종양 분류에 쓰이는 Microarray 분석 기반 머신러닝 기법

초록

종양이란 세포의 DNA 에 유전자 변이가 형성되고 이로 인하여 유전자 발현에 변화가 일어나 세포의 형질이 변화하여 세포분열이 통제되지 않는 것을 말한다. 이러한 종양의 발생에서 유전자 발현은 종양의 유형을 결정하기에 이에 대한 분석은 정확한 종양의 진단과 치료에 매우 중요하다.

따라서 전체의 종양 유전자 발현을 조사하기 위한 방법으로 microarray 분석이 사용되었으며, 이때 microarray 는 발현하는 mRNA 와 상보적인 DNA probe 를 가지고 혼성화하여 측정하는 기법으로 대량의 데이터와 병렬화 과정에 의해서 여러 측정치와 함께 노이즈가 혼재해 있다. 따라서 이러한 대량의 데이터에서 종양 세포에서의 발현만을 특이적으로 분석하기 위해서는 통계적인 방식을 통해 데이터의 규칙을 추론할 수 있는 머신러닝을 사용하게 된다.

지금까지 다양한 머신러닝 알고리즘이 microarray 데이터 분석에 적용되었으며, 이에 따라 본 리뷰 논문에서는 종양을 진단하고 분류하기 위해서 microarray 에 대해 적용되었던 머신러닝 알고리즘 중 k-Nearest Neighbors (k-NN) 알고리즘과 Support Vector Machine (SVM)에 대해 알아보고 그 사례를 소개하려 한다.

서론

1. 종양 발생 및 유전자의 발현

세포가 자외선, 일부 화학물질, 감염원 등의 발암물질에 노출되어 중요한 유전자의 돌연변이가 일어나면 통제되지 않는 세포분열로 인해 종양을 형성한다. 악성 (malignant) 종양은 양성 (benign) 종양과 달리 다른 조직에 나쁜 영향을 주며, 이를 암이라고 한다. 이때 정상세포의 증식과 성장을 조절하는 원발암유전자 (proto-oncogene)가 활성화되거나 종양의 형성을 억제하는 종양억제유전자 (tumor suppressor gene)가 불활성화되어 발암과정이 전개된다. 돌연변이가 일어난 유전자는 조직 내 주변 세포와 그 후손 세포에 영향을 미쳐 비정상적 성장을 유발한다 (Lewin, 2007).

2. Microarray 원리 및 분석 방법

Microarray 는 특정 RNA 시퀀스에 대한 수천에서 수만개의 cDNA 혹은 oligonucleotide 를 유리 슬라이드 등의 고체 표면에 고정한 칩이다. 이를 이용해 목적 유전체의 유전자 발현 패턴을 폭넓게 관측할 수 있다.

(**Figure.1**)에 그 과정이 묘사된 대로, 대조군과 실험군에서 분리한 mRNA 를 역전사시켜 cDNA 로 합성하고, 이를 각각 Cy3 (녹색 형광), Cy5 (빨간색 형광)으로 표지한다. 이를 Microarray 에 고정된 cDNA 나 oligonucleotide 에 혼성화시킨다. 이때 방출되는 형광의 상대적 세기를 공초점 현미경 등으로 측정하여 그 정도를 수치화

하면, 특정 조건 하에서 상대적인 mRNA 발현량을 동시에 대량으로 모니터링할 수 있으므로 이를 통해 해당 세포의 유전자 발현 방식과 생화학적 특성을 이해할 수 있다 (Butte, 2002).

각각의 종양은 고유한 유전자 발현 패턴을 갖는다. 이러한 패턴은 종양의 분자 지문 (molecular fingerprint) 역할을 할 수 있고, 다양한 종양의 종류를 구분하거나 새로운 분류체계를 구성하는 데에 도움을 줄 수 있다 (Yeang et al., 2001). 그러므로 종양의 분자적 특성을 결정하는 유전자를 정확히 찾아내는 것이 중요하다.

Cy3 녹색 형광, Cy5 빨간 형광으로 각각 표지된 정상세포와 암세포의 cDNA 를 microarray 에 혼성화시키면 칩 위의 DNA 중 상보적인 서열을 찾아 결합하게 된다. 특정 유전자가 녹색 형광으로 표지되면, 그 유전자는 정상세포에서 많이 발현되고, 빨간 형광으로 표지되면 암세포에서, 노란색일 경우와 검은색일 경우는 각각 정상세포와 암세포 모두에서 발현되거나 발현되지 않는 유전자이다. 이를 통해 정상세포와 암세포 간 유전자 발현 차이를 알 수 있다 (George and Raj, 2011).

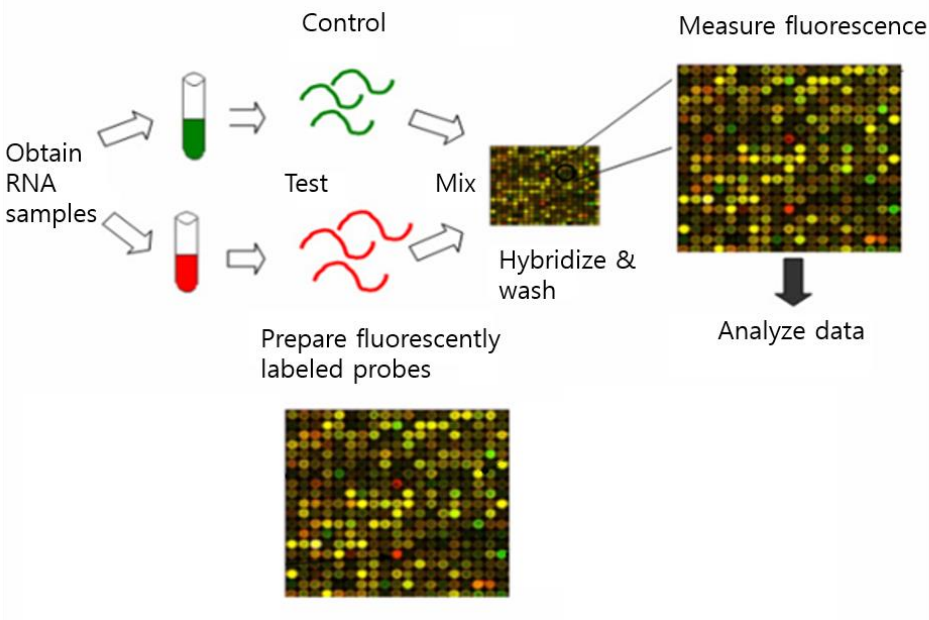

Figure.1. Microarray 분석 과정 (George and Raj, 2011)

$$M = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \leftarrow s_i, i = 1, \cdots, n$$

$$\uparrow$$

$$g_j, j = 1, \cdots, m.$$

Figure.2. Microarray 데이터 표현 (George and Raj, 2011)

