2002년도 대한전기학회 하계학술대회 논문집 2002.7.10-12

신경회로망과 유전 알고리즘을 이용한 유전자 추출법과 이의 암 분류법에의 적용

조현성*, 김태선**, 전성모*, 위재우*, 이종호*
인하대학교 정보통신공학과*, 가톨릭대학교 컴퓨터.전자공학부**

Gene selection method using neural networks and genetic algorithm and its applications to classification of cancers

Cho, Hyun Sung*, Kim, Tae Seon**, Jeon, Sung Mo*, Wee, Jae Woo*, Lee, Chong Ho*
Department of Information Technology and Telecomunications, Inha Univ.*
School of Computer Science and Electronic Engineering, Catholic Univ. of Korea**

Abstract - Classification method of cancers using cDNA microarrays data was developed using genetic algorithms and neural networks. For gene selection, 2308 genes were ranked using genetic algorithms, and selected by frequency number of selection from 1000 of genetic iterative runs. To calculate fitness values, artificial neural networks are used as classifier. The small, round blue cell tumors (SRBCTs) which is difficult to distinguish via pathological single test was used as test diseases for classification, and the test results showed the 96% of exact classification capability for 25 test samples.

1. 서 론

현재 알려진 질병은 그 종류에 따라서 치료 방법이 각각 다르다. 특히 병리학적 진단에 있어 조직적으로 유사한 종류의 암에 대해서는 그 구분이 쉽지 않을 뿐 아니라 진단 시간이 많이 걸리기 때문에 빠르고 정확한 진단 방법이 필요하다. 그러한 시도 중의 하나가 cDNA Microarray data, 즉 유전자 발현 정보를 이용한 것으로 현재 세계적으로 많은 연구가 진행 중이다. 그러한 시도를 보면 먼저 유전자 발현에서 비슷한 패턴을 가지는 유전자들끼리 그룹을 짓는 데 사용한 hierarchical clustering 방법이 있었고,[1] 두 종류의 백혈병(AML/ALL)을 분류하는 데 이용된 neighborhood 분석법[2] 등 여러 통계적 방법이 있었다. 이와 비슷한 많은 통계적인 방법이 제시되었지만 암을 구별하는 데 있어 아직까지는 이러한 방법이 암을 정확하게 분류하는 능력을 가졌다고는 인정받지 못하고 있다.

현재 새로운 진단 방법 중 신경망을 이용한 분류를 많이 시도하고 있다. 특히 신경망을 이용한 패턴 분류에서 중요한 부분은 수집된 패턴 벡터 중 분류에 영향을 미치는 중요한 특징만 추출하는 부분이다. 패턴 벡터의 양이 너무 많으면 분류하는 데 연산 시간이 증가될 뿐 아니라 분류할 패턴들 간의 특징이 잘 표현되지 않아 오히려 성능이 저하되는 경우도 있다. 특히 인간의 유전자는 그 수가 너무나 많고 그 유전자 모두가 암에 관련되어 있지는 않다고 알려져 있다. 그래서 인간의 유전자 중 특정 암에 관련된 의미 있는 유전자를 찾는 방법이 중요한 것이다.[3] 지금까지 제안된 알고리즘 중에는 우리가 본 논문에 사용한 것과 동일한 데이터(SRBCT)를 이용한 것으로써 주성분분석(PCA)와 신경회로망(ANN)을 이용해서 특정 유전자를 찾아내고 그 유전자 정보를 이용해서 분류한 방법이 있었다.[4] 그리고 두 가지 백혈병(AML/ALL)을 분류하는 데 관련된 유전자를 찾는 방법에 유전 알고리즘과 k-Nearest Neighbors (KNN)을 같이 사용한 방법도 있었다.[5]

본 논문에서는 cDNA microarray data를 이용해서 암을 분류하는 데 관련된 유전자를 추출하는 방법에 유전 알고리즘과 신경망을 이용하는 방법을 제안했다. 이것은 암과 같은 패턴 분류 방법으로는 신경망을 쓰고 탐색 도구로써 유전 알고리즘을 사용한 것이다. 마지막으로 제안된 알고리즘을 이용해서 나온 실험 결과에 대해 본 실험과 같은 데이터를 PCA와 신경망을 이용해서 암 관련 유전자를 추출, 분류한 결과(96개의 유전자로 분류)와 비교하였으며, 더 적은 유전자(49개의 유전자) 정보만으로 더 나은 결과를 얻을 수 있었음을 보였다.

