가하고 있다. 하지만 특징적인 유전자를 선택하는 Gene Selection 기법의 종류는 많지가 않으며 주로 통계적인 방법에 의존하여 유전자를 선택하는 기법을 많이 사용하고 있다. 본 논문에서는 유전자 알고리즘과 신경망의 결합을 통한 데이터마이닝을 기반으로 신뢰성 높은 특징적인 유전자를 선택하는 Gene Selection 기법에 대하여 연구을 진행하였다.

1. 서 론

DNA Chip의 연구는 최근 몇 년 동안 많은 발전을 거듭하면서 수천에서 수십만 종류의 유전자 배열을 배치할 수가 있게 되면서 DNA Chip 뿐만 아니라 다양한 유전자 Chip 분석의 중요성이 증가고 있다. DNA Chip은 방대한 양의 정보를 칩 실험을 통한 유전자의 발현정보를 이용하여 짧은 시간에 유전자의 기초연구 및 각종 질병의 원인을 규명할 수 있으며 연구의 결과에 대하여 많은 시간과 비용을 절약할 수 있게 되었다[1].

하지만 DNA Chip의 수많은 정보를 이용하여 분석하기란 쉬운 일이 아니며 때로는 불필요한 유전자 정보들로부터 정확한 분석의 어려움을 가져오게 된다. 이러한 불필요한 정보를 제거하고 특징적인 유전자만을 선택하는 Gene Selection(유전자 선택) 방법에는 일반적으로 통계적인 계산법을 이용한 Gene Selection 방법이 주로 사용되고 있으며 최근에는 데이터마이닝 기법을 적용하는 Gene Selection 기법이 연구되고 있다. 특징 유전자를 선택해야 하는 이유는 분류기(classifier)의 성능을 향상시키고 복잡도(complexity)를 감소시키는 등 효율적인 패턴분류를 가능하게 하기 위해서이다. 특징추출은 패턴분류나 패턴인식 문제에 있어서 매우 중요하게 다루어져 왔던 주제로 오랫동안 많은 연구자들에 의하여 폭넓게 연구되어 왔다.

본 논문에서는 진화적인 데이터마이닝을 적용하여 효과적인 분석을 할 수 있도록 특징적인 유전자를 선택할 수 있는 방법을 소개하고자 한다.

2. Gene Selection 관련 동향

DNA Chip에는 수많은 유전자가 존재하지만 연구에 불필요한 유전자가 상당수 존재하고, 이러한 유전자는 정확한 결과를 원하는 DNA Chip 분석에 혼란을 초래하는 요인이 된다. 따라서 최근에는 DNA Chip 분석 목적에 따라 중요한 유전자를 선택 할 수 있는 시스템이 요구되어지고 있으며 보다 정확한 분석을 위해 방대한 유전자들의 정보 속에서 특징적인 유전자만을 선택하여 분석하기 위하여 많은 연구가 진행되고 있다.

2.1 RankGene

Gene Selection을 위한 일반적인 통계적 기법으로는 t-statistic, Twoing rule, information gain, gini index, max minority, sum minority, and sum of variances, 등의 다양한 방법들이 사용되고 있다. 이러한 방법들은 T-statistic의 계산 법을 사용하여 발현값의 절대값을 내림차순으로 정렬하여 유전자를 선택하여 분석하는 방법이 있으며 유전자 발현 값을 up-regulation과 down-regulation 두 개의 부분으로 분류하여 가장 정확하게 class를 분류하는 유전자를 선택한 다음 분류된 샘플들은 모두 같은 class에 속한다고 가정하여 정확도를 측정하고 유전자를 선택하는 일반적인 계산법이 사용되고 있다.

2.2 Decision Tree

의사결정규칙(decision rule)을 도식화하여 관심대상을 몇 개의 소집단으로 분류(Classification)하거나 예측(Prediction)을 수행하는 분석방법으로 과정이 나무구조에 의해서 표현되기 때문에 다른 방법들(Neural Networks,Discriminant Analysis,Regression Analysis)에 비해 이해와 설명이 쉽다는 장점을 가지고 있다. 나무구조에서 각 노드는 Gene에 해당하며 각 Gene의 발현값에 의하여 분기가 결정되며 나무구조의 마지막 노드는 의사 결정값에 해당한다. 이 나무구조에서 선택되어진 Gene을 선택하는 방법으로 Gene Selection을 한다.
하지만 의사결정나무에서는 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 경계점 근방에서는 오류의 가능성을 가지고 있다.