Microarray 데이터는 벡터 또는 행렬로 표현되며, 이때 열벡터 g 는 각각의 유전자 (gene), 행벡터 s 는 개별 사례 (sample), 행렬 M 의 성분 w_{11}, ... w_{nm} 은 유전자의 발현 패턴을 나타낸다 (George and Raj, 2011).

cDNA microarray 에 사용되는 probe 는 일반적으로 250-750 bp 정도의 길이를 가져 cross-hybridization 발생 가능성이 있어 정밀도가 낮기 때문에 cDNA microarray 분석은 대체로 유전자가 어떤 상황에서 더 많이 발현되는지 전반적인 패턴을 분석하는 데에 쓰인다. oligonucleotide microarray 의 경우, probe 의 길이가 대부분 25-mer 이하로

짧아 염기 단위로 정밀하게 점돌연변이 (point mutation)까지 찾을 수 있어서 특정 유전자의 돌연변이에 의해 유발되는 종양이나 유전병 등의 진단 및 분류에 사용된다 (이정문, 2003).

3. 종양에서의 Microarray 활용

해당 종양 세포에서 전사되는 mRNA 양과 그 패턴을 정상 세포와 비교 분석하는 것은 정적인 유전체를 파악하는 방식에 비해 해당 종양 세포의 단백질 발현량과 생화학적 속성을 보다 직접적으로 알려준다. Microarray 분석을 활용하여 특정 종양의 유전적인 표지를 식별해 각 종양의 고유한 속성과 유형을 보다 정확하게 분류할 수 있다. 이러한 분류는 개별 종양에 대한 적확한 치료에 도움을 준다. 하나의 예로써, 급성 림프구성 백혈병 acute lymphoblastic leukemia (ALL)과 급성 골수성 백혈병 acute myeloid leukemia (AML)은 종양의 형태 분석, 조직 화학, 면역표현형 검사 등의 단일 테스트로 구분하기에 까다로웠고 종종 틀리는 경우가 있었다. 또한 각각 효과적인 치료 방법이 달라 틀린 방식으로 치료하게 되면 치유 속도가 현저히 감소하거나 독성이 발생할 수 있으므로 보다 정확한 종양 분류가 요구되었다 (Golub et al., 1999).

Microarray 로 만들어진 데이터는 방대한 량의 유전자가 관련되어 있고, 고가의 비용 등으로 인해 그 샘플의 수가 적어 유전자의 수가 샘플의 수에 비해 압도적으로 많기

때문에 (leukemia AML-ALL 이진분류의 경우, 총 7,129 개 유전자와 관련된 72 개의 샘플을 대상으로 한다 (Ben-Dor et al., 2000).) 분석하기 까다로운 특성을 갖는다.

샘플의 개수에 비해 다루는 유전자 수가 많아질수록 벡터 공간의 차원이 커져 학습이 어려워지는데, 이를 차원의 저주 (Curse of dimensionality)라고 한다. 그러므로 (**Figure.4**)와 같이 종양 분류 과정 이전에 feature selection 혹은 gene selection 이라는 과정을 거쳐 해당 종양과 보다 상관성이 높은 유전자를 잘 선별하는 과정이 요구되었고, 이는 데이터의 차원을 줄여 종양 분류 모델의 성능을 높여준다. 이후 분류에서도 방대한 연산량 등을 이유로 다양한 머신러닝 기법이 도입되어 (Dudoit et al., 2002; Ressom et al., 2008) 종양 유형 분류 분야를 크게 발전시켰다.

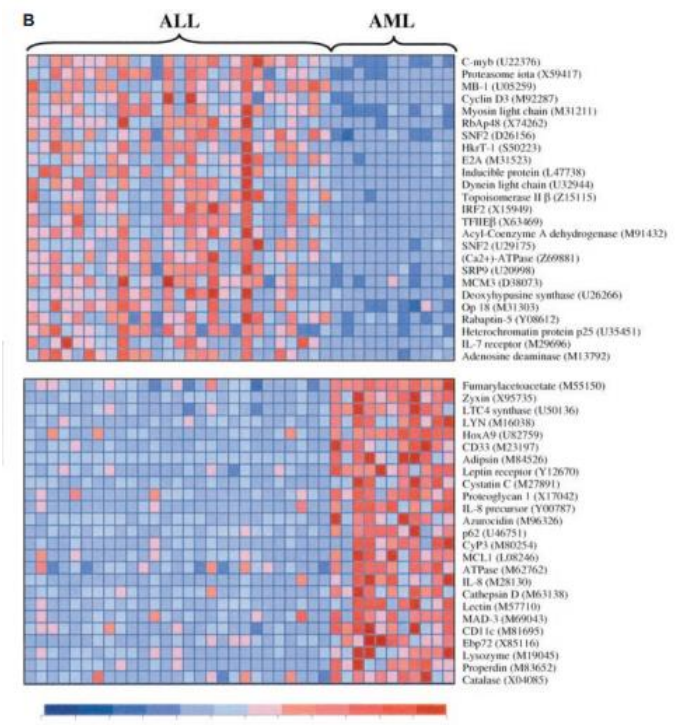

Figure.3. ALL 과 AML 인 경우의 유전자 발현 유전자 발현 정도에 따라 두가지 유형의 백혈병 ALL AML 을 분류할 수 있다. 각 행은 개별 유전자로 빨간색이 짙을수록 많이 발현되었고, 파란색이 짙을수록 적게 발현되었다 (Golub et al., 1999).

따라서 본 리뷰논문에서는 종양 분류에 쓰이는 microarray 분석 기반 머신러닝 기법, 특히 종양 분류에 초기부터 최근까지 꾸준히 채택되고 있는 k-Nearest neighbors (k-NN, 최근접 이웃) 알고리즘과 Support Vector Machine (SVM)에 대해서 알아보려 한다.

