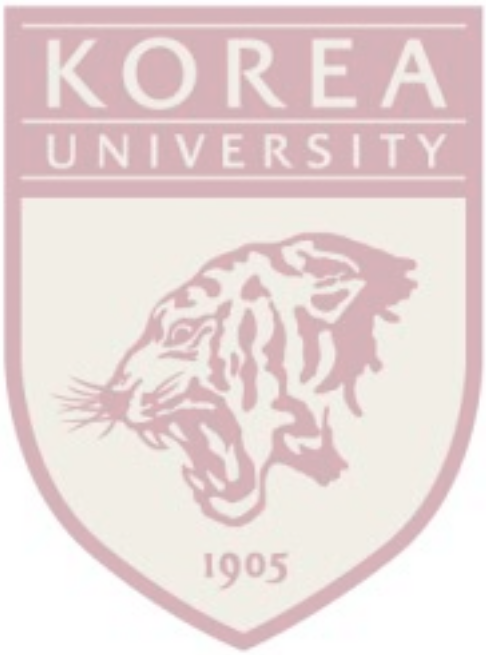
석 사 학 위 논 문

불균형 집단데이터를 위한 분류방법에 관한 비교연구



고려대학교 대학원

통 계 학 과

진 한 샘

2013년 12월 일

이 재 원 교수지도
석 사 학 위 논 문

불균형 집단데이터를 위한 분류방법에 관한 비교연구

이 논문을 통계학 석사 학위논문으로 제출함.

2013년 12월 일

고려대학교 대학원
통 계 학 과
진 한 샘

진한샘의 통계학석사 학위논문

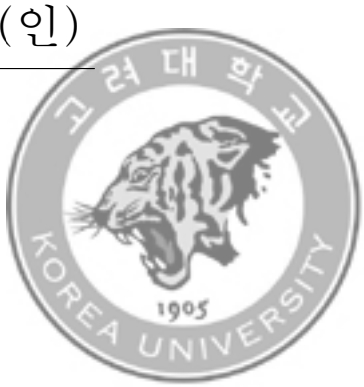
심사를 완료함.

2013년 12월 일

위원장 (인)

위 원 (인)

위 원 (인)



개 요

고차원 자료인 마이크로어레이 자료를 통한 분류는 다양한 방법을 통하여 진행된다. 특히 불균형 집단데이터에 대한 분류 방법들이 최근 주목받고 있지만, 자료의 특성을 고려한 체계적인 가이드라인이 제시되어 있지 않다. 따라서 본 연구에서는 최근에 개발되고 비교 가치가 있는 분류방법들에 집단별 분류정확도의 기하평균 방법을 적용한 후, 실제자료를 통해 분류기의 성능을 비교, 분석하였다.

주요용어: 분류분석, 불균형 집단 데이터, 고차원 자료

차 례

표 차례

제 1 장

서론

마이크로어레이 데이터 분석에서 가장 중요한 목적 중 하나는 환자의 유전자 발현 정보를 가지고 환자를 분류하는 데 있다. 마이크로어레이 데이터는 표본의 크기보다 공변량(covariate) 수가 더 큰 형태인 고차원 데이터(high-dimensional data)로 저차원 데이터(Low-dimensional data)에서 사용되는 분류기법들을 그대로 적용하는 것이 어렵고, 유전자 선택(gene selection) 과정을 요구하게 된다. 또한 대부분의 분류기가 집단 분포가 불균형인 데이터에서 성능을 보장하기 어렵기 때문에 최근 불균형 집단데이터에 대한 연구가 활발하게 진행되고 있다.

본 연구는 불균형 집단데이터에서 한계를 지닌 마이크로어레이 분류기의 성능을 비교하기 위해 다양한 실제 데이터를 여러 분류방법에 적용시켜 비교해보고 상황별로 좋은 성능을 보이는 방법을 알아보고자 한다.

비교하고자 하는 방법은 LDA에 기반을 둔 방법으로써 Nearest shrunken centroids(PAM, 2002), Classification to nearest

centroids(ClaNC, 2005), Shrunken centroids regularized discriminant analysis(SCRDA, 2007)를 고려하고, lee et al.(2005)의 비교연구에서 좋은 성능을 보였던 기계학습(Machine learning) 방법인 Supporter vector machine(SVM-Lin & SVM-Rad, 1995), 앙상블 학습(Ensemble learning) 방법인 Random forest(RF, 2001)를 고려했으며, Partial least squares discriminant analysis(PLSDA)를 고려하였다. 또한 불균형 집단데이터를 처리해 주기 위해 각 분류기의 분류 정확도의 기하평균을 이용한 soft-thresholding(2002) 유전자 선택 기법을 적용하였다.

따라서 본 연구에서는 고차원 데이터이면서 불균형 집단데이터에 여러 가지 방법을 적용하여, 앞으로의 연구에서 고차원 불균형 집단데이터에서 분류분석을 시행하는 데 유용한 정보를 주고자 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에 사용된 분류 방법들을 소개하고, 3장에서는 불균형 집단데이터를 처리해 주기 위한 방법을 소개하고. 4장에서는 방법들을 실제 데이터에 적용한 결과를 정리하며, 5장에서는 본 연구에 대한 결론 및 향후 연구 과제를 제시한다.

제 2 장
분류 방법

제 1 절 Nearest shrunken centroids(NSC)

Nearest shrunken centroids(NSC) 방법은 PAM(Tibshirani et al. 2002)이라고도 불리는 방법으로, 'Soft-thresholing'을 이용한 기법이다. 분류 결과에 대한 해석이 용이하고, 이해하기 쉬운 장점이 있다. 특히 잡음이 많은 데이터(noise data)에 적용함에 있어 잡음을 제거하는 효과를 가지는 방법이다. PAM을 수식으로 나타내면 다음과 같다.

유전자 i, 표본 j의 유전자 발현정보를 x_{ij} (i = 1,2, ... ,p , j = 1,2, ..., n), 집단 k안의 n_k개 표본들을 C_k (k = 1, 2, ..., K)라 하면,

$$\overline{x_{ik}} = \sum_{j \in C_k} \frac{x_{ij}}{n_k} \tag{2.1}$$

$$\overline{x_i} = \sum_{j=1}^{n} \frac{x_{ij}}{n} \tag{2.2}$$

여기서 식 (2.1)은 유전자 i의 집단 k안에서의 평균 발현 값이며, 식 (3.2)은 i번째 유전자의 총체적중심(overall centroid)이다.

