마이크로어레이 데이터를 이용한 점증적 유전자 선택기반 암 분류

권형태 홍진혁 조성배
연세대학교 컴퓨터과학과
생체인식연구센터
kwonht@naver.com, hjinh@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Incremental Gene Selection-based Cancer Classification Using Microarray Data

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

요 약

마이크로어레이 데이터는 매우 많은 수의 유전자로 구성되며, 암 분류 성능을 높이기 위해서는 대상 암과 관련된 유용한 유전자를 선택해야 한다. 기존 필터 기반 유전자 선택 기법은 유전자를 개별적으로 평가하여 암 분류에 사용하기 때문에, 유전자 사이의 관계나 분류기와의 상관성을 고려하지 않으며, 비슷한 특성의 유전자를 중복해서 선택하는 경향이 있다. 본 논문에서는 필터와 래퍼 방식을 결합하여 분류결과를 반복적으로 반영하며 유전자를 선택하는 기법을 제안한다. 필터 기법으로 유전자의 순위를 계산할 때 이전 분류에서 틀린 샘플의 가중치가 높도록 설계하고, 분류를 반복하면서 각 단계에서 유용한 유전자를 추가로 선택한다. 제안하는 방법을 대표적 암 분류 데이터인 림포마 암과 대장암 데이터에 적용하여 유용성을 검증하였다.

1. 서론

최근 마이크로어레이 데이터를 이용한 암 분류 기술이 활발히 연구되고 있다. 마이크로어레이 데이터는 보통 수천 개 이상의 유전자로 구성되지만 특정 암과 관련된 유전자는 많지 않다. 또한 샘플의 수가 매우 적기 때문에 모든 유전자를 사용한 암 분류기의 제작은 매우 어렵고 비효율적이다. 보통 암 분류에 유용한 유전자를 선택하여 관련이 없거나 중복되는 정보를 제거한다. 유전자 선택은 암 분류 성능을 높여줄 뿐만 아니라 연산량도 줄여주고 암에 대한 이해를 보다 쉽게 한다[1,2].

유용한 유전자를 선택하기 위해 표 1과 같이 매우 다양한 기술이 연구되고 있다. Lee 등과 Bae와 Mallick은 MCMC(Markov Chain Monte Carlo) 방법과 Gibbs 샘플링을 바탕으로 한 베이지안 유전자 선택 기법을 사용하였다[2,3]. Tan과 Pan은 Fisher 점수를 이용하여 유전자를 선택하였고[4], Wang 등은 Relief-F, 정보이득, x^2 통계 등을 이용하여 유전자의 가치를 측정하고 클러스터링을 통해 중복된 유전자를 제거하는 방식을 제안하였다[5]. Li 등[6]과 Li 등[7]은 유전자 알고리즘을 이용한 유전자 선택 기술을, Zhu 등[8]은 MBEGA(Markov blanket-embedded genetic algorithm)를. Banerjee 등[9]은 진화 기술과 Rough 집합 이론을 사용한 유전자 선택 기술을 제안하였다. 분류 성능에 따라 유전자를 추가하거나 제거하는 방식으로 Tang 등[10]은 SVM-RFE를, Ruiz 등[11]은 BIRS(Best Incremenral Ranked Subset) 알고리즘을 제안하였다.

본 논문에서는 이전 분류 결과를 고려한 필터 기법을 이용하여 우수한 유전자를 점증적으로 선택하는 암 분류 방법을 제안한다. 고차원 데이터에 대해 비교적 적은 연산량을 요구하는 필터 기법을 도입할 뿐만 아니라 유전자 사이의 상관관계를 고려하는 래퍼 방식을 적절히 혼용한다. 림포마 암과 대장암 데이터에 대해 제안하는 방법의 유용성을 검증하였다.