본론

1. 머신러닝

머신러닝이란 사람이 직접 명령을 작성하지 않아도 기계가 데이터로부터 규칙을 학습하고 기능할 수 있도록 하는 알고리즘으로, 사람이 일일이 규칙을 입력하는 방식이 아니라 주로 통계적인 접근 방법을 이용해 데이터로부터 규칙을 추론해내는 과정이다. 기본적인 머신러닝 알고리즘을 크게 구분하면 지도학습과 비지도학습으로 나눌 수 있는데 지도학습은 정답이 있는 데이터를 이용해 학습한 다음 새로 입력된 데이터의 정답을 추론해 내는 방식이고, 비지도학습은 정답이 주어져 있지 않은 데이터의 패턴을 다른 것들과 구분하는 방식이다 (Alpaydin, 2010).

Microarray 분석 기반의 종양 분류 사례에서는 이미 알려진 종양의 클래스를 분류하는 모델이 지도학습이며, k-NN, SVM 등의 알고리즘을 사용한다. 한편, 알려지지 않은 새로운 종양의 패턴을 다른 것들과 구분하는 군집화 모델이 비지도학습으로 (Dudoit et al., 2002), K-means, Hierarchical clustering, Self-organizing map 등의 알고리즘이 있다.

Microarray 분석을 통해 얻은 유전자 발현 데이터를 토대로 종양을 분류하는 과정 (**Figure.4**)에서 종양 분류에 유의미한 후보 유전자를 골라내는 feature selection 이 요구된다. 해당 종양과의 상관성이 높아 선별된 유전자를 토대로 k-NN 이나 SVM 등의

분류 모델을 학습하면 모델의 성능 개선을 기대할 수 있기 때문이다. 이를 통해 정상세포-종양 분류, 또는 종양간 하위 클래스 분류 등의 과정을 수행한다.

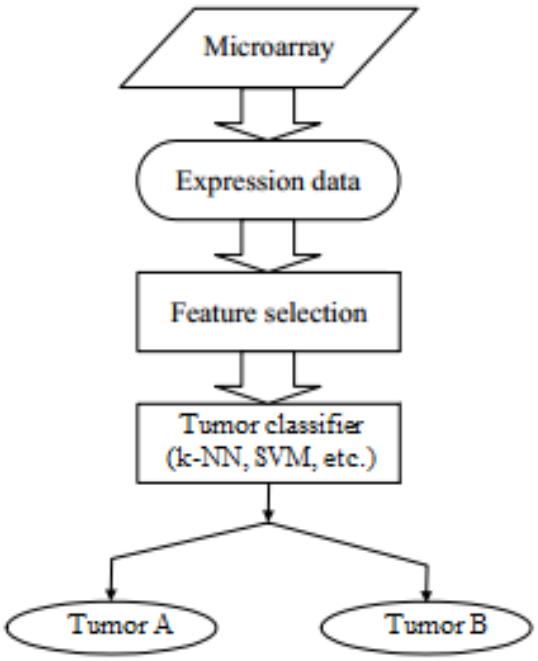

Figure .4. Microarray 유전자 발현 데이터를 이용해 종양을 분류하는 일반적인 과정 (Cho and Won, 2003)

2. k-Nearest neighbors (k-NN, 최근접 이웃)

1) k-NN 알고리즘

k-NN 은 비교적 간단한 방식의 기계학습 알고리즘으로, 분류나 회귀에 쓰이는 인스턴스 기반 학습(instance-based learning)의 비모수적 모형(non-parametric model)이다. 분류 모델로써 k-NN 은 다른 알고리즘에 비해 적용하기 쉽고 빠르면서도 성능이 보장된다는 장점을 갖고 있다.

이미 클래스가 분류된 훈련 데이터를 벡터 공간에 사영하고 새로 입력된 데이터가 어떤 클래스에 속할지 k 개 주변 데이터와의 거리 정보를 이용해 결정하는 작동 방식으로, 클래스가 분류되지 않은 새로운 데이터가 입력될 때 일반적으로 각 데이터 벡터 간 유클리드 거리를 척도로 하여 사용자가 설정한 k 개의 가장 근접한 이웃 벡터들을 가장 유사한 데이터라고 인식한다. k 개의 가장 근접한 데이터의 다수결

(대체로 k 는 홀수로 설정한다)을 통해 새로 입력된 데이터를 같은 클래스로 분류하고 이를 반복 확장하여 전체 데이터를 분류한다. 인스턴스 기반 학습의 장점으로, 새로운 데이터가 들어왔을 때 전체 데이터를 새로 학습하지 않고 기존 데이터 벡터와의 거리를 이용해 새로운 데이터의 클래스로 할당할 수 있다 (Harrington, 2012).

2) 사례

머신러닝을 도입한 초기의 사례 (Golub et al., 1999)에서는 k-NN 이나 SVM 보다 간단한 weighted voting 등의 분류방식을 채택해 acute lymphoblastic leukemia (ALL, 급성 림프구성 백혈병)과 acute myeloid leukemia (AML, 급성 골수성 백혈병) 두가지 유형의 종양을 분류했다.

분류 모델의 학습 데이터로 이미 진단이 내려진 환자 38 명 (27 ALL, 11 AML)의 골수 샘플에서 채취한 단핵세포(mononuclear cells)에서 RNA 를 준비하였고, 이를 high-density oligonucleotide microarray (probe 로 사람 유전자 6,817 개 사용)에 혼성화시킨 후 그 형광을 측정했다. 그 다음 AML-ALL 구분에 강한 상관관계를 갖는 유전자를 선별하기 위해 측정된 값의 Pearson correlation 을 구해 1,100 개의 feature 를 구했다.

여러 차례 통계적 방식을 수행하여 그 중 가장 뚜렷한 상관관계를 보이는 10 - 200 개의 유전자를 선정 분류 모델을 만들었고, 새로 입력한 34 개의 샘플을 테스트 데이터로 그 분류 정확도 (accuracy)를 측정하였다.

분류 결과 34 명의 환자에서 29 명을 100%의 정확도로 진단할 수 있었다. 이 성과를 토대로 microarray 데이터를 이용해 종양을 분류하는 본격적인 방법론을 제시하였다.

Index	Tumor type
0	Breast
1	Prostate
2	Lung
3	Colorectal
4	Lymphoma
5	Bladder
6	Melanoma
7	Uterus
8	Leukemia
9	Renal
10	Pancreas
11	Ovary
12	Mesothelioma
13	Brain

Table.1. 14 가지 종양 (Yeang et al., 2001).