PAM은 집단 내 표준편차(within-class standard deviation)를 이용하여 표준화하고, 이후 집단중심을 총체적중심으로 축소하는 방법을 적용한다. 집단 k의 총체적중심과 비교한 유전자 i에대한 t통계량은 다음과 같다.

$$d_{ik} = \frac{\overline{x_{ik}} - \overline{x_i}}{m_k(s_i + s_0)} \tag{2.3}$$

s_i는 유전자 i의 합동 집단내 표준편차(pooled within-class standard deviation)이며. s_0는 총체적중심과 집단중심사이의 발현값의 차이는 적으나 표준편차가 커져 유의하게 발현될 경우를 보정하기 위한 보정계수이다. s_0는 유전자 i에 대한 s_i들의 중앙값(median value)을 통해 계산한다.

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} (x_{ik} - \overline{x_{ik}})^2 \tag{2.4}$$

$$m_k = \sqrt{\frac{1}{n_k} + \frac{1}{n}}\ ) \tag{2.5}$$

식 (2.1)에서 d_{ik}를 이용하면 식 (2.6)과 같은 $\overline{x_{ik}}$이 도출되며 soft-thresholding을 이용하면 식 (2.7)과 같은 축소된 중심을 구할 수 있다.

$$\overline{x_{ik}} = \overline{x_i} + m_k(s_i + s_0)d_{ik} \tag{2.6}$$

$$\overline{x_{ik}}' = \overline{x_i} + m_k(s_i + s_0)d_{ik}' \tag{2.7}$$

각 d_{ik}는 절댓값과 △의 차이 정도에 따라 축소되고, 이 절댓값이 0보다 작으면 0으로 취급한다.

$$d_{ik}' = sign(d_{ik})(|d_{ik}| - \triangle)_+ \tag{2.8}$$

이 때 △는 교차타당(cross-validation)을 통해 CV 오분류율(CV error)을 최소화하는 값으로 추정되며, △값에 따라 선택되는 유전자의 수가 결정된다.

추정 된 △를 통해 선택 된 유전자들에서 축소된 중심을 이용한 분류 과정은 다음과 같다. 먼저 평가 표본(test sample)에 대해 식 (2.9)과 같은 판별 점수(discriminant score)를 계산한다.

$$\delta_k(x^*) = \sum_{i=1}^{p} \frac{(x_i^* - \overline{x_{ik}}')^2}{(s_i + s_0)^2} - 2\log\pi_k \tag{2.9}$$

π_k는 각 집단의 사전확률(prior probability)을 통해 구하거나 동등한 값을 부여한다. 최종적으로는 판별점수를 최소화하는 집단으로 분류된다.

$$C(x^*) = l \qquad where \;\; \delta_l(x^*) = \min_k \delta_k(x^*) \tag{2.10}$$

제 2 절 Classification to nearest centroids

Classification to nearest centroids(ClaNC) 방법은 PAM의 결점을 보완한 방법으로(Dabney, 2005) PAM과 네 가지 점이 다르다. (1) 집단중심을 축소하지 않으며, (2) 보정계수 s_0를 사용하지 않으며, (3) 집단별 특징에 맞는 유전자 선택과정을 거치며, (4) 선택된 유전

자들이 기껏해야 하나의 집단에서 유의하게 발현된다. 네 가지 점을 수정한 이유는 다음과 같다. (1) 축소(shrinkage)가 분류 정확도를 증가시킨 다는 것을 확신할 수 없고, (2) 보정계수의 사용은 t통계량의 정의에 반하는 것이며, (3) 모든 검정통계량에 동일한 △을 사용하는 것은 집단별 특징에 맞는 유전자 선택과정을 거치기 어려우며, (4) 다수의 유전자 선택은 불필요한 정보들을 포함할 가능성이 있고, 선택된 유전자들이 한 집단을 유의하게 분류한다는 것이 직관적이기 때문이다. 이 과정을 수식으로 표현하면 다음과 같다.

식 (2.3)에서 s_0를 제거한 d_{ik}값을 이용하여 식 (2.8)을 축소개념을 제거하고 집단별 특징에 맞는 △을 사용하면

$$\widetilde{d_{ik}} = sign(d_{ik})I(|d_{ik}| - \triangle_k) \qquad (2.11)$$

으로 표현되며 이 $\widetilde{d_{ik}}$를 이용하여 집단중심을 계산한다.

이 때 하나의 유전자 안에서 유의한 t통계량을 가지는 집단이 하나 이상이 될 경우에는 가장 상위의 t통계량을 가지면서 하나의 집단에서만 유의하게 발현되는 유전자를 선택하고 나머지 유전자들은 제거한다. 이후 판별과정은 PAM과 축소되지 않은 집단중심, s_0은 제거했다는 점을 제외하면 동일하다.

제 3 절 Shrunken centroids regularized discriminant analysis

Shrunken centroids regularized discriminant analysis(SCRDA)방법은 PAM 방법을 확장한 방법(Guo et al., 2007)이다. PAM 방법에서 사용하였던 축소(shrinkage) 방법과 정칙 방법(regularization)을 혼합한 방법이다. 기존 LDA에서 분류기에 사용하는 공분산행렬은 특이성(singularity)문제를 지니기 때문에 가역성(invertibility)을 가질 수 없다. 또한 고차원 데이터에서 행렬 계산과정이 복잡하기 때문에 SCRDA방법에서는 능형회귀(ridge regression)에서 사용하는 정칙 방법을 사용한다. 분류과정에서 사용하는 추정된 공분산 행렬을 $\hat{\Sigma}$라고 한다면 SCRDA는 정칙 방법을 적용한 공분산 행렬을 추정하여 사용 한다.

$$\tilde{\Sigma} = \alpha\hat{\Sigma} + (1-\alpha)I_p \tag{2.12}$$

판별점수는 축소된 집단중심을 이용하여 다음과 같이 계산된다.

$$\delta_k(x^*) = (x^* - \overline{x_k'})^T\widetilde{\Sigma^{-1}}(x^* - \overline{x_k'}) - \log\pi_k \tag{2.13}$$

이 판별점수를 이용하여 (2.10)과 같이 평가 표본의 집단을 추정한다.