2. 본 론

2.1 SRBCT

SRBCT란 small, round blue-cell tumor의 준말로서 우리나라에선 아동 소원형 남색 세포 악성 종양이라고 알려진 암의 한 종류이다. 이 암은 다시 세부적으로 4가지로 분류되어지는데 그 종류는 the Ewing family of tumors (EWS), non-Hodgkin lymphoma (NHL), neuroblastoma (NB), rhabdomyosarcoma (RMS) 이다. 이렇게 분류된 4가지 각각마다 그 치료방법이 다르다. 특히 생명을 다루는 문제이기 때문에 조직적인 진단의 결과를 기다리기에는 너무나도 긴 시간이 걸릴 뿐 아니라 현미경을 통한 조직적 검사에도 어느 정도 한계가 따르고 있는 실정이다. 그래서 이 암에 대한 cDNA microarray data를 이용하여 새로운 진단 방법을 제안하려 하는 것이다. 본 실험에 사용한 SRBCT 데이터는 standard NHGRI protocol에 따른 것으로 데이터의 샘플 수는 총 88개가 이용되었으며, 그 중의 63개 샘플(23 EWS, 8 Burkitt lymphomas (BL; NHL의 일종), 12 NB, 20 RMS)은 유전자 추출을 위한 학습 데이터로 사용하고, 나머지 25개 샘플(이 중 5개 샘플은 SRBCT와 상관없는 즉, non-SRBCT 샘플이다)은 검증하는 데 사용했다. 그리고 cDNA microarray data의 최초 6567개의 유전자 발현 정보에서 발현 수치가 미약하거나 발현 정보 분석에 실패한 유전자들을 제거한 나머지 2308개의 필터링된 유전자 데이터를 가지고 실험하였다. 이 유전자 발현 수치 분석 값은 DeArray software를 사용해서 얻어진 값들이다. 그리고 이렇게 필터링된 2308개의 유전자로 구성된 88개의 샘플들의 데이터 값을 다시 0과 1사이 값으로 정규화하여 사용하였다.

2.2 암 관련 유전자 추출

유전 알고리즘을 구현하는 데 있어 가장 먼저 염색체(Chromosome) 표현 형태로는 정수 코딩(인덱스 코딩)을 사용했다. 이는 필터링된 2308개의 유전자들을 위에서부터 순서대로 1번부터 2308번까지 인덱스를 주었으며 결국 찾고자 하는 것은 관련된 유전자들이기 때문이다. 또한 염색체 길이(length)는 20으로 고정하였고, 집단 수(pop size)는 100으로 하였으며 초기 집단의 값은 1~2308까지 임의의 정수 값을 갖도록 하였다

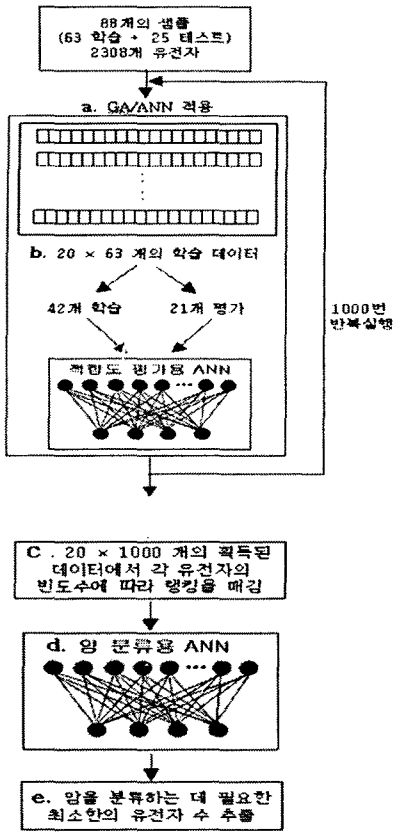

그림 1. 암 관련 유전자 추출 흐름도

(그림 1의 a 부분). 이렇게 생성된 초기 집단에서 각각의 염색체에 대하여 적합도 값을 평가함에 있어 신경망 모델을 이용하였다. 선택된 하나의 염색체(1~2308 사이의 유전자 번호를 나타내는 20개의 길이를 가진)의 정보 값(정규화된 데이터 값)을 입력 값으로 가지고 출력은 4bit으로 구현된 신경망을 구현했다. 여기서 목표 출력 값은 EWS-'1000', BL-'0100', NB-'0010', RMS-'0001' 으로 정의하였다. 그리고 학습 패턴으로는 63개의 샘플 중 임의로 선택된 42개의 샘플 데이터를 이용한다. 그리고 나서 나머지 21개의 샘플 데이터를 앞에서 학습된 신경망의 입력 값으로 넣은 다음 그 신경망을 통해서 나오는 결과 값을 목표 출력 값과 비교해서 일치하는 개수를 적합도 값으로 사용했다(그림 1의 b 부분). 여기서의 신경망은 단층 퍼셉트론으로 구현하였고, 학습 주기(epoch)는 100으로 하였으며 활성함수로는 시그모이드 함수를 사용하였다.

