Principal Component Analysis를 이용한 Gene Selection

임수홍 손기락 홍성용
한국외국어대학교 대학원 컴퓨터 및 정보통신학과
soohong@korea.com, ksohn@hufs.ac.kr,hongvspark@parn.com

Gene Selection using Principal Component Analysis for Molecular classification

Soo-Hong Lim, Kirack Sohn, Sung-Yong Hong
Dept. of Computer and Information Communication Engineering,
Hankuk University of Foreign Studies

요 약

수천개의 Gene Expression Measurement를 생성해 내는 DNA Microarray 연구는 조직과 세포의 표본으로부터 진단에 유용한 Gene Expression 정보를 모으게 된다. 이런 종류의 Data를 분석하기 위하여 SVM(Support Vector Machine)을 사용한 새로운 방법이 연구되어왔다. 본 논문에서는 Gene Expression Data에 대한 고유벡터(Eigen Vector)를 이용하여 SVM의 성능을 향상시키고 질병진단에 유용한 Gene을 찾아 내는 알고리즘을 기술한다. 고유벡터를 통하여 Gene을 선택적으로 SVM Learning에 참가 시키고, 분류의 결과를 통하여 추가된 Gene이 질병 진단에 미치는 영향력을 알아냄으로써 질병에 대한 Gene 역할을 파악 하는데 활용할 수 있다..

1. 서 론

일반적인 Data Mining에서 사용되는 Dataset과 DNA Microarray Expression 연구에 사용되는 Dataset의 가장 큰 차이는 객체들이 가지는 특성의 개수이다. 그림.1에서 보여지듯이 일반적인 Data Mining에서 사용되는 Dataset은 충분이 많은 객체들과 객체들이 가지는 여러 개의 특성들로 구성되지만 DNA Microarray Expression 연구에 사용되는 Dataset은 위와는 반대로 수백개의 객체들과 그 객체들이 가진 수천개의 특성들로 구성된다. Microarray Expression 실험은 동시에 수천개의 Gene의 Expression Level을 저장하게 되는데 이 Gene Expression Level의 Dataset으로부터 정보를 뽑아내기 위하여 SVM, Clustering Method, Self-Organizing Maps, Weighted Correlation Method 등 여러 방법들을 사용하여 이 Data를 분석한다.

Supervised Machine Learning technique 인 SVM은 Microarray Expression Data 분석, 단백질 동질성의 감지 등 여러 생물학 분석에 있어서 우수한 성능을 보여주고 있으며, 특히 전통적인 방법으로 증명하기 어려운 수천개의 Gene의 측정치를 포함하고 있는 고차원의 Expression Dataset의 분석에 유용하게 사용된다.

본 논문은 Microarray Expression Data의 효율적인 분석을 위하여 Principle Component Analysis를 이용하여 Data의 차원을 감소시키고, 특성에 관한 고유벡터로부터 환자(Case)를 구별하는 특성(Gene)을 선발하여 SVM Learning에 참가 시킴으로써 SVM Learning에 비협조적인 특성을 제거하여 분류의 정확성 향상과 Learning에 참가한 특성에 의미를 부여 할 수 있는 알고리즘을 제안한다.

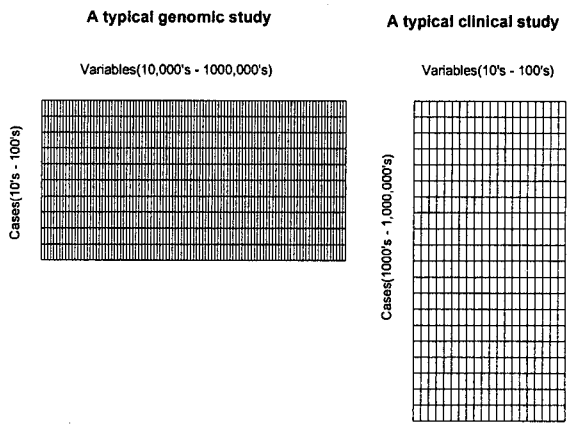

그림 1 Microarray dataset 와 일반 dataset의 차이
: High dimensionality

2. 유전자 선택 (Gene Selection)

2.1. Support Vector Machine(SVM)

Supervised Machine Learning 기술 중 하나인 SVM은 기본적으로 두 Class를 갖는 객체들을 분류하는 방법이다. 이 방법은 1976년 Vapnik에 의해 발표된 바 있으나 최근에 와서야 그 성능을 인정 받아 각광을 받게 되었으며, Vapnik(1995)에 잘 소개되어 있다.[1]

두개의 Class로 구성된 N개의 객체가 P차원 공간에 위치하는데 하나의 하이퍼플레인(Hyperplane)으로 구분되는 (separable) 경우, 두 Class사이에는 무수히 많은 하이퍼플레인이 존재 할 수 있으나 SVM에는 각 Class의 경계를 유지하는 객체(Support Vector)들을 지나는 하이퍼플레인(H1, H2)이 존재하는데 이 두 개의 하이퍼플레인과 두 Class를 구분하는 Hyperplane(SeparableHyperplane :H)간의 거리인 Margin m이 최대가 되는 Hyperplane인 H를 선택하게 된다.