SCRDA는 PAM 알고리즘에서 사용한 축소된 집단중심을 이용하여 선택된 유전자들을 통해 공분산행렬을 추정한 후 정칙 방법을 적용하는 방법론이다. 그러므로 조율 파라미터(tuning parameter)는 $\triangle$와 α 두 가지가 되는데, 조율 파라미터를 정하는 기준은 Min-Min Rule, One Standard Error Rule이 있다. 본 논문에서는

Min-Min Rule을 적용한다. Min-Min Rule에 대한 설명은 다음과 같다.

1. 먼저 훈련집단으로부터 CV 오분류율을 최소화 하는 $(\alpha, \triangle)$쌍을 찾는다.

2. 그 중, 선택되는 유전자 수를 가장 적게 하는 $(\alpha, \triangle)$쌍을 선택한다.

제 4 절 Support vector machine

Support vector machine(SVM)은 통계적인 기계학습 이론으로부터 구조적인 위험(structural risk)을 최소화하는 원리에 기반을 둔 방법(Vapnik, 1998)이다. 이 방법은 회귀분석, 분류분석 등 다양한 문제에 적용할 수 있으며 기본적인 원리는 가장 낮은 확률의 오류를 주는 가설(hypothesis)을 찾는 것에 있다. 분류분석에서는 집단데이터를 분류하는데 있어, 마진(margin)을 최대화 하는 초평면(hyperplane)을 통해 목표를 달성한다. 본 논문에서는 선형 커널 트릭을 이용한 방법(SVM-Lin)과 비선형 커널 트릭을 이용한 방법을 고려하였다. 비선형 커널 트릭을 이용한 방법 중 방사 함수(radial function)를 이용하였다(SVM-Rad). 선형 SVM을 집단이 2개인 경우에 맞추어 설명하면 다음과 같다.

학습 데이터(learning data)는 출력데이터 $y\,(y \in (1,-1))$의 값에 따라 부분집합을 구성하고 있다. 이 부분집합은 다음과 같은 초평면에 의해 구분된다.

$$w \bullet x - b = 0 \qquad (2.14)$$

이 때 w는 초평면을 이루는 하중 벡터(weight vector), b는 편향(bias)이며 다음 부등식을 만족한다.

$$w \bullet x - b \geq 1 \quad \text{for } y_i = 1 \tag{2.15}$$

$$w \bullet x - b \leq 1 \quad \text{for } y_i = -1 \tag{2.16}$$

식 (2.15)과 식 (2.16)은 다음과 동일하며

$$y_i(w \bullet x_i - b) \geq 1 \quad \text{for } all\ i \in (1,2,..,n) \tag{2.17}$$

(2.18)을 만족하는 초평면은 무수히 많고, 마진을 최대화 하는 최대 마진 초평면을 구하는 문제는 결국 최적화(optimization)문제가 된다. 식 (2.17)이라는 제약조건 하에서 최대마진 초평면을 구하는 방법은 다음과 같다.

$$\min \quad \frac{1}{2}\|w\|^2 \quad s.t. \quad y_i(w \bullet x_i - b) \geq 1 \tag{2.18}$$

그러나 (2.18)은 오차를 고려하지 않고 있다. 그러므로 오차를 고려하기 위해 양의 값을 가지는 새로운 변수 ξ_i(slack variable)를 추가하면 제약조건은 다음과 같이 변경된다.

$$y_i(w \bullet x_i - b) \geq 1 - \xi_i \quad \text{for } all\ i \in (1,2,..,n) \tag{2.19}$$

이제 이 제약조건 하에서 최대마진 초평면을 구하는 방법은 다음과 같이 변경된다.

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad s.t. \quad y_i(w \bullet x_i - b) \geq 1 - \xi_i \tag{2.20}$$

(2.20)을 통해 얻은 초평면을 이용하여 분류분석을 실시한다. 최적화 문제를 푸는 방법 중에서는 라그랑즈 승수(lagrange multiplier)를 이용한 쌍대 형태(dual form)를 통한 방법이 있는데, 방사 함수를 이용한 비선형 SVM은 쌍대 형태에서 가우시안 방사 기저함수(Gaussian radial basis function) 커널(kernel)을 이용한 것이다.

제 5 절 Random forest

Random forest(RF)는 분류 트리를 이용한(classification tree)들의 전체적인 효과를 이용한 방법으로 독립적인 임의변수선택과 bagging을 이용한 방법(Breiman, 2001)이다. Random forest의 분류 정확성은 Adaboost(Breiman, 1998)와 비슷한 성능을 지니거나 더 좋은 분류 정확도를 가진다. 또한 이상치(outlier)와 잡음(noises)이 있는 데이터에서 더욱 강건(robust)한 분류성능을 지닌다. Random forest에서 숲을 이루는 각각의 트리들은 붓스트랩 표본(boostrap sample)에서 만들어지고 각 분할에서 지원 변수들은 전체 변수들 중 임의적인 부분 집합이다. 또한 각 트리들은 여분의 변수들을 제거하지 않기 때문에 편향(bias)이 낮은 장점을 지니고 있다. 즉, 각 트리들은 상관관계(correlation)가 낮다.

각 트리에서 사용되지 않은 표본들을 OOB(out of bag)라고 하며, OOB의 오류율을 OOB 오류율(OOB error rate)라고 한다. Random forest의 유전자 선택과정은 결국 OOB오류율이 가장 작은 집합의 유전자들을 선택하는 것이다. Random forest는 평가 집단의 표본을 각각의 트리에 적용하여 집단들 중 과반수의 집단으로 투표(voting)하는 방식이다. 이러한 적용 방식 때문에 bagging을 이용한 알고리즘이라고 할 수 있다.

제 6 절 Partial least squares discriminant analysis

Partial least squares(PLS)는 독립변수와 종속변수 사이의 관계를 이용한 다변량 회귀 방법(multivariate regression method)이다. 독립변수와 종속변수는 독립변수안의 잠재변수(latent variable)를 찾기 위해 동시에 모형화 되며 독립변수는 종속변수를 예측하게 된다. PLS는 독립변수(X)와 종속변수(Y)의 공분산을 최대화 하는 선형 결합을 찾는 것이다.

$$\arg max_{b,c}\, COV(Xb,\ Yc) \tag{2.21}$$

이러한 과정은 차원축소(data reduction) 기법인 주성분분석(Principal component analysis)과 유사한 부분이 있으나, 종속변수가 독립변수와 함께 모형화가 되는 측면에서 차이가 있다. 이 때 잠재변수들은 위계적으로(hierarchically) 계산되기 때문에 마지막 잠재변수는 임의 변동(random variation)과 실험 오류(experimental error)를 의미한다. 잠재변수들이 위계적으로 계산되기 때문에 최적의 잠재변수의 수를 계산하는 것이 중요한데, 이 과정에서는 교차타당(cross validation)을 이용한다. PLS를 분류방법에 적용하면 Partial least squares discriminant(PLSDA, Barker and Ravens, 2003)가 된다. 집단이 2개인 경우 종속변수 Y는 집단 1의 경우는 1, 집단 2의 경우는 -1의 값을 지니며, 집단이 3개 이상인 경우 가변수(dummy variable)로 이루어진 행렬(matrix)형태이다.