표 1. 유전자 선택 관련 최근 연구 동향

저자	유전자 선택 방법	데이터	유형
Lee 등 (2003)	베이지안 기법	루케미아	필터
Bae & Mallick (2004)	베이지안 기법	루케미아	필터
Tan & Pan (2005)	Fisher 점수	루케미아, 대장암	필터
Wang 등 (2005)	Relief-F, 정보이득, x^2 통계	루케미아, 대장암	필터
Li 등 (2001)	유전자 알고리즘	림포마, 대장암	래퍼
Li 등 (2005)	유전자 알고리즘	림포마	래퍼
Zhu 등 (2007)	MBEGA	루케미아, 대장암	래퍼
Banerjee 등 (2007)	유전자 알고리즘	루케미아, 림포마, 대장암	래퍼
Tang 등 (2007)	SVM-RFE	루케미아, 림포마, 대장암	래퍼
Ruiz 등 (2006)	BIRS	루케미아, 림포마, 대장암	래퍼

2. 배경

2.1 유전발현 데이터를 이용한 암 분류

최근에 개발된 마이크로어레이 기술은 대량의 유전발현 정보를 획득하여, 질병 등에 관련된 다양한 정보를 제공한다. 패턴 인식 분야의 많은 연구자들은 획득된 유전발현 데이터를 이용한 암 분류 기술을 개발하고 있으며, 보통 유전자 선택기술과 분류기술로 구분된다. 매우 많은 수의 유전자 정보로 구성되는 유전발현 데이터에서 특정 문제와 연관된 유전자의 수는 적기 때문에 유전자 선택이 분류 성능향상에 도움을 주며, 유전발현 데이터를 직접 분석하는 것이 매우 어렵기 때문에 신경망, 베이지안 기법, SVM, 결정트리 및 k 최근접 이웃 등의 많은 기계학습 기법이 활발히 적용되고 있다[1,6,10].

2.2 유전자 선택 기법: 필터 vs. 래퍼

암 분류를 위해 마이크로어레이 데이터로부터 유용한 유전자를 선택하는 기술은 크게 필터 방식과 래퍼 방식으로 구분된다. 필터 방식은 사전에 설계된 평가식에 따라 개별 유전자의 가치를 계산하고 높은 순위의 유전자를 선택하는 방식으로 비교적 적은 연산량에 비해 양호한 성능을 보인다. 하지만 유전자를 개별적으로 평가하기 때문에 비슷한 유전자가 중복적으로 선택되기도 하고, 유전자 사이의 관계나 분류기와의 상관성을 고려하지는 않는다. 반면에 래퍼 방식은 분류 성능을 토대로 유전자를 선택하는 방식으로 유전자 집합을 주로 평가하며 유전자 사이의 관계와 분류기와의 상관성을 충분히 고려하기 때문에 보통 필터 방식보다 높은 성능을 보인다. 하지만 많은 연산량을 요구하며 학습 데이터가 충분하지 않을 경우 분류 성능에 대한 신뢰도가 떨어진다[12].

3. 점증적 유전자 선택 기반 암 분류

본 논문에서는 그림 1과 같이 반복적으로 암 분류를 수행하면서 분류 결과를 반영하여 유전자를 선택하는 기법을 제안한다. 먼저 원본 학습 데이터(OD)로부터 필터 기반 유전자 선택 기법을 통해 고순위의 유전자를 선택한다. 선택된 유전자는 암 분류를 위한 유전자 집합(GS)에 추가되고 이들 유전자로 구성된 학습 데이터를 이용하여 분류기를 학습한다. 분석 단계에서 학습 데이터에 대한 분류 결과(TR)를 이용하여 잘못 분류된 샘플이 많이 구성되도록 원본 학습 데이터로부터 임시 학습 데이터(TD)를 구성한다. 임시 학습 데이터에 대해서 다시 유전자 선택 과정을 통해 새로운 유전자를 선택하고 이들 유전자를 유전자 집합에 추가된다. 학습 데이터에 대한 분류 성능이 더 이상 높아지지 않을 때까지 이러한 과정을 반복한다. 그림 2는 제안하는 방법의 의사 코드를 보여준다.