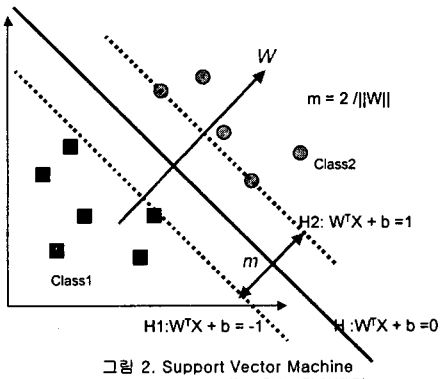

그림 2. Support Vector Machine
W: Hyperplane과 직교하는 단위벡터

두개의 class를 정확하게 분리하는 H가 존재 하지 않을 경우 오차를 인정하는 하이퍼플레인을 선택할 수 있으며 개별적인 Application에 적당한 Kernel Function을 사용하여 임의의 차원으로 객체들을 Mapping 시킨 후 그 차원에서의 Separable Hyperplane을 구하여 두 class를 분류 할 수도 있다.

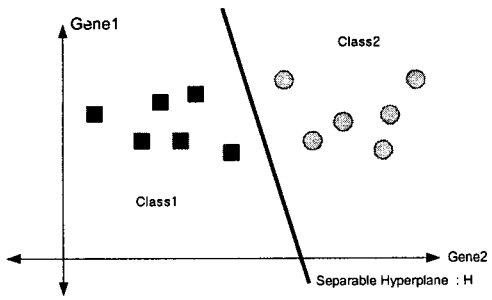

그림 3.Gene1과 Gene2의 공간에서의 객체의 분포

이 SVM을 이용하여 Microarray Data의 분석을 위하여 Learning Data의 생성 또한 중요하게 여겨지는데 예를 들어 그림[3]에서 Gene2는 두 Class를 구분할 수 있지만 Gene1은 두 Class를 구분하는 능력이 없다는 것을 알 수 있다. 만약 Gene1과 같은 Data가 SVM Learning에 참가한다면 이는 불필요한 특성의 첨가로 SVM Learning의 성능을 저하 시키는 요인이 된다.

본 논문에서는 불필요한 특성(Gene)을 제거하고, 유력한 특성을 선발하기 위한 방법으로 주성분분석 (Principal Component Analysis)을 사용하여 Learning Data를 만든다

2.2. 주성분분석 과 고유벡터

주성분분석(Principal Component Analysis)은 다차원적인 변수를 축소, 요약하는 차원의 단순화와 더불어 일반적으로 상관되어 있는 변수들 상호간의 복잡한 구조를 분석하는 것이 목적이다.

이를 위하여 주성분 분석은 변수들을 변환시켜 고유벡터(Eigenvector)라는 서로 상관되어 있지 않은 혹은 독립적인 새로운 인공 변수를 유도하는데 이 때 고유벡터가 보유하는 변이 즉 고유값(Eigenvalue)의 크기를 기준으로 그 중요도를 고려하게 된다. 일반적으로 P변량의 경우 P개의 고유값이 얻어진다. 이때 주성분의 수를 몇 개까지 선택하느냐가 문제가 되는데 이 논문에서는 공분산 행렬을 이용하였기 때문에 평균 고유값보다 큰 고유값을 갖는 고유벡터 선택하고 이를 주성분(Principal Component)이라고 하며 이 주성분이 선택되면서 고차원의 Data- set은 선택된 주성분의 개수 차원으로 변환 된다.

예를 들어 AML과 ALL의 Dataset[3][4]에서의 고유벡터와 고유값을 나타낸 그림.4에서 고유 평균값 이상의 고유벡터를 선택할 경우 Comp1~Comp7까지 선택된다.

이 분석법은 다변량 자료의 탐색적 연구, 차원축소를 통한 자료의 단순화 내지 요약, 순차적으로 독립적인 성분의 구축, 종속관계에 있는 변수들의 식별에 장점을 가지므로, 고차원 Micro-array Data분석에 유용한 분석법이다.