제 3 장

기하 평균을 이용한 불균형데이터 처리

불균형 집단데이터에서는 대부분의 분류기들의 성능이 균형 집단 데이터에 비해 좋지 않은 것으로 알려져 있다. 특히 PAM과 같이 LDA에 기반을 둔 방법들은 분류과정에서 집단의 사전분포(prior distribution)가 사용되기 때문에 소수 집단(minority class)보다는 다수 집단(majority class)에 더 유리하게 분류된다. Blagus and Lusa는 이러한 문제점에 착안해 CV 오분류율을 최소화 하는 △대신 각 교차타당 과정에서 집단별 분류정확도의 기하평균(geometric mean)을 최대화하는 △를 추정하는 GM-PAM(2013)을 제안했다. 각 k개의 집단에서 집단별 분류정확도의 기하평균은 다음과 같다.

$$GM = \sqrt[k]{\prod_{k=1}^{K} PA_k} \tag{3.1}$$

분류정확도의 기하평균은 각 집단에 동일한 가중치(weight)를 주고, 평가 집단의 집단 분포와 무관하며, 각 집단별 균형정도가 불균형

일 때 벌점(penalty)을 부과하는 특징을 가지고 있다. 또한 고정된 총합($\sum_{k=1}^{K} PA_k$)에서, 완전한 균형을 이루는 데이터 일 때 최댓값을 주기 때문에, 불균형 집단데이터의 성능평가 측도로 자주 사용된다.

Blagus and Lusa가 제안한 GM알고리즘은 기하평균의 이러한 특징에 맞추어 제안된 방법으로 다음과 같은 순서로 진행된다.

1. 입력 모수 : 훈련 집단과 집단변수(X, Y), 폴드(fold)의 수(F), △의 수(T), 분류기($C(\bullet)$)
2. 훈련 집단을 F만큼 나눈다. 이 때 나누어지는 폴드의 집단 불균형 정도는 X와 근사적으로 같아야 한다.
3. $\triangle_{\max} = \min(\triangle : \overline{x_{kj}'} = \overline{x_j} \text{ for } j=1,2,...,p,\ k=1,2,...,K)$를 정의한다.
4. 각각의 $\triangle_t = \triangle_{\max} \bullet t/(T-1)$에서, CV 기하평균을 최대화 하는 △을 선택한다.

본 연구에서는 불균형 집단데이터를 보정하기 위해 GM기법을 적용하였으며 각각의 방법론에서 적용한 방식은 다음과 같다.

PAM과 ClaNC, SCRDA의 경우에는 CV과정에서 오분류율을 최소화하는 조정 모수(tuning parameter)를 선택하는 것 대신, 기하평균을 최대화 하는 조정 모수를 선택하였다. SVM, RF, PLSDA의 경우에는 사전처리 과정(pre-processing)으로 'Soft-threshold'을 이용하여 200개 이하 유전자에서 분류기에 맞는 유전자 선택(gene selection) 과정을 진행하였는데, 해당 과정에서 오분류율을 최소화 하는 유전자 선택과 기하평균을 최대화 하는 유전자 선택을 고려하였다.

제 4 장
실제 데이터 적용

제 1 절 사용된 실제 데이터

본 연구에서 사용된 방법을 비교하기 위해 집단의 개수, 불균형 정도를 고려하여 7개의 실제 마이크로어레이 데이터를 사용하였다. 각 데이터에 대한 설명은 아래와 같다.

1. Prostate data

Prostate data는 52명의 전립선 종양(prostate tumor)집단과 50명의 정상 집단으로 구성되어 있다. 6033개의 유전자 발현정보로 구성되어 있고(Dettling et al., 2002), 두 집단의 거의 유사한 균형 상태를 지니는 데이터이다.

2. Colon data

Colon data는 40명의 대장 종양(colon tumor)집단과 22명의 일반 대장(normal colon)집단으로 나누어져 있다. 6500개 이상의 유전자 발현정보로 구성되어 있고(Alon et al., 1999), 그 중 표본들 사이의 가장 높은 최소 강도(highest minimal intensity)를 가진 2000개의 유전자를 선별하여(Alon et al., 1999) 분석에 적용하였다. 다수 집단과 소수집단의 비율은 1.8:1로 약간의 불균형 상태를 지니는 데이터이다.

3. Lymphoma data

이 데이터는 The Non-Hodgkin's Lymphoma Classification Project에 의해서 제작되었다. 비호지킨림프종(non-Hodgkin lymphoma) 중 흔히 성인에게 나타나는 종류는 diffuse large B-cell lymphoma (DLBCL) 과 follicular lymphoma(FL)가 있다. Lymphoma data는 58명의 DLBCL 환자와 19명의 FL 환자로 나누어져 있다. 6817개의 유전자 발현정보로 구성되어 있고, 다수집단과 소수집단의 비율은 3.1:1로 심한 불균형 상태를 지니는 데이터이다.

4. Ginseng 1 data

대사체 데이터(metabolite data)는 고차원 데이터이며, 대사체사이의 상관관계가 높은 점 등 마이크로어레이 데이터와 매우 유사한 특징을 가지고 있다. 본 논문에서는 액체크로마토그래피(liquid chromatography) 접근 방식을 이용한 인삼 데이터를 분석에 적용하였다. 본 데이터는 4년근 인삼 10개, 5년근 10개, 6년근 인삼 10개로

30개의 표본과 240개의 선별된 대사체로 구성된 데이터이다. 각 집단의 표본 크기가 완전히 동일한 균형 상태를 지니는 데이터이다.

