편향된 DNA 칩 유전자 발현 데이터셋을 위한 표지 유전자 선택 알고리즘 개발에 관한 연구

공현중[1], 김희찬[2,3]
[1]서울대학교 대학원 협동과정 의용생체공학 전공
[2]서울대학교 의과 대학 의공학교실
[3]서울대학교 의학연구원 의공학연구소

Marker Gene Selection Algorithm for the Imbalanced DNA Microarray Gene Expression Dataset

H.J. Kong[1] and H.C. Kim[2,3]
[1] Interdisciplinary Program, Biomedical Engineering Major, Graduate School
[2] Department of Biomedical Engineering, College of Medicine and
[3]Institute of Medical & Biological Engineering, Medical Research Center,
Seoul National University, Seoul, Korea

Abstract—**We propose new algorithm for the selection of marker genes from DNA Microarray gene expression dataset, which is skewed and imbalanced. Genes are sorted according the its statistic significance and stepwisely divided into two parts at each significance ranking. And first half of sample-by-gene matrix is used as a dataset to build classifier for this special situation and find its error rate. Since this iterative process provides error rates of classifier at each ranking, we can select the specific ranking where error rate is minimal and from first to this ranking call marker genes of the gene expression data**.

Keywords—**DNA Microarray, Imbalanced data, marker gene**

I. INTRODUCTION

DNA microarray 실험을 통해 얻은 유전자 발현 데이터로 패턴 분류 문제에 직면했을 때, 분류에 영향을 미치는 maker gene 들만 선별하는 작업은 매우 중요하다. 하지만 기존의 marker gene selection 방법은 구분하려는 각 class 들의 sample 수의 비율이 비슷하지 않은 경우 maker gene 들에 대한 신뢰성이 의문시 되고 그로 인해 분류의 정확성을 떨어뜨리고 있다. 본 논문에서는 DNA microarray 를 이용한 연구 상황에서 자주 직면하게 되는 imbalanced dataset 즉, 한 class 의 샘플수가 다른 class 의 샘플수와 극명하게 차이가 나는 경우를 위한 marker gene selection algorithm 을 제안하였다.

II. METHODOLOGY

주어진 dataset 은 p 개의 gene 과 n 개의 sample, 그리고 K 개의 class 를 가지고 있다고 가정한다. 여기서 class 를 나누는 데 영향을 미치는 중요 marker gene 들만 뽑아내는 과정을 아래의 block diagram 으로 나타내었다.

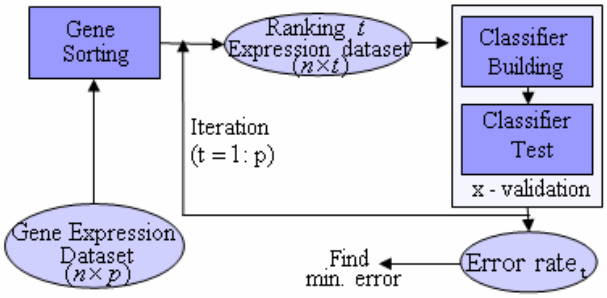

Fig. 1. Overall process of marker gene selection

A. Gene Sorting

x_{ij} : expression value of gene i for jth sample
$\bar{x}_{i.}^{k}$: mean expression value of gene i for class k
$\bar{x}_{i.}$: overall mean expression value of gene i
y_j : class label for jth sample

유전자발현(gene expression)값과 sample 의 class label 을 위와 같이 정의하기로 했을 때, i 번째 gene 에 대해 class 간 sum of squares(BSS)와 class 내 sum of squares(WSS)의 비율을 식 (1)과 같이 구할 수 있다 [1].

$$BW(i) = \frac{BSS(i)}{WSS(i)} = \frac{\sum_j \sum_k I(y_j = k)(\bar{x}_{i.}^{k} - \bar{x}_{i.})^2}{\sum_j \sum_k I(y_j = k)(\bar{x}_{i.}^{k} - \bar{x}_{i.})^2} \quad (1)$$

$(I(.)$: indicator function,
i.e. $I(\text{true}) = 1, I(\text{false}) = 0)$

BW score 클수록 class 를 구분하는 데 크게 영향을 미치는 gene 즉, marker gene 이라 할 수 있다. 이 BW

score 에 따라 각 gene 에 ranking 을 매긴 후(Fig. 2), 그 ranking 에 따라 gene-by-sample matrix 를 sorting 하였다(Fig. 2, 3).