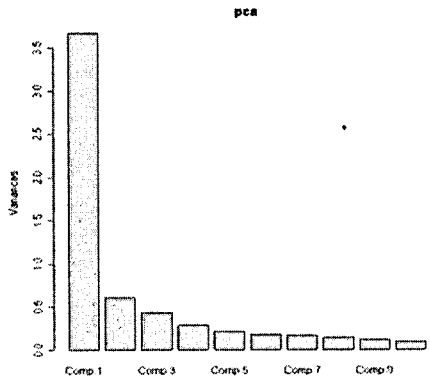

그림 4. ALL AML Microarray Dataset에서 PCA에 의해 선택된 Eigenvector의 Eignevalue의 분포

그림.4는 Golub에 의해 연구된 급성 림프구성 백혈병(Acute lymphoblastic Leukemia : ALL)과 급성 골수성 백혈병(Acute Myeloid Leukemia : AML)의 Micro-array Data에서 고유벡터와 고유값의 분포를 나타낸 그림이다.[3][4]

본 논문에서는 ALL과 AML에 대한 Dataset을 실험대상으로 주성분 분석을 통해 선택된 고유벡터로부터 각 특성에 대한 정보 즉 Variable Vector의 크기와 방향성으로 SVM Leaning에 유용한 특성을 추출하여 Learning Dataset을 만든다.

2.3. Data Set과 실험방법

본 논문에서 사용된 DataSet은 Golub의 실험에서 사용된 72명의 급성 림프구성 백혈병, 또는 급성 골수성 백혈병을 앓고 있는 환자로부터 골수 샘플과, 말초부의 혈액에서 채취한 Data Set이다. Golub의 실험에서 Data는 총 38명의 환자 (27명의 ALL 환자, 11명의 ALL 의 환자)의 Data를 가지고 Training Set을 만들고, 총 34명의 환자 (20명의 ALL 환자, 14 AML)의 Data를 가지고 Test Data set을 만들었으며, 각 DataSet은 Affymetrix사에서 만들어진 Oligonucleotide Microarrary를 사용하여 7129개의 Human Gene에 대한 Expression Level을 측정한 값이다. 측정한 값들은 각 Chip에 대한 종합적인 세기를 만들기 위하여 Rescale 한 Gene Expression Level이다.

우리는 이 DataSet의 Score를 각 Gene에 대한 Expression Level의 합을 구하고, 구해진 합으로 각 Entry를 나누는 Normalize 하였다. 실험을 위한 Gene의 선택은 50개 150개 500개를 본 논문에서 제시한 기법을 사용하여 선택하였으며, Training Set을 사용하여 Full hold-one-out cross-validation Test를 하였다.

2.4. 주성분 분석과 Variable Vector에 의한 Gene Selection

2.3.에서 설명한 Dataset을 가지고 정확한 SVM Learning을 위한 Dataset의 구성이 이 논문의 주요 내용이다.

step 1> Learning Data와 Test Data로써 각각 38명(AML 27, ALL 11)과 34명(AML 20, ALL 14)으로 환자(Case , Object)를 구분하고

step 2> 고유벡터를 구하기 위하여 본래의 Training Data의 7129개의 Gene에 대한 Covariance Matrix를 만든다.

Covariance :

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \quad (\bar{x} = \text{평균})$$

Covariance Matrix :

$$C = \begin{pmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z) \end{pmatrix}$$

step 3> Covariance Matrix에서 7129개의 고유값과 고유벡터를 구하고 고유평균값보다 큰 고유값들을 선택하고 그에 해당하는 고유벡터를 주성분으로 지정한다.

Eigenvector matrix :

$$E = \begin{pmatrix} v11 & v12 & v13 & \dots & \dots & v17129 \\ v21 & v22 & v23 & \dots & \dots & v27129 \\ v31 & v32 & v33 & \dots & \dots & v37129 \\ . & . & . & \dots & \dots & . \\ . & . & . & . & . & \end{pmatrix}$$

Eigenvector matrix는 Gene의 개수 * Gene의 개수의 크기로 만들어지며 고유평균값보다 큰 고유벡터만 남기고 나머지 벡터는 삭제된다. Eigenvector Matrix에서 주성분으로 선택된 Vector만을 남기고 만들어진 Matrix의 Column의 값으로 생성되는 벡터를 Variable Vector라고 한다.

step 4> 그림 5에서 보는 것과 같이 각 벡터들은 각각 다른 크기와 방향을 갖는다. 이를 고려하기 위하여 내적을 사용하였다. 모든 Variable Vector간의 내적을 구하고 추출하고자 하는 Data의 개수 n/2를 결정하여, 내적이 n/2 번째의 최소값을 선택하여 그에 해당한 Gene으로 Dataset을 완성한다.

