Potential SVM을 이용한 압타머칩에서의 바이오마커 탐색

김병희[o1], 김성천[2], 장병탁[1]
[1]서울대학교 컴퓨터공학부 바이오지능 연구실
{bhkim, btzhang}@bi.snu.ac.kr
[2](주) 제노프라
kimgp@cotech.co.kr

Biomarker Detection on Aptamer-based Biochip Data by Potential SVM

Byoung-Hee Kim[o1], Sung-Chun Kim[2], and Byoung-Tak Zhang[1]
[1]BI Lab, School of Computer Sci. & Eng., Seoul National University
[2]GenoProt Inc.

요 약

압타머칩은 혈청(serum) 내의 지정된 단백질의 상대적 양을 직접 측정할 수 있는 바이오칩으로서, 의학적 질병 진단에 유용하게 사용할 수 있는 툴이다. 압타머칩 데이터 분석에는 기존의 마이크로어레이 분석 기법을 그대로 적용할 수 있다. 본 논문에서는 Potential SVM(PSVM)을 이용하여, 심혈관질환 샘플 기반의 압타머칩 데이터에서 바이오마커 후보 단백질을 선정한 결과를 정리한다. PSVM은 분류 알고리즘으로서 뿐만 아니라 자질 선택(feature selection)에서도 우수한 성능을 보이는 알고리즘으로 알려져 있다. 심혈관 질환의 단계에 따라 구분한 4개 클래스, 135개 샘플로 구성된 3K 압타머칩 데이터에 대해 PSVM을 적용하여 자질을 선택하고 분류성능을 측정한 결과, 마이크로어레이에서의 자질 선택에 많이 사용되는 Gain Ratio 기법과 비교하여 보다 적은 수의 단백질 정보로 보다 나은 분류 성능을 보임을 확인하였다. 더불어, PSVM을 이용해 선택한 단백질군을 심혈관 질환 진단을 위한 바이오마커 후보로 제시한다.

1. 서 론

바이오마커란, 특정 질병 상태에 대한 지표가 되는 생체내 성분이다. 마이크로어레이(microarray) 기술이 소개된 이후[1-3] 유전자(gene) 수준에서 바이오마커를 탐색하는 연구는 생물정보학(bioinformatics)/의료정보학(medical-informatics)의 주된 연구 대상 중 하나가 되었다. 즉, 여러 종류의 샘플에 대해 수천 개의 유전자의 발현 패턴을 종합적으로 분석하여, 샘플의 종류를 구분해내는 데 있어 많은 정보를 포함하고 있는 유전자군을 골라내고, 바이오마커의 후보군인 이 유전자 군에 대해 추가로 생물학적/의학적 분석을 통해 검증하는 과정이 유용한 프로토콜로 자리잡고 있다.

마이크로어레이 데이터에서 유전자군을 골라내는 방법으로는 기계학습(machine learning) 분야에서 자질 선택(feature selection)이라 불리는 기법이 많이 사용된다. 자질 선택 기법은 통계적 패턴 인식, 기계학습, 데이터 마이닝 등의 분야에서 핵심 연구 주제 중 하나이며, 텍스트 분류, 이미지 추출, 고객관리, 유전데이터 분석 등의 응용분야와 함께 많은 기법들이 제시되고 있다[4]. 마이크로어레이 데이터를 이용한 마커 유전자(marker gene) 탐색 연구는 특히 질병 진단을 목표로 많은 연구가 수행되고 있으며, 간암[5], 폐암[6], 위장암[7] 등 암과 관련된 마커 탐색에 특히 많이 활용되고 있다. 최근에는 기존의 DNA 마이크로어레이 외에 ChIP(chromatin immunoprecipitation) 마이크로어레이[8], miRNA 마이크로어레이[9], 단백질 마이크로어레이[10] 등 마이크로어레이 기술의 다변화와 함께, 보다 다양한 각도에서의 바이오마커 탐색이 가능하게 되었다.

본 논문에서는 단백질 마이크로어레이의 일종인 압타머칩(aptamer-chip)을 이용하여 심혈관 질환 진단을 위한 단백질 마커(protein marker)를 탐색하고자 한다. 압타머칩을 이용한 심혈관 질환 진단 가능성은 [11]에서 시험적으로 살펴본 바 있으며, 본 논문에서는 보다 많은 샘플에 보다 우수한 기계학습 기법을 적용하여 질병 분류 뿐만 아니라 마커 탐색을 수행한 결과를 정리한다. 마커 탐색을 위한 자질 선택 기법으로는 최근 소개된 Potential Support Vector Machine(PSVM)을 선정하였다. PSVM은 기존 SVM과 비교하여 볼 때 분류기(classifier)로서 뿐만 아니라 자질 선택기(feature selector)로서도 뛰어난 성능을 보이는 알고리즘으로 알려져 있다[12].