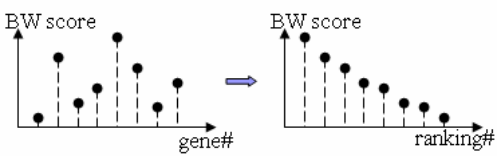

Fig. 2. 각 gene 의 BW score 와 그것의 내림차순 정렬

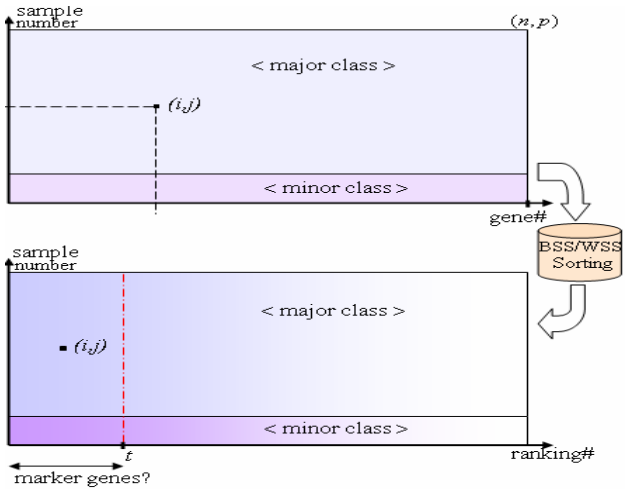

Fig. 3. Gene expression dataset before and after gene sorting according to BW score

B. Classifier & Cross-validation

앞서의 과정을 통해 BW score 에 따라 재정렬한 유전자 발현 dataset 에서 BW score 상 1 등부터 몇 등까지를 marker gene 으로 볼 것인지를 즉, 최적의 등 수, t 값을 결정해야 한다. 여기서는 Imbalanced dataset 에 특화된 classifier 를 구축하여 1 등부터 t 등까지의 dataset (n-by-t)을 잘라내어 classifier 를 구축하고 error rate 를 구하기로 하였다. 대부분의 DNA 칩 dataset 이 그렇듯이 sample 수 즉, n 값이 그리 크지 않기 때문에 cross-validation, 그 중에서도 leaved-one-out cross-validation 을 이용해 error rate 를 구하기로 하였다. n 개의 sample 중 n-1 개로 classifier 를 training 하고 나머지 1 개로 classifier 를 test 하는데, 전체 sample 수가 n 이므로 n 가지 경우 수만큼 이 과정을 수행한다.

하지만 imbalanced dataset 이라는 특성상 기존의 classifier 는 낮은 성능을 보여주기 때문에 그대로 사용하는 것은 매우 위험하다. 이에 대한 해결책으로 major class 의 sample 수를 인위적으로 줄여 minor class 의 sample 수와 비슷하게 만들어주거나, 반대로 minor class 의 sample 들의 복사본들을 만들어 역시 인위적으로 major class 의 sample 수와 비슷하게 만들어주는 방법들이 있다. 하지만 이 방법들은 data 의 분포 상황을 왜곡하거나 overfitting 을 일으키는 등의 문제가 있어 그리 신뢰를 받고 있지 않다[2].

1) *Classifier Building:* 본 논문에서는 이런 상황을 극복하기 위해 아래 그림과 같은 classifier 를 구축하였다.