2.3 Genetic Algorithm and k-Nearest Neighbor

kNN(k-Nearest Neighbor) 분류자는 표본의 분포 상태에 영향을 받지 않는 non-parametric 학습 방법의 하나로서,표본이 n-차원 공간상의 점들로 대응된다고 가정한다. n-차원 공간에서 자신과 가장 가깝게 위치하는 k개의 다른 표본들의 클래스중에서 가장 많은 것으로 분류되는 것이 kNN 알고리즘이다.
GA/kNN 방법은 GA(Genetic Algorithm)와 kNN를 이용한 Gene Selection 방법으로서 GA의 적합도 함수(fitness function)로써 kNN을 사용하는 방법이다.
GA/kNN 은 Gene들의 구분없이 단순히 Gene들간의 거리만을 이용하여 Gene Selection을 함으로써 잘못된 결과를 도출할 수 있다[2].

3. GA – Artificial neural networks(ANNs)를 이용한 Gene Selection 방법

유전자 칩 분석을 통하여 정확한 분류와 예측을 하기 위해서는 다양한 알고리즘의 상호 연동을 통하여 특징적인 유전자만을 정확히 알아내는 연구의 필요성을 갖게 되었으며 이러한 연구를 위하여 진화적인 알고리즘을 통하여 신경망을 이용한 Gene Selection 기법을 연구하였다. 본 장에서는 유전자 알고리즘과 신경망 알고리즘을 간단히 소개한 후 서로간의 유기적인 융합을 통한 Gene Selection 기법을 설명 하고자 한다.

3.1 Genetic Algorithm

Michigan 대학의 John Holland에 의해서 개발된 유전자 알고리즘은 주어진 환경에 잘 적응하는 유전자만을 선택(selection)하고 교배(crossover)하고 때에 따라서는 돌연변이(mutation)도 하며 다음 세대에 우수한 유전 형질이 전달(reproduction)되게 된다[3].

3.2 Neural Network

신경망(Neural Network)은 인간의 신경세포를 모형화 한 것으로써 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법이다.

[수식1] Combination Fuction

$$H_1 = f_1(b_1 + w_{11}X_1 + w_{21}X_2 + \Lambda + w_{p1}X_p)$$
$$H_2 = f_2(b_2 + w_{12}X_1 + w_{22}X_2 + \Lambda + w_{p2}X_p)$$
$$Y = g(b_0 + w_{10}H_1 + w_{20}H_2)$$

[수식1] 에서 보는 바와 같이 H는 은닉층의 은닉 노드를 가리키며 f는 활성함수, w는 연결강도를 나타내 주고 있다.

[수식2] Activation Fuction

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

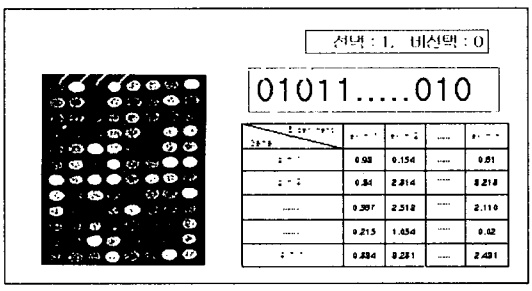
신경망은 은닉층과 은닉마디가 많으면 많을수록 추정해야 할 계수의 수가 급격히 증가하여 최적화가 어려울 수 있이며 적절한 은닉층과 은닉마디의 수를 결정하기 위해서는 여러번의 시행착오를 거쳐 최적의 값들을 분석 시스템에 저장하여 이용할 수 있다[4].

3.3. GA 와 Artificial neural networks(ANNs)

본 절에서는 GA와 Anns의 상호 연동을 하여 방대한 유전자 칩 데이터들 중에서 특징적인 유전자만을 선택할 수 있는 Gene Selection 방법에 대하여 설명 하고자 한다.

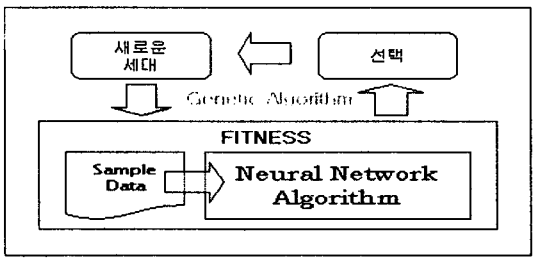
3.3.1 시스템 설계

본 논문에서는 GA와 ANNs 이용하여 Gene Selection을 하고자 한다. GA 알고리즘에서 표본과 적합도 함수의 디자인은 유전자 알고리즘의 성능을 좌우하는 매우 중요한 부분을 차지한다. 아래의 그림은 Chip이 포함하고있는 Gene 개수만큼의 길이를 갖는 이진수(선택:1, 비선택:0)로 표본을 디자인하는 그림이다.



[그림 1] 표본 디자인

아래의 그림은 DNA Chip 의 특징적인 유전자를 찾기 위한 시스템의 구조를 나타내고 있다.