이렇게 나온 각 염색체의 적합도 값을 기준으로 집단을 재생산한다. 이 때 선택연산자로서 Tourment selection을 사용하여 재생산하였다. 그리고 n=1인 elitism을 적용하여 그 전에 가장 높은 적합도 값을 가지는 염색체를 다음 세대에도 유지되도록 했다. 이렇게 재생산된 집단으로부터 새로운 자손을 생성하기 위한 교배 방법으로는 Uniform crossover를 사용했다. 교배율은 0.7로 고정하였다. 그리고 공간의 미세 탐색과 국소해로부터로 탈출하는 데 기여하는 돌연변이로는 simple mutation을 사용했다. 돌연변이율은 0.01로 고정하였다. 여기서 돌연변이 시킬 때는 1~2308 사이의 값 중 임의의 값을 갖도록 하였다. 마지막으로 항상 적합도를 평가하기에 앞서 중복되는 유전자 번호가 있는 지 확인하고 있다면 중복되는 값이 없도록 그 값을 다시 임의의 값을 갖도록 하였다. 이 일련의 유전 알고리즘 과정은 염색체의 적합도 값 중에서 최고의 값(출력 값 4bit × 평가용 21개 샘플 = 84)을 가질 때까지 반복시켰다. 그리고 최대 유전 알고리즘 반복 세대수를 100으로 제한하였다. 이렇게 해서 나온 최고의 적합도 값을 가지는 염색체 하나를 얻도록 하였다. 하지만 최종적으로 나오는 하나의 염색체의 정보 즉 선택된 20개의 유전자들만을 암에 관련된 특정 유전자로 보기는 어렵다. 하지만 4가지 암을 분류하는 데 관련된 특정 유전자가 다량 선택되었다고 볼 수 있다. 그래서 하나의 염색체가 선택되는 유전 알고리즘 과정을 여러 번 반복함으로 해서 2308개의 유전자 중에서도 많이 선택되는 유전자를 찾아냄으로 해서 그 선택된 수에 따라 암을 구별하는 데 관련된 유전자로 우선순위를 주었다. 우리는 이런 과정을 1000번 반복 수행하였다. 이렇게 해서 나온 길이 20인 염색체 1000쌍이 결과적으로 나오게 된다. 이 데이터에서 각 유전자 번호가 선택된 빈도수를 그림 2에 나타내었다.

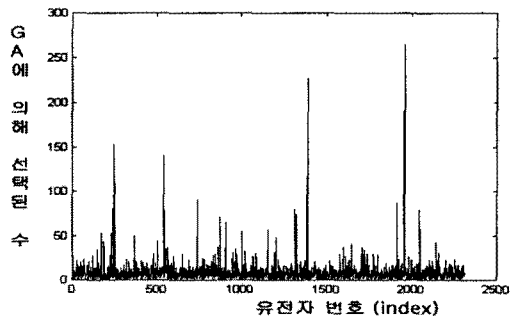

그림 2. GA/ANN에 의해 선택된 유전자 빈도수

이 빈도수를 기준으로 암 관련 유전자 랭킹을 매겼다(그림 1의 c 부분). 표 1은 같은 SRBCT 데이터로 PCA와 신경망을 이용해서 구한 암 관련 유전자 랭킹 20과 비교한 것이다. 표에서 보면 굵게 쓴 부분이 랭킹 20안에 든 유전자 중 두 방법 모두에 포함된 유전자를 나타낸다. 이렇게 서로 나타내는 랭킹이 다르긴 하지만 랭킹 100안에 드는 유전자들만 놓고 볼 때 상당 부분 같은 유전자가 선택된 것을 알 수 있었다.