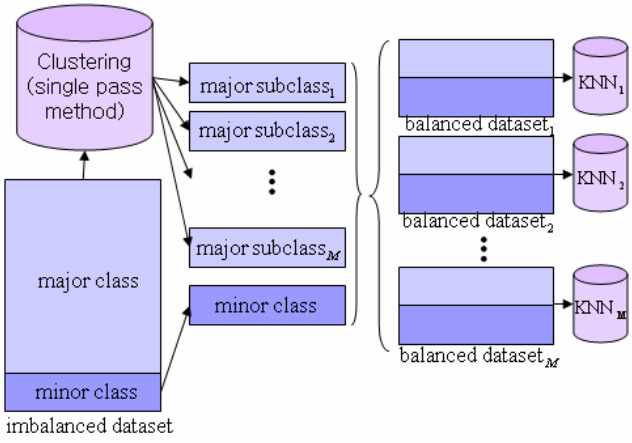

Fig. 4. Building classifier

major class 의 sample 들을 minor class 의 sample 수와 비슷한 크기를 갖는 여러 개의 subclass 들로 나누는 clustering 과정을 거친 후, major class 의 subclass 들을 각각을 minor class 와 하나의 set 로 합해서 여러 개의 balanced dataset 들을 만들어 각각의 dataset 별로 conventional 한 classifying algorithm 을 적용하게 된다.

major class 를 여러 개의 subclass 들로 나누는 clustering 기법에 대해 좀더 자세히 기술하면 다음과 같다.

1. 첫 번째 sample vector 는 첫 번째 cluster 로 할당한다.
2. k 번 째 sample vector 와 각 cluster 의 centroid vector (e.g. mean or median)들과 유사도를 측정한다.
3. 2 번의 유사도들 중 최대값을 찾아 그 값이 미리 정한 threshold 보다
    3.1. 큰 경우; k 번째 sample vector 와 최대 유사도를 보였던 cluster 에 할당한다. 그리고 나서 해당 cluster 의 멤버 sample 들의 개수가 미리 정한 기준(e.g. minor class 의 sample 들 수의 110%)보다
        3.1.1. 작은 경우; 다음 과정으로 넘어가고,
        3.1.2. 큰 겨우; 비슷한 sample 수를 갖는 두 개의 cluster 로 쪼갠다.
    3.2. 작은 경우; 해당 sample vector 를 새로운 cluster 에 할당한다.
4. major class 남아 있는 sample 들이 있으면 2 번으로 돌아간다.
5. 2~4 번까지의 과정을 cluster 들이 안정될 때까지 즉, 각 cluster 의 centroid vector 의 위치에 중요한 변화가 없을 때까지 반복 수행한다. 그리고 이렇게 해서 확정된 cluster 들을 major subclass 들이라 명명한다.

2) *Classifier Test:* 앞서 과정을 통해 형성된 classifier 에 새로운 test sample vector 를 입력하여 그것의 class label 즉, minor class 인지 major class 인지를 알아보는 과정이다. test sample vector 와 major subclass 들의 centroid 벡터들과 유사도를 측정하여 가장 큰 유사도를 보인 major subclass 와 원래의 minor class 가 합쳐져서 만들어진 balanced dataset 만이 activation 되고 그 dataset 만으로 training 된 conventional classifying 알고리즘, 여기서는 KNN 을 이용하여 class label 정보를 획득한다(Fig. 5).

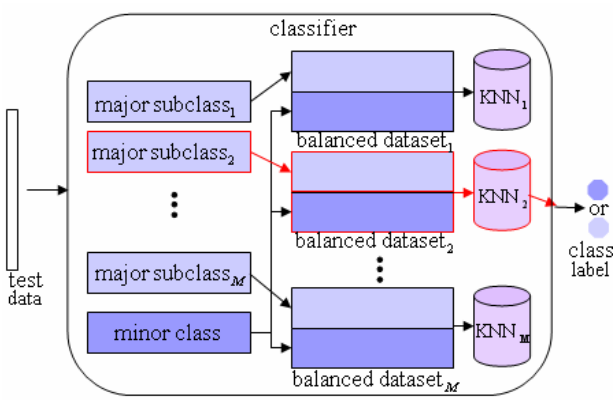

Fig. 5. Testing classifier

여기서 구축한 classifier 는 supervised learner 이기 때문에 예측이 맞았는지 틀렸는지 확인 할 수 있다. Leave-one-out cross-validation 을 수행하므로 총 n 번의 test 과정이 있고, 이를 바탕으로 classifier 의 error rate 를 계산할 수 있다.