이들은 후속 연구 (Yeang et al., 2001)에서 분류 모델로 k-NN 알고리즘과 SVM 등을 도입하였는데 16,063 개의 알려진 사람 유전자를 probe 로 한 high-density oligonucleotide microarray 를 이용해 190 종양 샘플에서 14 가지의 종양을 분류했다.

(Ben-Dor et al., 2000)는 colon cancer, ovarian cancer, leukemia 데이터를 기반으로 k-NN 을 비롯해 SVM, AdaBoost 등의 종양 이진 분류 모델을 검증하고 성능을 비교하였다. 특히 분류 모델이 사용한 유전자의 수를 2,000 개까지 늘려 최대 200 개로 제한되었던 (Golub et al., 1999)에 비해 유전자간 동시 발현 패턴을 더욱 폭넓게 활용할 수 있었다.

colon cancer 데이터 셋은 tumor 와 normal 두 가지 클래스로 구성되어 있으며, 샘플의 수는 총 62 개로 결장암 환자들의 결장 상피세포이다. Tumor 에 속하는 데이터는

환자의 종양에서, normal 에 속하는 데이터는 같은 환자의 건강한 결장 부위에서 수집하였다. High-density oligonucleotide array probe 로 약 6,000 여 유전자를 사용했으며, 이 중 2000 개 유전자를 선별하여 분류 모델링을 진행했다.

Ovarian cancer 데이터 셋은 15 개 난소암 샘플과 13 개 건강한 난소 샘플로 구성된 총 32 개이며, 마찬가지로 tumor 와 normal 두 가지 클래스로 이루어져 있다.

Leukemia 데이터 셋은 (Golub et al., 1999)사례와 동일한 구성이나 샘플의 수는 72 개로 AML 25 개, ALL 47 개로 구성되었다. High-density oligonucleotide array probe 로 총 7,129 개 유전자를 사용했다.

		Percent		
Data set	*Method*	*Correct*	*Incorrect*	*Unclassified*
Colon	Nearest Neighbor	80.6	19.4	0.0
	SVM, linear kernel	77.4	12.9	9.7
	SVM, quad. kernel	74.2	14.5	11.3
	Boosting, 100 iter.	72.6	17.7	9.7
	Boosting, 1000 iter.	72.6	17.7	9.7
	Boosting, 10,000 iter.	71.0	19.4	9.7
Ovarian	Nearest Neighbor	71.4	28.6	0.0
	SVM, linear kernel	67.9	3.6	28.6
	SVM, quad. kernel	64.3	3.6	32.1
	Boosting, 100 iter.	89.3	10.7	0.0
	Boosting, 1000 iter.	85.7	10.7	3.6
	Boosting, 10,000 iter.	85.7	14.3	0.0
Leukemia	Nearest Neighbor	91.6	8.4	0.0
	SVM, linear kernel	93.0	1.4	5.6
	SVM, quad. kernel	94.4	1.4	4.2
	Boosting, 100 iter.	95.8	2.8	1.4
	Boosting, 1000 iter.	95.8	2.8	1.4
	Boosting, 10,000 iter.	95.8	2.8	1.4

Table.1. Colon, ovarian, leukemia 데이터 셋에서의 각 분류 모델 성능 요약 (Ben-Dor et al., 2000).

분류 모델의 성능은 다음과 같았다. Colon cancer, ovarian cancer 분류에서 80.6%, 71.4%의 정확도로 SVM 의 77.4%, 67.9%에 비해 보다 높은 성능을 보였다. Leukemia 에 대해 91.6% 정확도를 보였으나 SVM 의 94.4%에 미치지 못했다. 다만 k-NN 알고리즘의 장점으로 분류하지 못한 데이터 (unclassified)가 없었다는 것이 주목할 만하다 (Ben-Dor et al., 2000).

	sample size	number of genes
ALL-MAL	72	7129
Breast-ER	49	7129
Breast-LN	49	7129
CNS	60	7129
Colon	62	2000
Lung	181	12533
Lymphoma	77	7129
Ovarian	253	15154
Prostate	102	12600

Table.3. 데이터 셋 개요 (Xiong et al., 2007)

(Xiong et al., 2007)에서는 8 가지 종양 분류에 대한 연구를 수행했는데, breast cancer 의 경우, Estrogen Receptor (ER), Lymph Node (LN) status 에 대한 각각의 데이터로 분류 모델을 구성했다. k-NN, SVM 과 함께 추가적인 두가지의 선형 분류기 (Diagonal Linear Discriminant Analysis, DLDA, Uncorrelated Linear Discriminant Analysis, ULDA)를 도입하여 성능을 비교했다. 특히 k-NN 알고리즘을 개량해 학습에 사용하는 데이터 resampling 방식을 고도화하여 학습의 편향을 줄인 kernel-based k-NN (KerNN) 분류 모델을

	KNN	DLDA	ULDA	SVM	KerNN
ALL-AML	5.78 (3.58)	**4.16** (2.57)	6.92 (4.13)	5.24 (3.20)	5.14 (3.10)
Breast-ER	10.32 (5.12)	**7.04** (5.63)	9.76 (6.86)	7.36 (4.07)	7.84 (5.11)
Breast-LN	14.16 (5.89)	9.74 (5.61)	15.52 (8.35)	9.76 (6.91)	**8.08** (5.08)
Colon	17.03 (4.66)	14.00 (4.59)	20.90 (8.04)	13.55 (5.05)	**12.90** (4.47)
CNS	21.94 (5.37)	23.61 (6.47)	17.10 (8.65)	**15.94** (7.24)	16.02 (6.36)
Lung	1.56 (1.20)	1.23 (0.91)	1.87 (1.59)	1.21 (0.81)	**0.84** (0.75)
Lymphoma	4.10 (3.88)	8.36 (3.90)	5.59 (7.27)	**3.38** (3.33)	**3.38** (2.95)
Ovarian	1.15 (1.30)	1.98 (1.05)	**0.39** (0.77)	0.44 (1.02)	0.44 (0.92)
Prostate	8.63 (3.43)	6.78 (3.00)	7.57 (5.59)	6.78 (3.02)	**6.43** (4.10)

Table.4. 평균 테스트 에러율 (1-정확도) (%)와 표준편차 (괄호 안) 비교 (Xiong et al., 2007). 에러율 **굵은 글씨**의 경우, 가장 우수한 성능을 보인 분류 모델.