표 1. GA/ANN 방법과 PCA와 신경망을 이용해서 나온 랭킹된 유전자 중 Top 20까지의 비교

	GA/ANN		PCA	
랭킹	Image Id.	Gene	Image Id.	Gene
1	784224	FGFR4	296448	IGF2
2	**770394**	FCGRT	**207274**	IGF2
3	**377461**	CAV1	841641	CCND1
4	1435862	MIC2	365826	GAS1
5	814260	FVT1	486787	CNN3
6	740801	BCKAD	**770394**	FCGRT
7	812105	AF1Q	**244618**	EST
8	898219	MSTH	233721	IGFBP2
9	**866702**	PTP	43733	GYG2
10	**244618**	EST	295985	EST
11	491565	CIT	629896	MAP1B
12	81518	apelin	840942	HLA-DPB1
13	839552	NRC1	80109	HLA-DQA1
14	814526	HSRNASEB	41591	MN1
15	796258	SGCA	**866702**	PTPN13
16	769716	NF2	357031	TNFAIP6
17	68950	cyclin E1	782503	EST
18	1473131	TLE2	**377461**	CAV1
19	143306	LSP1	52076	NOE1
20	**207274**	IGF2	811000	LGALS3BP

2.3 랭킹 유전자에 기초한 암 분류 신경망

앞에서 설명한 GA/ANN 과정을 거쳐 1번부터 2308번까지의 랭킹된 데이터를 얻었다. 여기서 나온 랭킹을 가지고는 몇 번째 랭킹된 유전자까지만 암에 관련된 유전자일지는 알 수가 없다. 그래서 랭킹된 순서대로 그 유전자 수를 늘려가면서 63개의 학습데이터로 신경망을 학습시킨 다음 나머지 25개의 테스트 데이터를 분류, 진단하였다(그림 1의 d 부분). 아래 그림 3은 랭킹된 유전자 수를 늘려가면서 각각 구현된 신경망을 통해 분류된 테스트 데이터의 결과 값과 목표치와 비교하여 일치하지 않는 개수를 나타낸 것이다. 이 개수는 최대 100까지(출력 값 4bit × 25개의 테스트 샘플 수)의 값이 나올 수 있다.

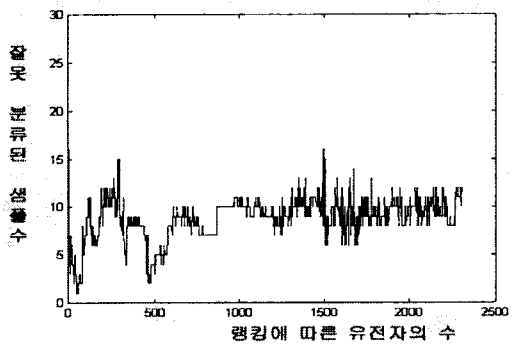

그림 3. 랭킹에 따른 유전자 수에 따라 일치하지 않은 개수 분포

이 결과 최초 랭킹 1번의 하나의 유전자만을 가지고 신경망을 구현했을 때 일치하지 않은 개수는 27이 나왔다. 그리고 점점 그 유전자 수가 늘어날수록 에러는 점점 줄어들었고, 결국 49~59번까지 랭킹된 유전자로 신경망을 구현했을 때 일치하지 않은 개수가 1로써 가장 적은 수치를 나타내었다. 하지만 그 이상 유전자 수가 늘어날 때는 다시 일치하지 않은 개수가 높아지는 것을 볼 수 있다. 이것은 곧 암을 분류하는 데 있어 모든 유전자가 관련되어 있지 않다는 것을 보이고 있는 것이다(그림 1의 c 부분).

표 2는 49번째까지 랭킹된 유전자만으로 새로 구현된 신경망을 이용해 25개의 테스트 샘플을 분류한 결과를 나타내었다. 그리고 또한 PCA와 신경망을 이용하여 랭크된 유전자 중 96번째까지의 유전자들을 이용해서 분류 진단한 것과 비교하였다. 표에서 보면 알 수 있듯이 테스트 샘플 12번을 보면 PCA 진단에서는 어떤 암인지 정확히 진단을 못한 반면 GA/ANN 진단에서는 올바르게 진단하였다. 그리고 5개의 non-SRBCT 샘플(3, 5, 9, 11, 13)들에 관해서는 어떤 암에도 진단되지 않은 결과를 보였다. 하지만 테스트 샘플 20번의 경우 둘 다 진단하지를 못한 결과가 나왔다. 단, 여기서 histological diagnosis는 조직적으로 진단된 것으로 테스트 샘플이 무슨 암에 속한 샘플인지를 나타낸 것이다.