5. Leukemia data

Leukemia data는 3571개의 유전자 발현(gene expression) 정보와 3개의 집단(class)으로 구성되어 있다(Golub et al., 1999). 유전자 발현 정보에 따라 급성 골수성 백혈병(acute myeloid leukemia, AML)과 급성 림프성 백혈병(acute lymphoblastic leukemia, ALL)로 구분되어 있고, ALL의 경우에는 B-cell과 T-cell로 구분되어 있다. 각각의 표본은 B-cell ALL - 38, T-cell ALL - 9, AML - 25로 불균형 데이터이다.

6. Ginseng 2 data

이 데이터는 액체크로마토그래피(liquid chromatography) 접근 방식을 이용한 인삼 데이터이다. 1년근 인삼 9개, 2년근 인삼 10개, 3년근 인삼 10개, 4년근 인삼 10개, 5년근 10개, 6년근 인삼 10개로 59개의 표본과 1183개의 대사체로 구성된 데이터이다. 각 집단의 표본 크기가 거의 동일한 균형 상태를 지니는 데이터이다.

7. NCI60 data

이 데이터는 The National Cancer Institute's anti-cancer drug screen project에 의해 제작되었다. 전체 데이터는 60명의 환자와 9709의 유전자를 지니며, 다양한 종양 조직을 지닌 표본으로 구성되어 있다(7 breast, 5 central nervous system(CNS), 7 colon, 6 leukemia, 8 melanoma, 9 non small cell lung carcinoma(NSCLC),

6 ovarian, 2 prostate, 9 renal and 1 unknown.). 어떤 집단에서는 표본크기가 분석을 실시하기에 너무 적어, 1375개의 선별된 유전자와 6개의 집단데이터로 축소된 데이터(Ross et al., 2000)를 이용하였다. 6개의 집단은 위계적 군집분석에 의해 정의되었고(Scherf et al., 2000), 각 집단의 표본의 크기는 8, 14, 9, 11, 10, 8명이다. 집단별로 표본의 크기가 다른 불균형데이터이다.

제 2 절 평가 측도

불균형 집단데이터에서 분류 예측 성능을 평가하는 방법은 두 가지로 생각할 수 있다. 첫 번째는 임의의 환자에 대한 올바른 집단군으로 판별하였는가를 측정하는 방법이고, 두 번째는 다수 집단(majority class)과 소수 집단(minority class)의 분류정확도를 비교할 때 다수 집단에 유리한 방향으로 분류가 예측되지 않았는지에 대해 측정하는 방법이다. 이 두 방법을 기반으로 오분류율(error rate)과 각 집단별 분류정확도, 각 집단별 분류정확도의 기하평균을 성능평가 측도로 사용하였다.

먼저 오분류율은 분류분석의 성능평가 측도로 가장 빈번하게 사용되는 개념으로 평가 집단 중 잘못 분류된 표본의 개수를 전체 평가 표본의 개수로 나눈 개념이다. 즉 전체 평가 집단 중에서 잘못 분류된 표본의 비율을 구하는 개념이다. 오분류율은 일반적으로 분류분석의 성능을 평가하기에 적합한 도구인 것은 사실이나 불균형 집단데이터에서는 오분류율을 그대로 신뢰하기는 어렵다. 만약 불균형 비율이 9:1이라고 가정한다면 평가집단을 모두 다수 집단으로 분류

할 때 오분류율은 항상 0.1이 되기 때문이다. 그러므로 다수 집단에 치우치지 않게 소수 집단을 잘 분류했다는 평가측도가 필요한데 이는 각 집단별 분류정확도의 기하평균을 이용할 수 있다. 각 집단별 분류정확도의 기하평균은 식 (3.1)과 같으며, 본 연구에서는 기하평균 이외에도 소수집단과 다수집단의 분류정확도를 정확히 보기 위해 각 집단별 분류정확도도 평가측도로 사용하였다.

제 3 절 분석 결과

본 논문의 분석과정은 다음과 같다. 각 데이터를 7:3의 비율로 훈련집합(training set)과 평가집합(test set)으로 나누고 각 분류 방법에 적용해 분석하였다. 이러한 과정을 50번씩 반복하였다.
방법들을 적용한 결과는 <표 4.1>부터 <표 4.7>까지 제시되어 있다. 각각의 데이터에서 가장 우수한 성능을 보였던 분류 기법을 나열하면 다음과 같다.

- Prostate data : PLSDA, GM-PLSDA
  (오분류율 4%, 기하평균 : 96%)
- Colon data : SVM-Rad, GM-SVM-Rad
  (오분류율 9%, 기하평균 : 90%)
- Lymphoma data : SVM-Lin, GM-SVM-Lin, PLSDA, GM-PLSDA
  (오분류율 3%, 기하평균 : 96%)
- Ginseng 1 data : SCRDA, GM-SCRDA, SVM-Lin, GM-SVM-Lin, SVM-Rad, GM-SVM-Rad,

PLSDA, GM-PLSDA

(오분류율 0%, 기하평균 : 100%)

- Leukemia data : SVM-Rad, GM-SVM-Rad

(오분류율 2%, 기하평균 : 99%)

- Ginseng 2 data : GM-SCRDA(오분류율 2%, 기하평균 : 99%)
- NCI60 data : GM-RF(오분류율 6%, 기하평균 : 91%)

실제 데이터 분석을 통해 얻은 주요 발견사항은 다음과 같다.

PAM, ClaNC, SCRDA와 같이 LDA 기반 분류기들은 균형 집단데이터에서는 다른 분류기에 비해 성능이 크게 떨어지지 않지만 불균형 집단데이터에스는 성능이 떨어진다. 이는 LDA에서는 알고리즘에서 집단의 사전분포(prior distribution)가 요구되기 때문이다. 즉, 새로운 데이터가 분류기에 입력될 때 다수 집단으로 분류예측 될 가능성이 증가함을 의미한다. 세 분류기는 일반적으로 우수하지 못한 성능을 보여줬지만 SCRDA는 세 분류기 중 가장 우수한 성능을 보였다. 이는 PAM, ClaNC의 경우 유전자 사이의 상관관계를 고려하지 않는 모형이나 SCRDA가 유전자 사이의 상관관계를 고려하는 모형이기 때문이라 판단된다.

SVM기반 방법과, PLSDA는 대체적으로 모든 데이터에서 좋은 결과를 보였다. Lee et al.(2005)의 연구에서는 대부분의 데이터에서 SVM기반 방법들이 가장 좋은 성능을 보여줬다. 불균형 집단데이터를 연구한 Lin and Chen(2012)의 연구에서도 대부분의 Data가 SVM기반 방법이 가장 좋은 성능을 보였다. 분석결과 PLSDA도 SVM기반 방법과 비교할 만한 우수한 성능을 보이고 있다.