PSVM을 이용하여 4단계로 구분된 심혈관 질환 샘플 135개의 3K 압타머칩(ABB, (주)제노프라) 데이터 분석 결과, state-of-the-art 자질 선택기법 중 하나인 Gain Ratio와 비교하여, PSVM은 보다 적은 자질군(feature set)으로도 보다 높은 분류 정확도(classification accuracy)를 얻을 수 있었다.

본 논문에서는 이 결과를 다음과 같은 순서로 정리한다. 2절에서는 관련 연구로서 압타머칩, 자질선택 기법 및 PSVM에 대해 정리하고, 3절에서는 분석에 사용된 데이터 및 전처리 과정, 분석목표 및 계획을 정리하고 4절에서 실험결과를 정리하여 분석을 병행한 후 5절에서 결론과 함께 차후 연구과제를 정리한다.

2. 관련 연구

2.1. 압타머칩

압타머(aptamer)란 단일염기서열인 DNA 나 RNA로 항원 항체(body-antibody) 반응과 같이 타겟 물질에 대한 특별한 친화력과 특이성을 나타내는 생체정보 감지 소재이다. 일반적으로 단일 핵산 가닥이 직선형(linear)인 것과 달리 압타머는 복잡한 3차원적 구조를 가지기 때문에, 이러한 타겟 특이성이 나타나게 된다. (주)제노프라는 여러조합의 RNA library에서 특이 질병에 관여하는 RNA 분자를 찾아내고 분리해 낼 수 있는 시험관증폭선택법(Selection Evolution of Ligands by Exponential Enrichment, SELEX)이라는 기술을 이용하여 표적 단백질과 특이적으로 결합하는 RNA 압타머를 선별해냈다. 이러한 RNA 압타머와 상보적으로 결합하는 capture DNA (anti-aptamer sequence)를 글라스 표면에 고정화하여 마이크로어레이화한 것이 압타머칩이다. 이러한 압타머칩을 통해 대량의 표적단백질의 패턴을 파악할 수 있으며, 2006년 초에 3,000개의 단백질 패턴을 하나의 칩으로 확인할 수 있는 3K 압타머칩이 개발되었다.

압타머칩은 현재 혈액 내 단백질 양의 변화를 통해 질병 유무를 판별하고, 질병 특이적인 단백질을 탐색하는 데 활용되고 있다.

2.2. 자질 선택(Feature Selection)

2.2.1. 개요

통계적 패턴 인식, 기계학습, 데이터 마이닝 등에서 다루는 대부분의 데이터는 일반적으로 수백~수천 이상의 자질로 설명되며, 문제의 복잡도를 줄이고(차원 축소) 비용을 줄이기 위해 의미있는 자질군을 선별하는 전처리 작업은 필수적이다. 자질군 선별기법은 크게 자질 선택(feature selection)과 자질 생성(feature construction)의 두 부류로 나뉜다. 자질 생성은 기존 자질의 조합을 통해 새로운 자질을 계산하는 방법이며, 문제공간의 차원 축소(dimensionality reduction)를 위해 많이 사용된다. PCA(principal component analysis), ICA(independent component analysis)는 기존 자질의 선형 조합을 통해 자질군을 생성하는 대표적인 기법이며, 커널 기법(kernel method) 및 정보병목기법(information bottleneck method)을 이용한 자질 생성 기법도 연구되고 있다. 반면, 자질 선택은 입력 자질 중에서 문제해결에 연관성이 있다고 판단되는 '부분집합'을 선택하는 방법이며, 선별된 자질군을 적절히 조합하여 문제를 해결하는 작업, 예를 들면 분류작업(classification)이나 군집화(clustering)는 다음 단계에 맡기게 된다.

자질 선택 기법은 각 자질에 순위를 매긴 후 상위 자질을 선택하거나 또는 자질 간의 연관성을 고려하여 자질군 부분집합을 선별하는 작업을 수행한다. 순위기반 자질선택 기법에서는 자질 간의 상관관계를 고려하지 않는 경우가 많은 반면, 자질군 선택 기법은 순위기반 관점에서는 정보가 적더라도 다른 자질과의 관계성을 고려하여 자질을 선택한다.

2.2.2. 자질선택기법의 분류

자질선택기법은 크게 필터(filter)기법과 포장(wrapper)기법으로 구분된다[4][12][13]. 필터 기법은 별도의 마이닝 기법 없이 데이터의 일반적은 특성을 기반으로 자질을 평가하고 선택하는 반면, 포장 기법은 지정된 마이닝 기법의 성능향상 기여도를 기준으로 자질을 평가한다[5]. 보통 포장 기법이 필터 기법에 비해 계산에 더 많은 비용을 필요로 한다. 이 외에 두 기법을 별도의 단계에 두어 차례로 수행하는 혼합 기법도 존재한다.