3) *Error Rate Plotting:* 최종 marker gene set 를 정하는 최적의 t 값을 결정하기 위해, t 값을 1 에서부터 전체 유전자의 개수는인 p 까지 바꿔가면서 각각의 경우에 있어 구축된 classifier 의 error rate 를 plotting 하고 최소의 error rate 를 보이는 t 값을 찾기로 했다. (Fig. 6 참조)

III. RESULTS

사용한 DNA 칩 dataset 은 AML(급성골수성백혈병) 환자에 있어 한 달여간 기존의 프로토콜화된 화학적 항암치료를 했을 때 완전 관해(Complete emission)가 오는 class 와 그렇지 않은 class 로 이루어져 있다. 전자의 경우가 major class 이고 후자가 minor class 이다. 각각의 sample 수는 28 과 7 로 확실히 imbalanced dataset 이었다.

사용한 유사도 측정방법은 상관계수였고, classifier 의 말단에 사용할 conventional classifying 알고리즘으로 비교적 빠른 속도와 좋은 성능을 자랑하는 K-nearest neighborhood(KNN) 기법을 사용하였다. t 값을 바꿔주며 구한 error rate 는 Fig. 6 과 같다.

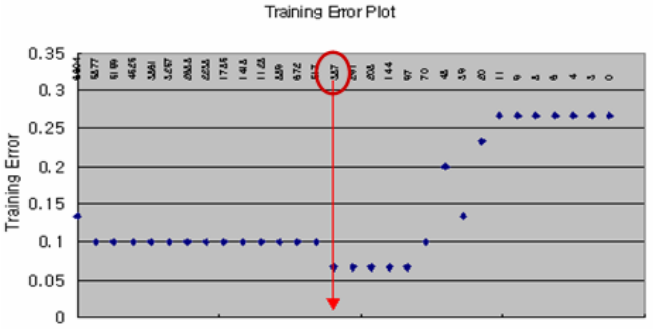

Fig. 6. Plotting error rate to find minimum error rate

Error rate 의 최소값은 0.05 였고 그때의 t 값은 97 에서 387 까지 였다.

IV. DISCUSSION

앞서 실험에서 t값이 한 개로 떨어지지 않고 넓게 나온 것은 아무래도 샘플수의 부족 때문인 것 같다. 샘플 수가 많아질수록, 비록 imbalanced dataset이라고 하더라도 좀더 나은 성능을 보이고 marker gene들에 대한 신뢰성이 향상될 것이라 생각한다. 이렇게 선정된 maker gene들로 특정 질환, 예를 들면 본 논문에서 실험한 AML 환자 진단용 DNA chip을 만들어 사용한다면, 범용 DNA 칩을 사용함으로써 오는 경제적 부담을 줄일 수 있고, 만약에 발생할지도 모르는 통계 처리의 오류에 대한 위험 부담을 줄일 수 있을 것이다.

ACKNOWLEDGMENT

This work was supported in part by Advanced Biometric Research Center (ABRC) Supported by the Korea Science and Engineering Foundation

REFERENCES

[1] Dudoit et al., “Comparison of recent classification tools applied to microarray data”, Journal of the American Association, March 2002

[2] A. S. Nugroho, S. Kuroyanagi, A. Iwata, “A Solution for Imbalanced Training Sets Problem by CombNET-II and Its Application on Fog Forecasting, IEICB Trans. INF. & Syst., vol. e85-d, no. 7, July 2002