또한 기하평균에 기반을 둔 방법이 오분류율에 기반을 둔 방법보다

동일하거나 우수한 성능을 보이는 것을 알 수 있었다. 특히 불균형 상태가 심해질수록, 집단의 개수가 많아질수록 그 효과가 두드러짐을 알 수 있다. 2개의 집단군을 가진 데이터 중 가장 불균형 상태가 심각했던 Lymphoma data에서는 PAM, ClaNC, SCRDA적 기법보다 GM방식을 사용한 것이 각각 Geometric means를 3%, 1%, 3%를 상승시켰다. 또한 집단의 개수가 많은 Ginseng 2 자료와 NCI60 자료에서 가장 우수한 성능을 보인 분류 방법은 모두 기하평균에 기반을 둔 방법이었다.

기하평균에 기반을 둔 방법은 Embedded하게 유전자를 선택하는 분류기인 PAM, ClaNC, SCRDA에서 큰 효과를 보였다. SVM, RF, PLSDA에서는 큰 효과를 보이지 못했지만 RF, SVM의 경우에는 기하평균을 사용한 방법이 결점을 보완한 구간이 존재했다. 그 효과는 class의 개수가 많아질수록 더 크게 작용했다. 불균형 상태가 심한 Lymphoma데이터에서 GM-RF가 RF에 비해 분류정확도의 기하평균을 1% 상승시켰다.

기하평균을 사용한 방법의 원리를 알아보기 위해 2개의 집단군 데이터를 자세히 살펴보면 기하평균에 기반을 둔 방법론은 기존방법들 보다 다수 집단의 분류정확도를 약간 감소시키면서 소수 집단의 분류정확도를 상승시키는 효과를 보였다. 그러므로 새로운 데이터를 다수 집단으로 분류할 가능성이 큰 분류기에는 기하평균을 이용한 방법을 반드시 고려해야 할 것이다.

<표 4.1> Prostate data 분석 결과

방법	Tumor PA	Normal PA	geometric means	test error
PAM	0.92/0.07	0.90/0.06	0.91/0.05	0.09/0.05
GM-PAM	0.93/0.05	0.90/0.06	0.91/0.04	0.08/0.04
ClaNC	0.92/0.09	0.90/0.07	0.91/0.05	0.09/0.05
GM-ClaNC	0.93/0.08	0.90/0.08	0.91/0.05	0.08/0.04
SCRDA	0.92/0.10	0.92/0.06	0.92/0.06	0.08/0.05
GM-SCRDA	0.95/0.10	0.92/0.06	0.94/0.04	0.06/0.04
SVM-Lin	0.97/0.04	0.94/0.06	0.95/0.04	0.05/0.04
GM-SVM-Lin	0.97/0.04	0.94/0.06	0.95/0.04	0.05/0.04
SVM-Rad	0.97/0.04	0.91/0.07	0.94/0.04	0.06/0.04
GM-SVM-Rad	0.97/0.04	0.91/0.06	0.94/0.04	0.06/0.04
RF	0.97/0.05	0.91/0.06	0.94/0.04	0.06/0.04
GM-RF	0.97/0.05	0.91/0.06	0.94/0.04	0.06/0.04
PLSDA	0.97/0.04	0.94/0.06	0.96/0.03	0.04/0.03
GM-PLSDA	0.97/0.03	0.94/0.06	0.96/0.03	0.04/0.03

* 평균/표준편차

<표 4.2> Colon data 분석 결과

방법	Tumor PA	Normal PA	geometric means	test error
PAM	0.91/0.07	0.81/0.16	0.84/0.09	0.13 /0.07
GM-PAM	0.87/0.08	0.86/0.15	0.86/0.09	0.13/0.07
ClaNC	0.92/0.07	0.76/0.18	0.83/0.11	014/0.07
GM-ClaNC	0.89/0.08	0.85/0.15	0.86/0.09	0.13/0.07
SCRDA	0.90/0.08	0.80/0.15	0.84/0.08	0.14/0.07
GM-SCRDA	0.88/0.09	0.82/0.17	0.84/0.10	0.14/0.08
SVM-Lin	0.94/0.06	0.86/0.14	0.89/0.07	0.09/0.05
GM-SVM-Lin	0.92/0.07	0.87/0.13	0.89/0.07	0.09/0.05
SVM-Rad	0.93/0.06	0.88/0.12	0.90/0.07	0.09/0.06
GM-SVM-Rad	0.92/0.06	0.89/0.12	0.90/0.07	0.09/0.06
RF	0.92/0.07	0.84/0.15	0.87/0.08	0.11/0.06
GM-RF	0.90/0.07	0.86/0.15	0.87/0.08	0.11/0.06
PLSDA	0.93/0.06	0.83/0.14	0.87/0.08	0.09/0.06
GM-PLSDA	0.92/0.07	0.84/0.14	0.87/0.08	0.09/0.06

* 평균/표준편차

<표 4.3> Lymphoma data 분석 결과

방법	DLBCL PA	FL PA	geometric means	test error
PAM	0.85/0.09	0.88/0.19	0.86/0.12	0.14/0.08
GM-PAM	0.82/0.09	0.97/0.09	0.89/0.07	0.14/0.07
ClaNC	0.86/0.09	0.76/0.24	0.80/0.15	0.16/0.08
GM-ClaNC	0.86/0.10	0.78/0.19	0.81/0.12	0.16/0.09
SCRDA	0.95/0.05	0.86/0.18	0.90/0.11	0.07/0.05
GM-SCRDA	0.95/0.05	0.91/0.12	0.93/0.06	0.06/0.04
SVM-Lin	0.98/0.03	0.94/0.09	0.96/0.04	0.03/0.03
GM-SVM-Lin	0.97/0.04	0.95/0.08	0.96/0.04	0.03/0.03
SVM-Rad	0.96/0.05	0.91/0.10	0.93/0.06	0.06/0.04
GM-SVM-Rad	0.95/0.06	0.92/0.10	0.93/0.05	0.06/0.05
RF	0.94/0.07	0.80/0.19	0.86/0.13	0.10/0.08
GM-RF	0.93/0.07	0.81/0.18	0.87/0.11	0.10/0.08
PLSDA	0.97/0.04	0.95/0.09	0.96/0.05	0.03/0.03
GM-PLSDA	0.97/0.04	0.95/0.08	0.96/0.04	0.03/0.04