제안했는데 이는 기존 k-NN 뿐만 아니라 다른 분류 모델에 비해 대부분의 데이터 셋에서 분류 에러율이 가장 낮은 편이었고, feature 의 개수가 늘어나도 꾸준히 안정적인 성능을 보였다.

Type	Number of samples
HBM	26
AML	259
ALL	197
CML	53
CLL	29

Table.5. Leukemia 의 5 가지 subtypes (Castillo et al., 2019)

k-NN 알고리즘은 핵심 아이디어가 단순하고 문제 상황에 적용하기 간편해 종양 분류 모델로써 초기에 도입되었으나 최근까지도 꾸준히 활용되고 있다. (Castillo et al., 2019)에서는 5 개의 클래스로 구성된 leukemia 를 분류하였는데, 구성은 (Table.5) 와 같다. SVM, Naïve Bayesian (NB), Random Forest (RF) 알고리즘을 써서 k-NN 과의 성능을 비교하였다.

	10 Genes	*20 Genes*	*30 Genes*	*40 Genes*
Classifier	ACC	ACC	ACC	ACC
SVM	95.64%	96.61%	98.14%	97.83%
k-NN	96.40%	98.56%	98.78%	98.87%
NB	94.76%	95.98%	95.34%	95.66%
RF	95.51%	95.05%	95.35%	95.42%

Table.6. 분류 모델들의 분류 정확도 (Castillo et al., 2019)

k-NN 분류 모델은 특히 feature 의 수가 40 개인 40 genes 케이스에서 leukemia 5 class 분류 정확도가 98.87%로 다른 분류 모델들에 비해 가장 좋은 성능을 보였다.

3. Support Vector Machine (SVM)

1) SVM

데이터의 집합이 주어졌을 때, 이를 벡터 공간에 사영하면 각각의 점으로 표현된다. 공간상의 점들이 두 가지 카테고리에 각각 속한다고 할 때, 이들 중 어디에 속하는가 결정하는 것이 분류 모델의 목표이다.

간단한 사례로써, (**Figure.5**)로 microarray data 로 종양을 분류하는 SVM 의 작동 방식을 설명하려 한다. (**Figure.5a**)는 AML-ALL 구분에 강한 상관관계를 보여 선택된 두 유전자 ZYX 와 MARCKSL1 의 유전자 발현량 데이터 벡터를 2 차원 평면에 사영한 것을 나타낸 것이다. 각 데이터 벡터는 x 값으로 MARCKSL1 의 형광 세기, y 값으로 ZYX 의 형광 세기 값을 갖는다. 샘플의 색깔은 초록색과 빨간색 중 더 강하게 나타난 형광으로 표현하였다. (**Figure.5b**)는 AML-ALL 를 구분하는 직선을 보여주며, 이를 초평면 (hyperplane)이라고 한다. 만일 고려할 유전자가 애초에 한 개였으면, 데이터는 (**Figure.5c**)와 같이 1 차원 직선 위에 존재할 것이고 검은 점 (hyperplane)으로 두 클래스를 분류할 수 있다. 기존의 두 유전자 외에 하나의 유전자가 더 추가되어 데이터가 3 차원 벡터 공간에 표현된다면, (**Figure.5d**)는 늘어난 차원에 맞춰 새로운 유전자 HOXA9 좌표축을 도입하여 벡터 공간에 데이터를 사영한 것을 나타내고, 적절히 분류하기 위해 3 차원 공간에 비해 한차원 낮은 평면 (hyperplane)이 필요하다.

(**Figure.5e**)는 다시 평면에서 AML-ALL 을 구분하는 여러 초평면들을 나타내는데, 그 중 AML-ALL 을 가장 효과적으로 분류하는 직선을 (**Figure.5f**)와 같이 구할 수 있다.

이는 초평면과 각 클래스에 속한 가장 가까운 점들과의 거리 (margin)가 최대가 되는 성질을 가진다 (Noble, 2006).

이제 일반화하여 SVM은 데이터가 n-차원의 벡터로 주어졌을 때, 이를 (n-1)-차원의 초평면으로 분류 가능한지 확인한다. 여러 초평면이 존재할 경우 가장 가까이에 있는 각 클래스의 벡터와 초평면 사이의 거리를 최대가 되는 초평면을 선정하면 이 초평면은 데이터를 효과적으로 분류했다고 할 수 있고 이 분류 모델을 최대 마진 분류기 (maximum margin classifier)라고 한다. (Harrington, 2012).

2) 사례

비교적 초기에 나온 연구로 ovarian cancer (cancerous, normal), leukemia (AML, ALL), colon tumor (tumor, normal)에 대해 분류 모델로 SVM을 이용한 이진분류 사례 (Furey et al., 2000)가 있다.

SVM 은 벡터가 희소하거나 노이즈가 많은 데이터에 안정적인 성능을 내는 분류 모델로 알려져 있으며, 적당한 선형 분류 초평면이 없을 때 kernel 함수를 이용해 데이터를 고차원 공간으로 사영해 비선형 분류 초평면으로 클래스를 구분할 수도 있다. 이 연구에서는 linear kernel, polynomial kernel 등 다소 간단한 kernel 함수들만 이용했다.