표 2. 두 가지 방법으로 나온 분류 결과 비교

테스트 샘플	EWS	BL	NB	RMS	GA/ANN 진단	PCA 진단	histological diagnosis
1	0.00	0.00	**0.98**	0.00	NB	NB	NB-C
2	**1.00**	0.01	0.00	0.00	EWS	EWS	EWS-C
3	0.00	0.00	0.00	0.00	-	-	Osteosarcoma-C
4	0.00	0.00	0.00	**1.00**	RMS	RMS	ARMS-T
5	0.00	0.00	0.00	0.00	-	-	Sarcoma-C
6	**1.00**	0.00	0.00	0.00	EWS	EWS	EWS-T
7	0.00	**1.00**	0.00	0.00	BL	BL	BL-C
8	0.00	0.00	**1.00**	0.00	NB	NB	NB-C
9	0.00	0.01	0.00	0.15	-	-	Sk.Muscle
10	0.00	0.00	0.00	**0.99**	RMS	RMS	ERMS-T
11	0.00	0.00	0.00	0.00	-	-	Prostate Ca-C
12	**1.00**	0.00	0.00	0.00	**EWS**	-	**EWS-T**
13	0.00	0.00	0.00	0.58	-	-	Sk.Muscle
14	0.00	0.00	**0.99**	0.00	NB	NB	NB-T
15	0.00	**0.99**	0.00	0.00	BL	BL	BL-C
16	0.00	0.00	**0.99**	0.00	NB	NB	NB-T
17	0.00	0.00	0.00	**1.00**	RMS	RMS	ARMS-T
18	0.00	**0.99**	0.00	0.00	BL	BL	BL-C
19	**1.00**	0.00	0.00	0.00	EWS	EWS	EWS-T
20	0.00	0.00	0.00	0.00	-	-	EWS-T
21	**1.00**	0.00	0.00	0.00	EWS	EWS	EWS-T
22	0.00	0.00	0.00	**1.00**	RMS	RMS	ERMS-T
23	0.00	0.00	**0.94**	0.00	NB	NB	NB-T
24	0.00	0.00	0.00	**1.00**	RMS	RMS	ERMS-T
25	0.00	0.00	**1.00**	0.00	NB	NB	NB-T

3. 결 론

본 논문에서는 cDNA microarray data를 이용하여 특정 암에 관련된 유전자들을 추출하는 데 있어 유전 알고리즘과 신경망을 이용한 GA/ANN 방법을 제안하였다. 제안된 GA/ANN은 넓은 유전자 탐색 공간에서 특정 암을 분류하는 데 관련된 유전자들을 여러 세대를 거쳐 스스로 탐색하였다. 그리고 이 과정을 통해 나온 유전자 우선순위를 이용하여 보다 적은 유전자들의 집합으로 암을 분류할 수 있음을 보였다.

차후에는 본 논문에 사용한 유전 알고리즘의 연산자를 다르게 하여 적용한 결과를 비교해 볼 것이다. 또한 신경망 구현에 있어서도 여러 가지의 다층 신경망을 적용해 볼 것이다. 더 나아가 인간의 질병 뿐 아니라 다른 생물에 관련된 새로운 데이터에도 시도를 해 봄으로써 제안한 알고리즘의 적용가능성의 유무를 보일 것이다.

[참 고 문 헌]

[1] Ben-Dor. A. Shamir. R. and Yakhini. Z. "Clustering gene expression patterns", J. Comput. Biol., p.281~297, 1999.

[2] Donna K. Slonim, Pablo Tamayo, Jill P. Nesirov, Todd R. Golub and Eric S. Lander, "Class Prediction and Discovery Using Gene Expression Data", In processing of the fourth annual international conference of computational molecular biology, p. 263~272, 2000

[3] Youngjun Kwon, Jungwon Ryu, and Sung-BAE Cho, "Classification of Cancer-related Gene Expression Data Using Neural Network Classifiers", Proc. Korea Information Science Society (B), 2000.4

[4] Javed Khan, Jun S. Wei, Markus Ringner, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R. Antonescu, Carsten Peterson & Paul S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine, 7(6), p.673~679, 2001

[5] Leping Li, Clarice R. Weinberg, Thomas A. Darden and Lee G. Pedersen, "Gene selection for sample classification based on gene expression data : study of sensitivity to choice of parameters of the GA/KNN method", Bioinformatics, 17, p.1131~1142, 2001