* 평균/표준편차

<표 4.4> Ginseng 1 data 분석 결과

방법	1년산 PA	2년산 PA	3년산 PA	geometric means	test error
PAM	1.00/0.00	0.90/0.20	0.99/0.09	0.94/0.16	0.04/0.07
GM-PAM	1.00/0.00	0.91/0.20	0.99/0.09	0.94/0.16	0.04/0.07
ClaNC	0.96/0.15	0.90/0.23	0.94/0.19	0.89/0.22	0.07/0.11
GM-ClaNC	0.97/0.12	0.92/0.22	0.95/0.17	0.91/0.21	0.06/0.09
SCRDA	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00
GM-SCRDA	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00
SVM-Lin	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00
GM-SVM-Lin	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00
SVM-Rad	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00
GM-SVM-Rad	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00
RF	0.95/0.15	0.98/0.10	0.99/0.05	0.97/0.09	0.02/0.08
GM-RF	0.97/0.12	0.98/0.10	0.98/0.10	0.97/0.08	0.02/0.06
PLSDA	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00
GM-PLSDA	1.00/0.00	1.00/0.00	1.00/0.00	1.00/0.00	0.00/0.00

* 평균/표준편차

<표 4.5> Leukemia data 분석 결과

방법	B-cell PA	T-cell PA	AML PA	geometric means	test error
PAM	0.97/0.04	0.97/0.10	0.95/0.10	0.96/0.05	0.04/0.04
GM-PAM	0.97/0.05	0.97/0.10	0.97/0.07	0.96/0.04	0.03/0.03
ClaNC	0.97/0.05	0.88/0.25	0.95/0.09	0.90/0.20	0.05/0.05
GM-ClaNC	0.96/0.05	0.92/0.17	0.96/0.08	0.94/0.07	0.04/0.04
SCRDA	0.96/0.06	0.82/0.25	0.91/0.12	0.88/0.12	0.07/0.06
GM-SCRDA	0.96/0.06	0.90/0.17	0.95/0.08	0.93/0.08	0.05/0.05
SVM-Lin	0.97/0.04	0.99/0.07	0.98/0.05	0.98/0.03	0.02/0.02
GM-SVM-Lin	0.97/0.04	0.99/0.07	0.98/0.05	0.98/0.03	0.02/0.02
SVM-Rad	0.98/0.04	1/0.00	0.98/0.05	0.99/0.02	0.02/0.02
GM-SVM-Rad	0.98/0.04	1/0.00	0.98/0.05	0.99/0.02	0.02/0.02
RF	0.97/0.05	0.93/0.14	0.99/0.04	0.96/0.06	0.03/0.03
GM-RF	0.97/0.04	0.93/0.14	0.99/0.05	0.96/0.06	0.03/0.03
PLSDA	0.97/0.04	0.97/0.10	0.98/0.05	0.97/0.05	0.03/0.03
GM-PLSDA	0.97/0.04	0.97/0.10	0.98/0.05	0.97/0.05	0.03/0.03

* 평균/표준편차

<표 4.6> Ginseng 2 data 분석 결과

방법	1년산 PA	2년산 PA	3년산 PA	4년산 PA	5년산 PA	6년산 PA	geometric means	test error
PAM	1.00/0.00	0.98/0.08	1.00/0.00	0.99/0.05	0.93/0.16	0.87/0.21	0.95/0.06	0.04/0.04
GM-PAM	1.00/0.00	0.97/0.09	1.00/0.00	0.99/0.07	0.95/0.13	0.89/0.20	0.96/0.05	0.03/0.04
ClaNC	0.98/0.08	0.96/0.11	0.98/0.10	0.91/0.16	0.89/0.17	0.97/0.07	0.93/0.06	0.05/0.06
GM-ClaNC	0.97/0.09	0.97/0.10	0.99/0.09	0.9/0.17	0.88/0.18	0.99/0.06	0.93/0.06	0.05/0.05
SCRDA	1.00/0.00	0.96/0.11	0.92/0.20	0.85/0.20	0.98/0.08	0.99/0.05	0.97/0.15	0.05/0.07
GM-SCRDA	1.00/0.00	1.00/0.00	0.97/0.11	1.00/0.00	1.00/0.00	0.97/0.11	0.99/0.05	0.02/0.05
SVM-Lin	1.00/0.00	1.00/0.00	0.99/0.07	0.96/0.11	0.95/0.12	0.97/0.09	0.98/0.03	0.02/0.03
GM-SVM-Lin	1.00/0.00	1.00/0.00	0.99/0.07	0.96/0.11	0.95/0.12	0.97/0.09	0.98/0.03	0.02/0.03
SVM-Rad	1.00/0.00	0.99/0.05	0.99/0.07	0.85/0.22	0.97/0.12	0.97/0.10	0.95/0.06	0.04/0.04
GM-SVM-Rad	1.00/0.00	0.99/0.05	0.98/0.08	0.87/0.19	0.96/0.13	0.96/0.11	0.95/0.05	0.04/0.04
RF	1.00/0.00	1.00/0.00	1.00/0.00	0.90/0.19	0.89/0.16	0.98/0.08	0.95/0.05	0.03/0.04
GM-RF	1.00/0.00	1.00/0.00	1.00/0.00	0.90/0.18	0.89/0.15	0.98/0.08	0.95/0.04	0.03/0.04
PLSDA	0.99/0.07	0.99/0.05	0.97/0.09	0.95/0.12	0.89/0.19	0.99/0.97	0.95/0.05	0.02/0.03
GM-PLSDA	0.99/0.07	0.99/0.05	0.97/0.10	0.95/0.12	0.90/0.17	0.98/0.08	0.96/0.05	0.02/0.03