[그림 2] GA - ANNs Method for gene selection

Gene Selection 의 방법은 다음과 같다. 우선, Chip 이 포함하고있는 Gene 개수만큼의 길이를 갖는 이진수로 표본을 디자인하고, 이러한 표본을 랜덤하게 많이 만들어서 하나의 세대군을 만든다.

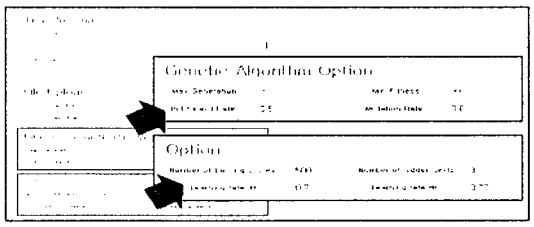
표본의 이진값에 따라서 Gene 의 선택여부를 결정하고 선택된 Gene 의 발현 정보를 이용하여 적합도 함수를 적용하여 표본을 평가하고, 만족스러운 표본이 나타날 때까지 표본들간의 교배,돌연변이 등의 방법으로 다음 세대군을 생성하고, 표본을 평가하여 만족스러운 표본을 찾으면 표본의 이진값에 따라서 Gene Selection 을 한다.

적합도 함수는 선택된 유전자의 정보만을 이용하여 ANN classification 알고리즘을 훈련시키고 분류한 결과의 정확도를 통해 결정된다. ANN Classification 알고리즘의 적합도 함수는 계산량이 많으므로 cross-validation 을 통해 정확도를 계산하지 않고, 전체 데이터를 통해 훈련시키고 분류하여, 정확도를 계산한다.

3.3.2 웹 기반 시스템 구현

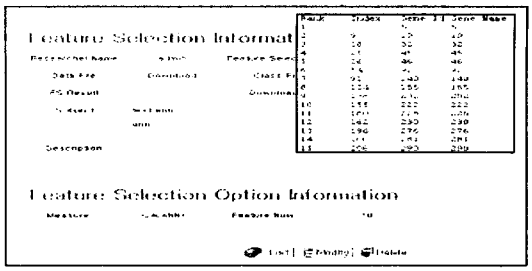
데이터마이닝 기반으로 하는 특징 유전자를 선택하는 시스템을 웹으로 구현을 하였다. 먼저 Gene Selection 을 하기 위해 다양한 옵션을 선택 할 수 있도록 구현하였다.

아래의 그림은 Gene Selection을 위한 입력 페이지이며 다양한 옵션 값들을 설정할 수 있다.



[그림 3] GA - ANNs Gene Selection

아래의 그림은 DNA Chip 의 샘플 데이터들을 입력 받아 GA-ANNs 시스템을 이용한 결과 값을 보여주고 있다.



[그림 4] GA - ANNs Result Page

4. 결론

Gene Selection 하는 방법에는 크게 Filter Method와 Wrapper Method[5] 을 이용한 여러 가지 방법들이 소개되어지고 있으며 최근에는 DNA Chip 뿐만 아니라 SNP Chip, BAC Chip등 다양한 종류의 유전자 Chip 이 개발되어지고 있는 추세이다. 바이오 기술에 대한 발전으로 대용량의 정보를 포함하고 있는 유전자 칩의 분석은 Normalization, Clustering, Classification, Text Mining, Data Mining 등 다양한 알고리즘과 분석 방법들이 개발되고 있다. 본 논문에서 제시한 진화적인 데이터마이닝 기법을 이용한 Gene Selection 방법은 유전자 분석에 있어서 많은 도움을 줄 것으로 기대되고 있다.

참 고 문 헌

[1] M. B. Eisen and P. O. Brown, " DNA arrays for analysis of gene expression," Methods Enzymbol, vol. 303, pp. 179-205, 1999.

[2] Yang, J., and Honavar, V. Feature subset selection using a genetic algorithm. Proceedings of the Genetic Programming Conference, pages 380-385. Stanford, CA, 1997

[3] Li, L., Pedersen, L.G., Darden, T.A., and Weinberg, C.R., Computational analysis of leukemia microarray expression data using the GA/KNN method, Lin, S.M. and Johnson, K.F. (eds.), Methods of Microarray Data Analysis (Proceedings of CAMDA' 00), Kluwer Academic Publishers, MA, pp. 81-95, 2002

[4] Javier Herrero1, Alfonso Valencia2, Joaquín Dopazo1*, A hierarchical unsupervised growing neural network for clustering gene expression patterns, 2000

[5] Jacek Jarmulak and Susan Craw, Genetic Algorithms for Feature Selection and Weighting, Appears in Proceedings of the IJCAI' 99 workshop on Automating the Construction of Case Based Reasoners, 1999.