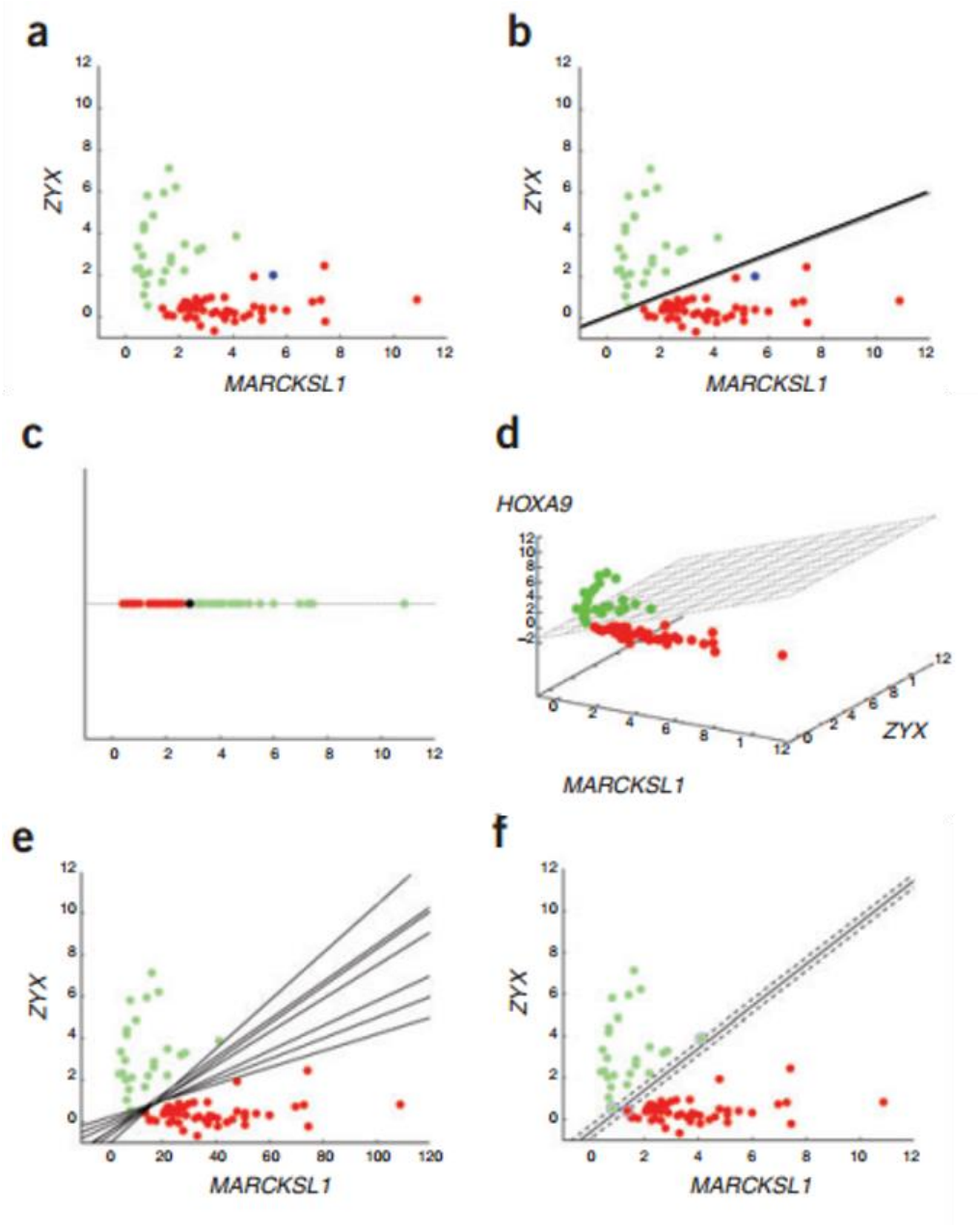

Figure.5. 간단한 SVM 동작 방식 설명 (a) 유전자 발현 데이터의 2 차원 표현. (b) 두 클래스를 구분하는 초평면. (c) 하나의 유전자 발현 데이터가 있을 때 표현된 1 차원. 이때 초평면은 점으로 표현된다. (d) 분류에 쓰이는 새로운 유전자 도입 시 데이터의 3 차원 표현. (e) 두 클래스를 구분하는 여러 초평면. (f) 그 중 두 클래스를 가장 잘 분류하는 최대 마진 초평면. (Noble, 2006)

Ovarian 데이터는 총 31 개의 샘플이고, leukemia 데이터는 총 72 개로 38 개의 훈련 데이터 (ALL: 27, AML: 11), 34 개의 테스트 데이터 (ALL: 20, AML: 14)로 구성되어 있다. High-density oligonucleotide microarray 를 이용해 7,129 개의 사람 유전자를 대상으로 유전자 발현량 데이터를 구했다. Colon 분류기는 40 개의 tumor, 22 개의 normal 로 훈련되었으며, 테스트 데이터는 62 개로 구성되었다. High-density oligonucleotide microarray 의 probe 로 6,500 개 사람 유전자를 이용했다.

Feature	FP	FN	TP	TN	Accuracy (%)
25	5	4	10	12	70.97
25	5	2	12	12	77.42
25	4	2	12	13	80.65
25	4	2	12	13	80.65
50	4	2	12	13	80.65
50	3	2	12	14	83.87
50	3	2	12	14	83.87
50	3	2	12	14	83.87
100	4	3	11	13	77.42
100	5	3	11	12	74.19
100	5	3	11	12	74.19
100	5	3	11	12	74.19
97 802	17	0	14	0	45.16
97 802	9	2	12	8	64.52
97 802	7	3	11	10	67.74
97 802	5	3	11	12	74.19

Table.7. Ovarian cancer 분류 성능. FP: normal 잘못 분류, FN: tumor 잘못 분류, TP: tumor 제대로 분류, TN: normal 제대로 분류 (Furey et al., 2000).

분류 성능을 개선하기 위해 한가지 분류 모델 SVM 을 다양한 방식으로 조정하였다. SVM 의 훈련 오차를 줄이기 위해 kernel 함수와 관련된 kernel matrix 의 diagonal factor 를 조절했고, soft-margin 을 적용했다. Feature 가 50 개, kernel matrix 의 diagonal factor 가 2, 5, 10 일 때 가장 높은 성능을 거두었다 (정확도 83.87%).