질병 진단과 같은 분류(classification) 문제의 경우를 예로 들면[12] 필터 기법에서는 별도의 분류기(classifier)에 의존하지 않고 다만 '클래스 변수'와 의존성이 높은 자질을 선택한다. 대표적인 필터 기법으로는 통계적 기법을 들 수 있으며, 클래스 변수와 각 자질 간의 의존성을 피어슨 상관계수, 윌콕슨 통계량(Wilcoxon statistics), t-통계량, SNR(signal-to-noise ratio) 등을 기준으로 평가한다. 포장 기법에서는 분류기의 분류성능을 목적함수(objective function), 즉 평가 기준으로 두어 자질을 선택한다. 학습데이터(training data)를 기준으로 분류오류를 최소화하는 모델 선별(학습) 과정을 거쳐 '분류기'를 얻은 후, 별도의 평가데이터(validation set)에 대한 분류기의 예측성능을 측정한다. 이 때, 학습데이터/평가데이터에서 사용되는 자질의 선별/제거에 따른 예측성능의 변화가 자질평가의 기준이 된다. 각 기법의 다양한 예는 [4][12][13]을 참조하기 바란다.

2.3. 마이크로어레이에서의 자질 선택

마이크로어레이에서 자질 선택을 하는 목적은 다음과 같이 세 가지로 요약된다[12]:

(가) 기계학습 기법의 예측성능 향상을 위한 데이터 전처리

(나) 지표(indicator) 마커 판별. 데이터의 해석과 이해를 돕기 위해 필요

(다) 비용 절감. 마이크로어레이 데이터가 진단 목적으로 사용되는 경우

(가)항목은 샘플의 수에 비해 자질의 수가 큰 마이크로어레이의 특성과 관련이 있다. 자질의 수가 많은 경우는 "차원의 저주(curse of dimensionality)"라 불리며, 예측성능의 일반화는 어렵게 된다. (나)항목은 샘플의 클래스에 따라 발현패턴이 바뀌는 자질을 선택하는 과정을 의미한다. 대조군과 실험군에서 다른 패턴을 보이는 자질을 선택한 후 해당 자질과 관련된 세포 수준의 기작이나 생체내의 패스웨이(pathway)를 밝히는 연구, 또는 실험군과 대조군의 차이를 유발하거나 줄이는 목표 자질(유전자 또는 단백질)을 탐색하는 연구가 후속작업으로 진행이 될 수 있다. (다)항목과 관련해서는 적은 수의 프로브(probe)를 사용한 저렴한 칩, 실험 대상군을 줄임으로서 인력 절감, 실험결과 해석의 용이성 등이 세부 목적으로 볼 수 있다.

기존에 많이 활용된 cDNA/올리고 마이크로어레이의 경우, 자질 선택의 의미는 유전자(gene) 선택을 의미하지만, 이 논문에서 분석하고자 하는 압타머칩의 경우 혈청 내의 '단백질(protein)' 선택을 의미한다.

2.4. Potential Support Vector Machine (PSVM)

2.4.1. PSVM 개요 및 특징

PSVM은 최근 소개된([12],[14]) 기법으로서, 분류, 회귀뿐만 아니라 변수 선택에도 사용되는 최대 경계면 탐색기법(large margin method)이다. 기존 SVM[15]과 차이점은 목적함수(objective function)와 최적화시의 제한조건(constraints)을 모두 새롭게 변경하여 데이터의 스케일 변화에도 보다 안정적으로 낮은 오류(empirical error)로 최적해를 찾도록 하였다는 점이다. 기존 SVM과 비교되는 PSVM의 큰 특징 중 하나는, negative-definite하고(기존 SVM은 positive semi-definite를 요구) 정방형이 아닌 행렬도 처리할 수 있다는 점이다. 표준적인 SVM에서는 분류/회귀 함수를 'support vector'라 부르는 데이터 부분집합으로 표현된다. PSVM에서는 데이터 행렬 자체를 커널행렬(이른바 Gram matrix)로 해석하여 변수와 데이터의 역할을 뒤바꿈으로써 변수(자질) 선택을 가능하게 한다. 경계면 함수에서의 가중치 벡터 w는 적은 수의 'support variables'의 선형 조합으로 표현이 되며, 이 변수들이 '의미있는' 자질로 선택된다.

PSVM을 자질 선택을 위한 필터 기법으로 적용하는 방법은 [16]에서 제시되었으며, 'NIPS 2003 feature selection challenge'에서 제시된 다른 자질 선택 기법과 벤치마킹한 결과 간결한 자질군을 선택하는 가장 좋은 방법 중 하나로서 판별되었다. PSVM을 통해 선정된 'support variables'는 다음 단계로서 임의의 분류기의 입력으로 사용할 수 있다.