* 평균/표준편차

<표 4.7> NCI60 data 분석 결과

방법	1집단 PA	2집단 PA	3집단 PA	4집단 PA	5집단 PA	6집단 PA	geometric means	test error
PAM	0.86/0.27	0.92/0.11	1.00/0.00	0.68/0.25	0.93/0.13	0.99/0.09	0.83/0.23	0.11/0.07
GM-PAM	0.86/0.28	0.93/0.11	1.00/0.00	0.73/0.28	0.97/0.10	0.99/0.09	0.83/0.26	0.10/0.08
ClaNC	0.77/0.32	0.92/0.14	1.00/0.00	0.82/0.20	0.88/0.19	0.96/0.14	0.81/0.26	0.11/0.08
GM-ClaNC	0.76/0.31	0.92/0.13	0.99/0.05	0.84/0.21	0.93/0.15	0.99/0.09	0.84/0.23	0.10/0.07
SCRDA	0.80/0.32	1.00/0.00	1.00/0.00	0.72/0.22	0.97/0.10	0.99/0.05	0.83/0.26	0.09/0.06
GM-SCRDA	0.82/0.30	0.98/0.07	1.00/0.00	0.72/0.22	0.97/0.10	0.99/0.05	0.85/0.23	0.08/0.06
SVM-Lin	0.84/0.30	0.94/0.11	1.00/0.00	0.90/0.15	0.95/0.12	0.99/0.05	0.88/0.23	0.06/0.06
GM-SVM-Lin	0.87/0.25	0.91/0.14	1.00/0.00	0.87/0.22	0.95/0.14	0.99/0.05	0.91/0.15	0.07/0.09
SVM-Rad	0.78/0.31	0.96/0.10	1.00/0.00	0.90/0.15	0.95/0.12	0.99/0.05	0.86/0.26	0.07/0.06
GM-SVM-Rad	0.85/0.25	0.94/0.12	1.00/0.00	0.86/0.21	0.92/0.17	0.98/0.08	0.90/0.15	0.08/0.09
RF	0.86/0.25	0.96/0.09	1.00/0.00	0.87/0.17	0.95/0.12	0.99/0.07	0.90/0.19	0.06/0.06
GM-RF	0.87/0.23	0.96/0.09	1.00/0.00	0.85/0.18	0.95/0.12	1.00/0.00	0.91/0.15	0.06/0.06
PLSDA	0.89/0.25	0.92/0.14	1.00/0.00	0.84/0.20	0.92/0.14	1.00/0.00	0.90/0.15	0.08/0.07
GM-PLSDA	0.91/0.24	0.92/0.14	1.00/0.00	0.83/0.20	0.92/0.14	1.00/0.00	0.90/0.15	0.08/0.07

* 평균/표준편차

제 5 장

결론

본 논문에서는 고차원 데이터인 마이크로어레이 데이터에서 집단별로 불균형한 분포를 가질 때 적합한 분류 방법들에 대하여, 집단의 개수, 불균형의 정도 등을 고려한 여러 가지 실제 데이터를 통해 분류 방법의 성능을 비교하였다. 성능에 대한 평가측도로는 각 집단의 분류정확도, 각 집단의 분류정확도의 기하평균, 오분류율을 사용하였다.

실제 데이터 분석을 통해 얻은 주요 발견사항은 다음과 같다. PAM, ClaNC, SCRDA와 같이 LDA에 기반을 둔 분류기들이 불균형 집단데이터에서 성능이 크게 떨어지는 것을 확인했고. 기하평균을 이용한 방법들이 이를 잘 보완해 주는 것을 확인할 수 있었다. 세 분류기 중에서는 유전자 사이의 상관관계를 고려하는 SCRDA가 가장 우수한 성능을 보였다. SVM기반 방법과, PLSDA는 대체적으로 모든 데이터에서 좋은 결과를 보였다. 이미 많은 연구를 통해 우수한 성능을 인정받은 SVM뿐만 아니라 오래된 방법론이지만

PLSDA가 일반적으로 우수한 성능을 보이는 것을 확인할 수 있었다. 기하평균에 기반을 둔 알고리즘은 다수 집단의 분류정확도를 낮추고 소수 집단의 분류정확도를 높이면서 전체적으로는 기하평균을 높이는 효과를 보였다. 기하평균에 기반을 둔 방법은 Embed하게 유전자를 선택하는 분류기에서 우수한 성능을 보였으며, 다른 분류기에서도 그 효과를 보였다. 또한 기하평균에 기반을 둔 방법은 데이터의 불균형 정도가 심할 때, 집단의 개수가 많아 질 때 더 큰 효과를 보였다.

표본의 크기가 공변량의 수보다 더 많은 데이터인 고차원 데이터에서 대부분의 분류기들이 집단의 사전 분포가 다른 불균형 집단데이터에서 그 효율이 떨어짐을 확인 할 수 있었다. 하지만 각 집단 분류정확도의 기하평균에 기반을 둔 방법은 불균형 집단데이터에서 더 정확한 분류를 가능하게 함을 알 수 있었다. 본 연구를 기반으로 앞으로의 마이크로어레이 데이터와 같은 고차원 데이터 중 집단의 크기가 불균형인 데이터의 분류분석에 있어 보탬이 되기를 기대한다.

참 고 문 헌

[1] Alon, U. and Barkai, N. and Notterman, D.A. and Gish, K. and Ybarra, S. and Mack, D. and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci,* **96**(12), 6745 - 6750.

[2] Blagus. R., and Lusa, L. (2013). Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC Bioinformatics*, **14**(64).

[3] Breiman, L. (2001). Random forest. *Machine Learning*, **45**(1), 5-32.

[4] Dabney A.R. (2005). Classification of microarrays to nearest centroids. *Bioinformatics,* **21**(22), 4148-4154

[5] Dettling M and Beuhlmann P. (2002). Supervised clustering of genes. *Genome Biology,* **3**, research0069.1-0069.15.

[6] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M.,Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

[7] Guo, Y., Hastie, T., Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Bioinformatics,* **8**(1), 86-100.

[8] Lee, J.W. and Lee, J.B. and Park M.R. and Song S.H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, **48**, 869-885.

[9] Lin W.J. and Chen J.J (2012). Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics,* **14**(1), 13-26.

[10] Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Spellman, P., Iyer, V., Jeffrey, S.S., de Rijn, M.V., Waltham, M., Pergamenschikov, A., Lee, J.C.F., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O. (2000.) Systematic variation in gene expression patterns in human cancer cell lines. *Natur. Genetics* **24**, 227 - .234.

[11] Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., Weinstein, J.N. (2000). A gene expression database for the molecular pharmacology of cancer. *Natur. Genetics*. **24**, 236 - .244.

[12] Tibshiranit, R., Hastie, T., Narasimhan, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci.* **99**, 6567 - .6572

[13] Vapnik, V., 1998. Statistical Learning Theory. Wiley, Chichester, GB.