유전발현 데이터기반 암 분류를 위한 순위기반 다중클래스 유전자 선택

홍진혁° 조성배
연세대학교 컴퓨터과학과
생체인식연구센터
hjinh@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Rank-based Multiclass Gene Selection for Cancer Classification based on Gene Expression Profiles

Jin-Hyuk Hong° Sung-Bae Cho
Dept. of Computer Science, Yonsei University
Biometrics Engineering Research Center

최근 활발히 연구가 진행 중인 유전발현 데이터를 이용한 다중클래스 암 분류는 DNA 마이크로어레이로부터 획득된 대규모의 유전자 정보를 분석하여 암의 종류를 판단한다[1]. 수집된 유전발현 데이터에는 대상 암과 관련이 없는 유전자도 포함되어 있기 때문에 높은 성능의 분류 결과를 얻기 위해서 유용한 유전자를 선택하는 것이 필요하다. 기존의 순위기반 유전자 선택은 이진클래스를 대상으로 고안되었고 이상표식 유전자(Ideal marker gene)를 이용하기 때문에 다중클래스 암 분류에 직접 적용하기에는 한계가 있다[2]. 본 논문에서는 이상표식 유전자를 사용하지 않고 유전발현 수준의 분포를 직접 분석하는 순위기반 다중클래스 유전자 선택 기법을 제안한다. 유전발현 수준을 이산화하고 학습데이터로부터 빈도를 계산하여 클래스 간 분별력을 측정한 후, 선택된 유전자를 이용하여 나이브 베이즈 분류기를 사용해 다중 암 분류를 수행한다. 제안하는 방법을 다수의 다중클래스 암 분류 데이터에 적용하여 그 유용성을 확인하였다.

본 논문에서는 다중클래스 암 분류에 적합하도록 그림 1과 같이 이산화, 빈도계산, 클래스구별력 및 영역밀집도 계산, 유전자 가치 및 순위 측정으로 구성된 유전자 선택 기법을 제안한다. 기존의 순위기반 유전자 선택 기법과 달리 제안하는 방법은 이상표식 유전자를 사용하지 않고 유전발현 수준을 직접 분석하여 유전자의 중요도를 측정한다. 먼저 유전자를 세분화하여 각 영역별 샘플의 빈도를 계산하고 클래스구별력과 영역밀집도에 따라 유전자의 중요도를 측정한다. 중요도가 높은 유전자를 선택하고 나이브 베이즈(NB) 분류기를 이용하여 다중클래스 암 분류를 수행한다.

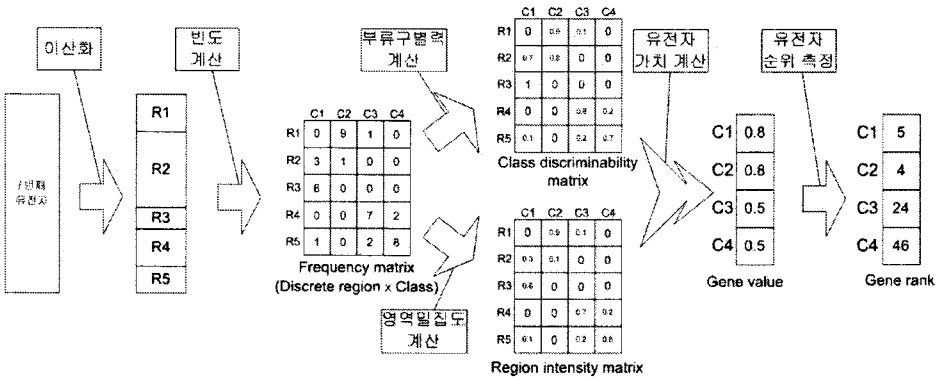

그림 1. 제안하는 유전자 선택 방법

제안하는 방법을 평가하기 위해서 대표적 다중클래스 유전발현 데이터인 GCM[3], Leukemia[4], NCI60[5]와 SRBCT[6] 데이터를 사용하였다. 이들은 적은 수의 샘플에 비해 매우 많은 수의 유전자로 구성되어 있다. 본 논문에서는 (클래스 수)×10개의 유전자를 선택하였으며, 유전발현 수준은 모두 0에서 1사이로 정규화하여 실험을 수행하였다.

표 1은 각 데이터에 대한 특징 선택 방법의 학습률을 보여준다. 아주 적은 수의 특징을 사용하여 대부분 학습 데이터를 거의 완벽하게 분류하는 NB 분류기를 획득하였다. 표 2는 테스트 데이터에 대한 분류율을 보여주는데, 제안하는 방법이 대체로 다른 특징 선택 기법에 비해 높은 분류율을 보여주었으며, 평균 79.3%의 분류율로 기존의 방법보다 높은 분류성능을 획득하였다.

표 1. 학습 데이터에 대한 분류율 (PP: 제안하는 방법)

데이터 (%)	PC	SC	ED	CC	IG	MI	SN	PP
GCM	99	97	94	98	97	99	99	97
Leukemia	98	98	100	98	97	97	100	100
NCI	100	100	100	100	100	100	100	100
SRBCT	100	100	100	100	100	100	100	100

표 2. 테스트 데이터에 대한 분류율

데이터 (%)	PC	SC	ED	CC	IG	MI	SN	PP
GCM	48	46	33	48	35	44	52	50
Leukemia	100	93	100	100	93	80	100	100
NCI	50	56	39	72	56	50	61	72
SRBCT	100	100	75	95	65	85	95	95
Avg	74.5	7.38	61.8	78.8	62.3	64.9	77	79.3

다중클래스 분류는 패턴인식에서 매우 도전적인 과제로 기존의 순위기반 특징 선택 방법을 직접 적용하기에는 한계가 있다. 본 논문에서는 이상표식 유전자를 사용하지 않고 유전자의 발현 수준을 직접 분석하는 방법을 제안하였고, 생물정보학의 대표적인 다중클래스 암 분류 데이터를 대상으로 다중클래스 암 분류에 적용하여 기존 방법보다 높은 분류 정확률을 획득하였다. 향후에는 보다 다양한 다중클래스 데이터에 적용할 것이다.

참고문헌

[1] J.-H. Hong and S.-B. Cho, "Efficient huge-scale feature selection with speciated genetic algorithm," Pattern Recognition Letter, vol. 27, no. 2, pp. 143-150, 2006.

[2] Y. Wang, F. Makedon, J. Ford and J. Pearlman, "HykGene: A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," Bioinformatics, vol. 21, no. 8, pp. 1530-1537, 2005.

[3] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander and T. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," Proc. National Academy of Science, vol. 98, no. 26, pp. 15149-15154, 2001.

[4] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den Boer, M. Minden, S. Sallan, E. Lander, T. Golub, and S. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nature Genetics, vol. 30, no. 1, pp. 41-47, 2002.

[5] D. Ross, U. Scherf, M. Eisen, C. Perou, P. Spellman, V. Iyer, S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Botstein, and P. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," Nature Genetics, vol. 24, no. 3, pp. 227-234, 2000.

[6] J. Khan, J. Wei, M. Ringnér, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," Nature Medicine, vol. 7, no. 6, pp. 673-679, 2001.