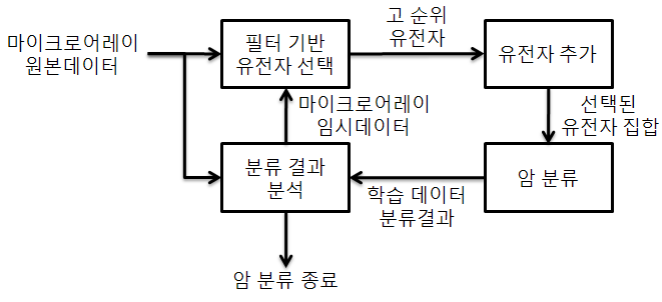

그림 1. 제안하는 방법의 동작 과정

```
FS := 특징선택함수
    (입력: 학습데이터+선택유전자개수, 출력: 유전자)
CL := 분류함수
    (입력: 학습데이터+유전자집합, 출력: 분류결과)
DA := 데이터구성함수
    (입력: 학습데이터+분류결과, 출력: 학습데이터)

IGS(OD, TD, GS, TR, a, inc, tr_r)
Initialize
    OD :={학습데이터 셋}
    TD :={∅}
    GS :={선택 유전자 셋}
    TR := {학습데이터 분류결과}
    a := 초기 유전자 선택 개수
    inc := 유전자 추가 개수
    tr_r := 학습데이터 분류 증가율
Begin
    GS := FS(OD, a)
    TR, tr_r := CL(OD, GS)
    TD := DA(OD, TR)
    while(tr_r > 0)
        GS := GS + FS(TD, inc)
        TR, tr_r := CL(OD, GS)
        TD := DA(OD, TR)
    end
End
```

그림 2. 제안하는 방법의 의사 코드

3.1 필터 기반 유전자 선택

표 2. g_i와 g_{ideal}의 유사도를 측정하기 위한 방법

$$PC(g_i, g_{ideal}) = \frac{\sum g_i g_{ideal} - \frac{\sum g_i \sum g_{ideal}}{N}}{\sqrt{\left(\sum g_i^2 - \frac{(\sum g_i)^2}{N}\right)\left(\sum g_{ideal}^2 - \frac{(\sum g_{ideal})^2}{N}\right)}}$$

$$SC(g_i, g_{ideal}) = 1 - \frac{6\sum (D_g - D_{ideal})^2}{N(N^2-1)},$$

(D_g, D_{ideal}은 g_i, g_{ideal}의 순위 행렬)

$$ED(g_i, g_{ideal}) = \sqrt{\sum (g_i - g_{ideal})^2}$$

$$CC(g_i, g_{ideal}) = \frac{\sum g_i g_{ideal}}{\sqrt{\sum g_i^2 \sum g_{ideal}^2}}$$

$$IG(g_i, c_j) = P(g_i \mid c_j) \log \frac{P(g_i \mid c_j)}{P(c_j)\cdot P(g_i)} + P(\bar{g}_i \mid c_i) \log \frac{P(\bar{g}_i \mid c_j)}{P(c_j)\cdot P(\bar{g}_i)},$$

(c_j: j번째 클래스)

$$MI(g_i, c_j) = \log \frac{P(g_i, c_j)}{P(c_j)\cdot P(g_i)}$$

$$SN(g_i) = \frac{\mu_{c1}(g_i) - \mu_{c0}(g_i)}{\sigma_{c1}(g_i) + \sigma_{c0}(g_i)}$$

순위기반 유전자 선택 기법은 사전에 정의된 이상표식 유전자와 유사한 유전자를 선택한다. 클래스 레이블 $y_i \in Y=\{0, 1\}$에 대해서 n개의 학습 샘플이 주어지면 다음과 같이 길이가 n인 두 개의 이상표식 유전자 $\{K^+, K^-\}$를 정의한다.

양의 이상표식 유전자 K^+ : $(k^1, k^2, \cdots, k^n)$

$$\begin{cases} k^i = 1, & \text{if } y_i = 1, \\ k^i = 0, & \text{if } y_i = 0. \end{cases}$$

음의 이상표식 유전자 K_j^- : $(k^1, k^2, \cdots, k^n)$ (1)

$$\begin{cases} k^i = 0, & \text{if } y_i = 1, \\ k^i = 1, & \text{if } y_i = 0. \end{cases}$$

e_j^i를 j번째 학습 샘플의 i번째 유전자의 발현 수준이라 할 때, 학습 샘플에 대해 i번째 유전자 g_i는 다음과 같이 정의된다.

$$g_i = (e_1^i, e_2^i, \cdots, e_n^i), \quad (2)$$

이상표식 유전자와 각 유전자의 유사도를 계산하여 유전자의 순위를 매긴다. 유사도는 표 2와 같이 피어슨 상관계수(PC), 스피어만 상관계수(SC), 유클리드 거리(ED), 코사인 계수(CC), 정보이득(IG), 상호정보(MI), 신호대잡음비(SN) 등의 다양한 방법을 사용하여 측정한다[1].