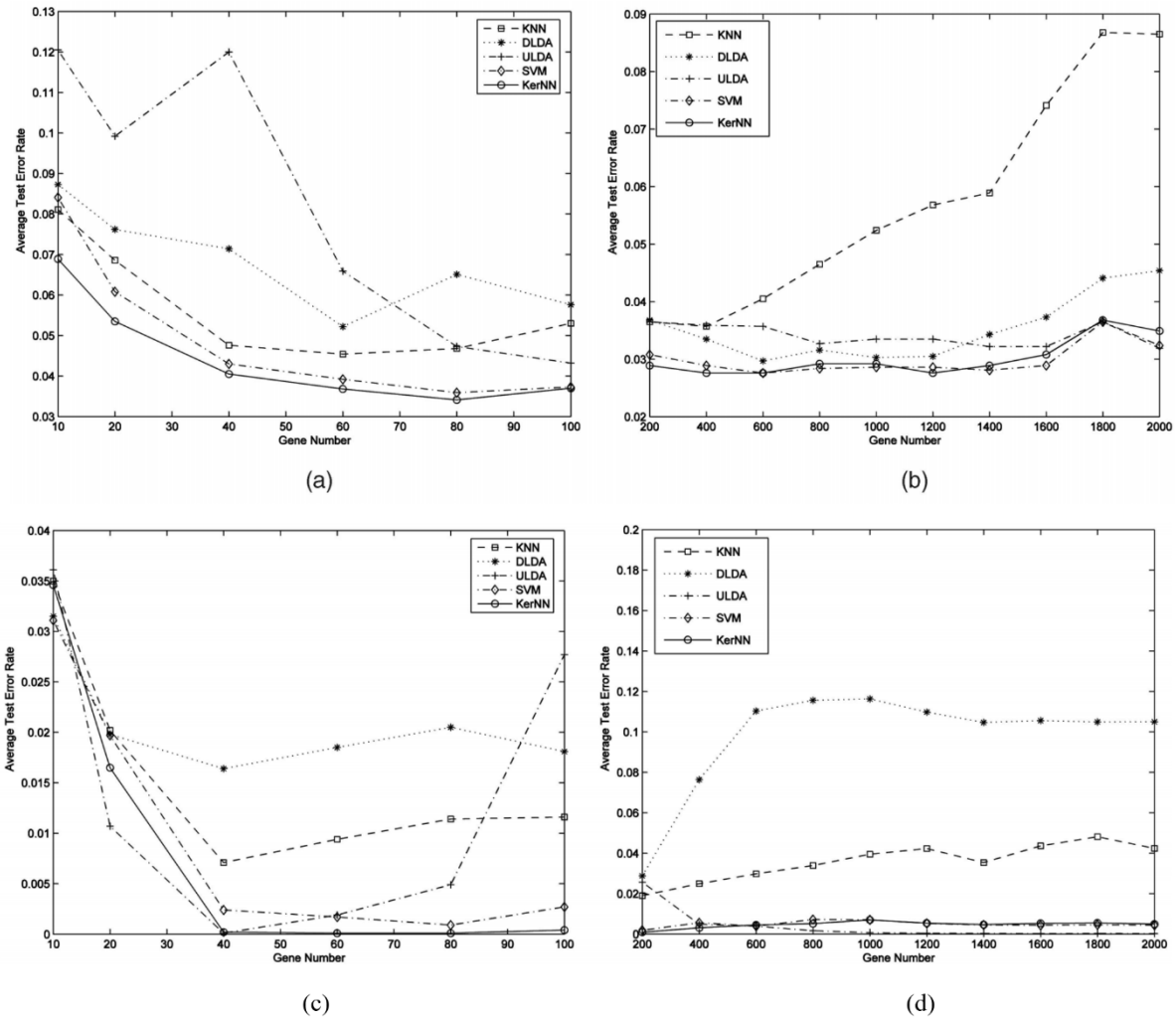

Figure.6. 각 모델들의 분류 성능. x 축은 유전자 수, y 축은 평균 에러율을 나타낸다. (a), (b)는 feature 수에 따른 각 ALL-AML 분류 모델의 평균 에러율. (c), (d)는 feature 수에 따른 각 ovarian cancer 분류 모델의 평균 에러율 (Xiong et al., 2007).

(Xiong et al., 2007) 사례에서는 SVM 을 비롯해 k-NN, DLDA, ULDA, KerNN 등의 분류

모델을 사용하였다. 데이터는 (**Table.3**), 분류 모델 성능은 (**Table.4**)에서 확인할 수 있다.

이 연구에서 SVM 은 에러율이 낮거나 에러율의 표준 편차가 작아 대부분의 클래스 구분에서 안정적인 성능을 보였다. 또한 feature 수가 10, 20, 40, 60, 80, 100, 200, 400, 600, 800, 1,000, 1,200, 1,400, 1,600, 1,800, 2,000 으로 늘어남에 따라 오차율 변동이 심한 다른 모델들에 비해 큰 변화폭 없이 안정적으로 종양을 분류할 수 있었다.

(Pirooznia et al., 2008)은 SVM 을 비롯, 다양한 분류 모델을 이용해 여덟 가지 종양 사례에 대해 이진 분류하였고 그 성능을 비교하였는데 이 연구에서 쓰인 데이터는 다음과 같다.

Dataset	Comparison	Variables (Genes)	Samples
1. Lymphoma (Devos et.al, 2002)	Tumor vs. Normal	7129	25
2. Breast Cancer (Perou et. al, 2000)	Tumor subtype vs. Normal	1753	84
3. Colon Cancer (Alon et. al, 1999)	Epithelial vs. Tumor	7464	45
4. Lung Cancer (Garber et. al, 2001)	Tumor vs. Normal	917	72
5. Adenocarcinoma (Beer et.al, 2002)	NP vs. NN	5377	86
6. Lymphoma (Alizadeh et al, 2000)	DLBCL1 vs. DLBCL2	4027	96
7. Melanoma (Bittner et. al, 2000)	Tumor vs. Normal	8067	38
8. Ovarian Cancer (Welsh et. al, 2001)	Tumor vs. Normal	7129	39

Table.8. 8 가지 종양. 분류 기준, feature 의 수, 샘플의 수 (Pirooznia et al., 2008).

(Xiong et al., 2007)의 경우 분류 모델이 다룬 데이터의 feature 의 수는 2000 개까지였는데, 이 연구에서는 그 크기가 대폭 증가하여 melanoma 분류 모델은 8,067 개의 유전자를 대상으로 학습을 진행했음에도 분류 성능이 향상되었다. 이를 통해 종양과 강한 상관관계를 갖는 유전자의 동시 발현 패턴을 더욱 폭넓게 활용해 분류모델을 학습시킬 수 있었다. 다양한 데이터셋, 다양한 분류 알고리즘에서 SVM 분류모델의 성능이 안정적이면서 가장 우수하다는 것을 보였으며, 예전 연구들에 비해 에러율을 상당히 낮춘 것을 볼 수 있다.