벡터의 내적연산은 벡터의 크기와 두 벡터사이의 각도에 대한 연산으로, 벡터의 크기는 Principal Component공간에서 객체를 표현하는 공헌도이며, 두 벡터의 각도는 두 벡터의 상이한 정도를 나타내는 것으로 그 정도는 Cosine값과 일치한다

이로써, 양성표본과 음성표본을 구별할 수 있는 잠재적 능력을 가진 유전자 쌍을 차례로 고려할 수 있는 기법을 제시하였다

3. 결론

본 논문에서 제시한 Gene Selection 기법으로 Gene을 선택하여 수행한 분류 실험 결과는 다음 표.1,2 와 같다.

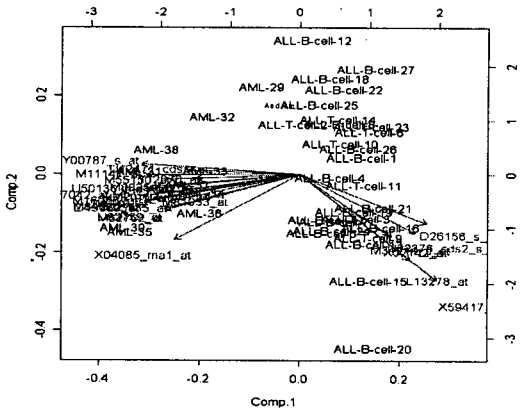

그림 5. 25개의 Gene을 선택하여 Principal Component 공간으로 변환한 객체와 Variable Vector를 표시한 예

	TRAINING RESULT	TEST RESULT
RAW DATA	FP=8 FN=1 TP=10 TN=19	FP=10 FN=3 TP=8 TN=17
50's GENE SELECTION	FP=0 FN=0 TP=11 TN=27	FP=0 FN=0 TP=11 TN=27
150's GENE SELECTION	FP=0 FN=0 TP=11 TN=27	FP=0 FN=1 TP=10 TN=27
500's GENE SELECTION	FP=0 FN=0 TP=11 TN=27	FP=1 FN=0 TP=11 TN=26

표 1 : GENE SELECTION 수행 후의 분류의 향상

	CORRECTNESS	PRECISION	RECALL
RAW DATA	68.78%	44.4%	73%
50's Gene Selection	100%	100%	100%
150's Gene Selection	99%	100%	90%
500's Gene Selection	97%	92%	100%

표 2 : CORRECTNESS PRECISION RECALL의 향상

표에서 보여지는 것과 같이 Gene Selection을 하지 않은 Raw Data의 경우 많은 오류가 발생함을 알 수 있다. 이로써 제시한 Gene Selection으로 Data의 Noise를 감소시키고, SVM을 사용하여 더욱 정확한 분류작업을 할 수 있음을 알 수 있다.

Training Result에서는 Raw Data에서는 두 개체를 정확히 분류하는 하이퍼플레인이 존재 하지 않았으나 Gene Selection을 통하여 우수한 하이퍼플레인을 찾아냈음을 알 수 있으며, 그로 인해 Test Result에서 정확한 분류가 수행되었다.

제시된 Gene Selection은 정확한 분류를 위하여 대립적인 성격의 Gene을 집단적으로 양분화하여 Learning에 참가시킨 경향이 있다. 향후에, 제시된 Gene Selection에 의해 선택된 Gene의 세분화 분류 작업을 통하여 각 Gene들의 역할 규명 연구의 범위를 대폭 줄일 수 있을 것이라 생각한다.

참고문헌

[1] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995

[2] Lindsay I Smith " A Tutorial on Principal Components Analysis" February 26, 2002

[3] Golub TR, Slonim DK, et al, " Molecular classification of cancer: class discovery and class prediction by gene expression monitoring " , Science,1999 Oct 15;286(5439):531-7.

[4] Furey TS, et .al, " Support vector machine classification and validation of cancer tissue samples using microarray expression data" , Bioinformatics, 2000 Oct;16(10):906-14.

[5]Mehmed Kantardzic, " Data Mining Concepts, Models, Methods, and Algorithms" , ISBN 0-471-22852-4.

[6] http://svmlight.joachims.org /.

[7] http://microarray.cpmc.columbia.edu/gist/ .