3.2 kNN을 이용한 암 분류

k 최근접 이웃 알고리즘(k Nearest Neighbor; kNN)은 입력된 샘플에 대해 학습 데이터에서 가장 유사한 k개의 개체를 선택하여 선택된 개체의 값으로 분류하는 알고리즘으로, 일반적인 기계 학습 기법과 달리 학습 과정이 따로 필요하지 않다[6].

테스트 샘플 te와 학습 데이터 셋 $TR = \{tr_1, tr_2, ..., tr_n\}$의 각 샘플과의 유사도는 수식 (3)과 같이 측정되며, 보통 유클리드 거리나 해밍 거리 등을 사용한다. 여기서 δ는 두 속성값의 차이이며, 두 속성이 같은 값을 가지면 1로, 다른 값을 갖는다면 0으로 정의한다.

$$\Delta(te, tr_i) = \sum_{j=1}^{m} \delta(te^j, tr_i^j), \ m\text{: 유전자 수} \quad (3)$$

학습 데이터 셋의 모든 샘플과의 유사도를 측정한 후, 가장 유사한 k개의 샘플을 선택하고 수식 (4)와 같이 가장 많이 나타난 클래스로 분류한다. 수식 (4)에서 $s_i(x)$는 x값이 i이면 1을 다른 값이면 0을 갖는다.

$$c_{te} = \arg\max_{i=0 \text{ or } 1} \left\{ \sum_{j=1}^{k} s_i(tr_j^{best}) \right\} \quad (4)$$

4. 실험 및 결과

4.1 평가 데이터

제안하는 방법을 평가하기 위해 림포종 암 데이터[13]와 대장암 데이터(http://www.sph.uth.tmc.edu:8052/hgc/default.asp)를 사용하였다. 모든 특징값은 0에서 1로 정규화하였다. 림프종 암(DLBCL)은 비홉킨스 림프종의 대표적인 질병이다. 림포마 암에는 다양한 종류가 있으며 각각 다른 치료방법이 필요하지만 임상적으로 이들을 구분하는 것은 쉽지 않다. 이 데이터 집합은 각 샘플이 4,026개의 유전 발현값으로 이루어진 47개의 샘플로 구성되어 있다. 24개 샘플은 germinal centre B-like group이고 23개 샘플은 activated B-like group이다. 대장암 데이터는 2,000개의 유전 발현값을 가지는 62개의 샘플로 구성되어 있다. 40개는 암 세포의 샘플이며, 나머지 22개는 정상 세포의 샘플로, 같은 환자의 암 부위와 정상 부위의 세포에서 채취하였다.

각 데이터 집합은 많은 특징에 비해 샘플 수는 매우 적기 때문에 10-fold cross-validation으로 평가하였다. 각 데이터 집합의 1/10은 테스트 데이터로, 나머지는 학습 데이터로 사용하며 모든 데이터가 한번 씩 테스트 데이터로 사용되도록 10회 실험한다.

4.2 결과분석

표 3과 4는 각 테스트 데이터에 대한 방법별 암 분류율을 보여준다. 모든 방법은 초기에 10개의 유전자를 선택하여 분류를 시작하였다. 림포마 암 데이터의 경우에는 유전자 수가 20개와 50개 일 때를, 대장암 데이터의 경우는 유전자 수가 50개와 100개 일 때를 각각 비교하였다. 기존 방법은 유전자 선택을 한번만 수행하였고, 제안하는 방법은 암 분류를 반복하여 2개와 4개씩 유전자를 증가시켰다. 대부분의 경우에 초기 10개의 유전자를 사용할 때보다 성능이 향상되었으며, 특히 대장암의 경우 제안하는 방법이 기존 방법에 비해 더 정확하였다. 유사도 측정 방법 면에서는 IG와 MI를 제외한 대부분의 경우 높은 분류율을 획득하였으며, PC와 SC, SN 등을 이용할 때 제안하는 방법이 보다 유용하였다. 기존 방법의 경우, 비슷한 성향의 유전자들이 선택되거나 데이터에 잡음이 있어 더 많은 유전자를 사용하더라도 분류 성능이 향상되지 않는 경우가 종종 발생하였다.