Dataset	SVM	RBF Neural Nets	MLP Neural Nets	Bayesian	J48 Decision Tree	Random Forest	Id3	Bagging
1. Lymphoma (Devos et.al, 2002)	96.0	84.0	68.0	88.0	64.0	76.0	48.0	52.0
2. Breast Cancer (Perou et. al, 2000)	97.6	97.6	96.4	92.9	92.9	96.4	94.0	96.4
3. Colon Cancer (Alon et. al, 1999)	95.6	91.1	91.1	93.3	91.1	80.0	88.9	93.3
4. Lung Cancer (Garber et. al, 2001)	97.2	97.2	97.2	95.8	94.4	95.8	97.2	97.2
5. Adenocarcinoma (Beer et.al, 2002)	96.5	94.2	75.6	75.6	74.4	79.1	66.3	79.1
6. Lymphoma (Alizadeh et al, 2000)	96.9	88.5	75.0	85.4	75.0	76.0	62.5	84.4
7. Melanoma (Bittner et. al, 2000)	94.7	81.6	84.2	76.3	81.6	81.6	52.6	81.6
8. Ovarian Cancer (Welsh et. al, 2001)	94.9	84.6	89.7	87.2	87.2	89.7	74.4	89.7

Table.9. 각 분류 모델의 분류 정확도 (accuracy). 분류 모델로 SVM, 두 종류의 neural nets, 두 종류의 decision tree 기반 알고리즘 (J48, ID3), Bayesian method, random forest, bagging 등을 사용하였다 (Pirooznia et al., 2008).

결론

세포가 발암물질로 인해 유전자 변이가 일어나 세포분열이 통제되지 않는 경우 이를 종양이라 한다. 이때 종양세포의 유전자 발현은 종양의 분자적 특성과 종양의 유형을 결정하므로 치료 방법 또한 구분된다.

종양의 형태 분석, 조직 화학 등의 검사 방법으로는 종양의 유형을 분류하기에 제한적이었으므로 종양의 정확한 진단과 치료를 위해 microarray 분석으로 서로 다른 종양세포 간 유전자의 발현 패턴 차이를 측정하여 이를 해결하려 했다.

Microarray 분석을 이용해 만들어진 유전자 발현 데이터는 feature 의 수가 방대한 데에 비해 샘플의 수가 훨씬 적다는 특성을 가져 분석이 용이하지 않았다. (Welford et al., 1998), (Golub et al., 1999) 등의 사례를 보면 제한적인 유전자 수를 이용해 폭넓은 유전자 발현 패턴을 관찰하거나 분류 모델을 연구하기에 한계가 있었음을 알 수 있다.

머신러닝은 주로 통계적인 접근 방법을 이용해 기계가 데이터로부터 규칙을 학습하고 기능하는 과정으로, 대량의 데이터를 효율적으로 다룰 수 있다. 분류 모델은 지도학습의 대표적인 사례로 정답이 주어진 데이터를 학습해 새로운 데이터의 정답을 예측하는 머신러닝의 일종이다. 이 분류 모델을 이용해 microarray 유전자 발현 데이터를 기반으로 종양 분류를 수행할 수 있었다.

k-NN, SVM 등의 머신러닝 기법을 도입하면서 종양 분류 연구 발전에 가속이 붙었으며 (Ben-Dor et al., 2000), 분류 모델의 성능을 점차 향상시켰고 각각의 머신러닝 알고리즘을 다양하게 조정해 사용 데이터의 차원이 커져도 성능을 개선시킬 수 있었다 (Furey et al., 2000).

k-NN 알고리즘을 이용해 leukemia 유형을 분류할 수 있었고, colon cancer, ovarian cancer 등을 종양과 정상세포로 분류할 수 있었으며 (Ben-Dor et al., 2000), 최근의 연구 성과로 564 개의 leukemia 데이터 (5 가지 하위유형)를 이용해 선별한 40 개의 유전자로 학습한 분류 모델은 98.87%의 정확도를 기록하였다 (Castillo et al., 2019).

(Furey et al., 2000)은 다양한 모델 조정으로 종양 분류 모델로써 SVM 의 훈련 오차를 줄일 수 있다는 것을 보였고, (Xiong et al., 2007)은 다른 분류 모델과 비교했을 때 SVM 의 성능이 안정적이라는 것을 증명하였다. 여덟 가지 종양을 이진분류한 (Pirooznia et al., 2008) 사례에서는 유전자 발현 데이터의 feature 개수를 줄이지 않고도 우수한 성능을 보이는 SVM 분류 모델을 학습시킬 수 있다는 것을 확인하였다.

많은 사례를 통해 분류 모델의 성능을 살펴보았는데 이는 문제 상황과 데이터의 속성에 따라 변동이 있으므로, 다양한 데이터 각각의 특성에 알맞은 알고리즘을 적용할 필요가 있다.

참고문헌

Alpaydin, E. (2010). Introduction to machine learning, 2nd edn (Cambridge, Mass.: MIT Press), 1-13.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., and Yakhini, Z. (2000). Tissue classification with gene expression profiles. J Comput Biol *7*, 559-583.

Butte, A. (2002). The use and analysis of microarray data. Nat Rev Drug Discov *1*, 951-960.

Castillo, D., Galvez, J.M., Herrera, L.J., Rojas, F., Valenzuela, O., Caba, O., Prados, J., and Rojas, I. (2019). Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level. PLoS One *14*, e0212127.

Cho, S.-B., and Won, H.-H. (2003). Machine learning in DNA microarray analysis for cancer classification. Paper presented at: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19 (Australian Computer Society, Inc.).

Dudoit, S., Fridlyand, J., and Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc *97*, 77-87.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics *16*, 906-914.

George, G., and Raj, V.C.J.a.p.a. (2011). Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile.

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

Harrington, P. (2012). Machine learning in action (Shelter Island, N.Y.: Manning Publications Co.), 18-20, 101-106.

Lewin, B. (2007). Cells (Sudbury, Mass.: Jones and Bartlett Publishers), 561-573.

Noble, W.S.J.N.b. (2006). What is a support vector machine? *24*, 1565.

Pirooznia, M., Yang, J.Y., Yang, M.Q., and Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics *9 Suppl 1*, S13.

Ressom, H.W., Varghese, R.S., Zhang, Z., Xuan, J., and Clarke, R. (2008). Classification algorithms for phenotype prediction in genomics and proteomics. Front Biosci *13*, 691-708.

Xiong, H., Zhang, Y., Chen, X.-W.J.I.A.t.o.c.b., and bioinformatics (2007). Data-dependent kernel machines for microarray data classification. *4*, 583-595.

Yeang, C.H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M., Lander, E., Mesirov, J., and Golub, T. (2001). Molecular classification of multiple tumor types. Bioinformatics *17 Suppl 1*, S316-322.

이정문 (2003). 유전 알고리즘을 이용한 마이크로어레이 표본 분류에 유용한 유전자 선택 (서울대학교 대학원).