표 3. 림포마 암 데이터에 대한 테스트 분류율

특징수	방법	PC	SC	ED	CC	IG	MI	SN
10	시작	93.6	95.7	93.6	91.5	87.2	89.4	93.6
20	기존	93.6	95.7	93.6	97.9	83.0	95.7	93.6
	점증+2	95.7	95.7	95.7	97.9	85.1	91.5	95.7
	점증+4	95.7	95.7	95.7	95.7	85.1	95.7	95.7
50	기존	100	97.9	97.9	97.9	93.6	91.5	100
	점증+2	100	97.9	**100**	95.7	83.0	95.7	100
	점증+4	100	**100**	97.9	**100**	91.5	**97.9**	95.7

표 4. 대장암 데이터에 대한 테스트 분류율

특징수	방법	PC	SC	ED	CC	IG	MI	SN
10	시작	85.5	82.3	83.9	85.5	82.3	85.5	85.5
50	기존	87.1	85.5	83.9	83.9	82.3	82.3	85.5
	점증+2	91.9	88.7	83.9	87.1	80.7	83.9	**93.6**
	점증+4	91.9	**91.9**	**88.7**	82.3	79.0	**90.3**	91.9
100	기존	85.5	87.1	87.1	87.1	82.3	79.0	87.1
	점증+2	**93.6**	**91.9**	85.5	83.9	82.3	82.3	**93.6**
	점증+4	**93.6**	**91.9**	85.5	85.5	82.3	83.9	90.3

그림 3은 대장암 데이터에 대해 유전자 수에 따라 기존 방법과 제안하는 방법(2개씩 증가한 경우, 모든 유사도 측정 방법을 적용한 결과의 평균값)을 비교한 결과로, 제안하는 방법이 기존 유전자 선택 기법에 비해 3% 정도 향상되는 것을 확인하였다. 이는 유전자 선택을 분류율에 따라 점증적으로 하는 것이 이전에 선택된 유전자와 다른 정보를 가진 유전자를 선택한다는 사실을 의미한다.

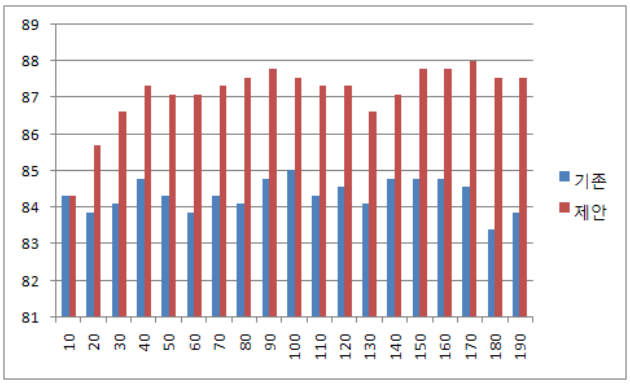

그림 3. 기존 방법과 제안하는 방법과의 성능 비교
(*x*축: 유전자 수, *y*축: 테스트 분류율)

그림 4와 5는 제안하는 방법(PC를 사용한 경우)의 유전자 추가 개수에 따른 각 데이터에 대한 테스트 분류율을 보여준다. 2개나 4개씩 유전자를 추가할 때 높은 성능향상을 보였으며, 대부분 기존 유전자 선택 방법보다 높은 분류율을 획득하였다. 다수의 유전자를 추가하는 경우에는 비슷한 특성의 유전자가 함께 선택되기 때문에 성능향상이 적은 것으로 생각된다.

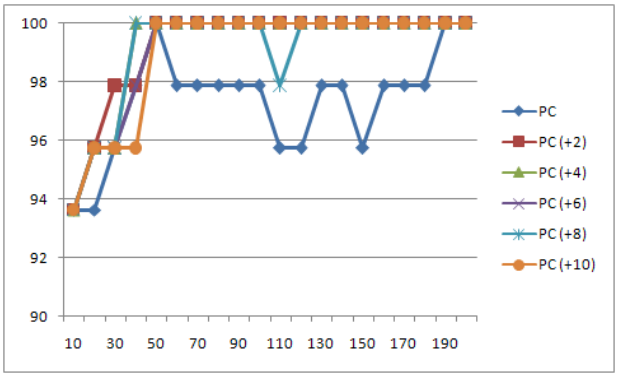

그림 4. 유전자 추가 개수에 따른 분류결과 (림포마 암)

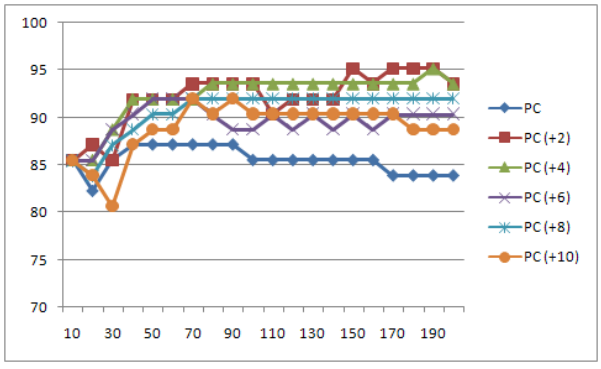

그림 5. 유전자 추가 개수에 따른 분류결과
(대장암)

림포마 암 분류의 경우, 30~50개의 유전자를 선택할 때 높은 분류 성능을 보였으며, 그 이상의 유전자는 기존에 선택된 유전자와 중복되거나 잡음으로 작용하였다. 대장암 분류에서는 70~90개 정도의 유전자 개수에서 전반적으로 높은 성능을 보였으며, 2개와 4개씩 추가할 경우에는 150~190개에서 95%의 최고 분류율을 보이기도 하였다.

5. 결론

마이크로어레이 데이터를 이용한 암 분류에서 유전자 선택은 생물정보학 분야의 매우 도전적인 과제로, 다양한 기법이연구되고 있다. 필터 기반의 유전자 선택 기법은 유전자 사이의 상관관계나 분류기와의 상관성을 고려하지 않기 때문에 비슷한 유전자가 중복되어 선택되는 등 성능 향상에 한계가 있다. 본 논문에서는 기존 필터 기법의 한계를 극복하기 위해 암 분류를 반복적으로 수행하면서 점증적으로 유전자를 선택하는 기법을 제안하였다. 필터 기법을 이용하여 유전자의 가치를 평가하지만 분류 결과를 적절히 반영하여 오분류 샘플의 분류에 유용한 유전자가 선택되도록 하였다. 제안하는 방법을 생물정보학의 대표적인 암 분류 데이터인 림포마 암 데이터와 대장암 데이터에 적용하여 기존 방법보다 높은 분류 정확율을 획득하였다. 향후에는 보다 다양한 데이터에 적용할 것이다.

감사의 글

본 연구는 생체인식연구센터(BERC)를 통해 한국과학재단(KOSEF)에서 지원받았음.

참고문헌

[1] S.-B. Cho and J. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. of the IEEE*, vol. 90, no. 11, pp. 1744-1753, 2002.

[2] K. Lee, N. Sha, E. Dougherty, M. Vannucci, and B. Mallick, "Gene selection: A Bayesian variable selection approach," *Bioinformatics*, vol. 19, no. 1, pp. 90-97, 2003.

[3] K. Bae and B. Mallick, "Gene selection using a two-level hierarchical Bayesian model," *Bioinformatics*, vol. 20, no. 18, pp. 3420-3430, 2004.

[4] A.-H. Tan and H. Pan, "Predictive neural networks for gene expression data analysis," *Neural Network*, vol. 18, no. 3, pp. 297-306, 2005.

[5] 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.

[6] L. Li, C. Weinberg, T. Darden, and L. Pedersen, "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131-1142, 2001.

[7] F. Li and Y. Yang, "Analysis of recursive gene selection approaches from microarray data," *Bioinformatics*, vol. 12, no. 19, pp. 3741-3747, 2005.

[8] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognition*, vol. 40, no. 11, pp. 3236-3248, 2007.

[9] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary rough feature selection in gene expression data," *IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 37, no. 4, pp. 622-632, 2007.

[10] Y. Tang, Y.-Q. Zhang, and Z. Huang, "Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 4, no. 3, pp. 365-381, 2007.

[11] R. Ruiz, J. Riquelme, and J. Anguilar-Ruiz, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognition*, vol. 39, no. 12, pp. 2383-2392, 2006.

[12] I. Inza, P. Larranga, R. Blanco, and A. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artificial Intelligence in Medicine*, vol. 31, no. 2, pp. 91-103, 2004.

[13] I. Lossos, A. Alizadeh, M. Eisen, W. Chan, P. Brown, D. Botstein, L. Staudt, and R. Levy, "Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas," *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 18, pp. 10209 - 10213, 2000.