2.4.2. PSVM의 수학적 표현 [17]

이 절에서는 PSVM알고리즘의 수학적 표현을 간결히 정리한다. 상세한 설명은 [12][14]를 참고하기 바란다.

목표문제는 이진 클래스 분류 문제이다. m개의 입력 벡터(d차원)과 클래스 레이블(label)은 각각 행렬 $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_m)$ 및 벡터 $\mathbf{y}$로 표기한다. 학습목표는 가중치 벡터 $\mathbf{w}$와 오프셋 b를 인자로 표현되는 분류기 집합

$$g(\mathbf{x}) = sign(\mathbf{w}\cdot\mathbf{x}+b). \quad (1)$$

중에서 risk functional R을 최소로 하는 하나의 분류기 g를 선택하는 것이다. 데이터 벡터 각각에 표준화(standardization, 즉 평균을 0, 표준편차를 1로 조정)를 적용한 후, 자질 선택을 위한 primal PSVM 최적화 문제는 다음의 수식으로 표현된다:

$$\min_{\mathbf{w}} \tfrac{1}{2}\left\|\mathbf{X}^T\mathbf{w}\right\|^2, \quad (2)$$

제한조건:

$$\mathbf{X}^T(\mathbf{X}^T\mathbf{w}-\mathbf{y})+\varepsilon\mathbf{1}\geq\mathbf{0}$$
$$\mathbf{X}^T(\mathbf{X}^T\mathbf{w}-\mathbf{y})-\varepsilon\mathbf{1}\leq\mathbf{0}. \quad (3)$$

이 최적화 문제의 dual 문제(Wolfe dual)는 다음과 같이 표현된다:

$$\min_{\boldsymbol{\alpha}^+,\boldsymbol{\alpha}^-} \tfrac{1}{2}(\boldsymbol{\alpha}^+-\boldsymbol{\alpha}^-)^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\alpha}^+-\boldsymbol{\alpha}^-)-\mathbf{y}^T\mathbf{X}(\boldsymbol{\alpha}^+-\boldsymbol{\alpha}^-)+\varepsilon\mathbf{1}^T(\boldsymbol{\alpha}^+-\boldsymbol{\alpha}^-) \quad (4)$$

제한조건:

$$\mathbf{0}\leq\boldsymbol{\alpha}^+,\mathbf{0}\leq\boldsymbol{\alpha}^-. \quad (5)$$

여기에서 ε은 support 변수의 수를 결정하는 인자이다(큰 값을 부여할수록 선택되는 변수의 수는 줄어든다). 두 벡터 $\boldsymbol{\alpha}^+$ $\boldsymbol{\alpha}^-$는 제한조건에 대한 Lagrange multiplier이다. 0이 아닌 성분 α_i 가 support 변수를 결정한다.

식 (4)는 새로운 SMO(sequential minimum optimization) 방법으로 풀 수 있다. $\boldsymbol{\alpha}=\boldsymbol{\alpha}^+-\boldsymbol{\alpha}^-$로 설정하면, 가중치 벡터 $\mathbf{w}$와 오프셋 b는 다음과 같이 결정된다:

$$\mathbf{w}=\boldsymbol{\alpha}, \quad b=\frac{1}{m}\sum_{i=1}^{m} y_i, \quad (6)$$

최종적으로 선택되는 분류기는 다음과 같이 주어진다:

$$g(\mathbf{x})=sign(\mathbf{w}\cdot\mathbf{x}+b)$$
$$=sign\left(\sum_{j=1}^{d}\alpha_j(\mathbf{x}\cdot\mathbf{e}_j)+b\right). \quad (7)$$

2.4.3. PSVM을 이용한 자질 선택

PSVM 기법이 가중치 벡터 $\mathbf{w}$를 0이 아닌 소수의 'support feature' 집합의 조합으로 표현하는 특성을 이용하여, 샘플의 클래스에 대해 '정보를 제공하는(informative)' 자질 선택에 최적화된 알고리즘을 얻을 수 있다[14]. 식 (3)에서 도입된 ε은 노이즈의 영향을 받은 자질이 분류에 영향을 주는 것을 방지하는 '상관관계 임계값(correlation threshold)'의 역할을 한다. 식 (3)의 첫 번째 항이 ε보다 작은 값을 가지도록 하는 자질군은 샘플 클래스와의 관계가 노이즈로 인해 임의로 발생한 것으로 판단하며, (3)의 조건에 의해 $\mathbf{w}$에서 해당 자질의 성분은 0이 된다. 최종적으로 'support feature' 집합을 자질로서 선택하게 된다. ε값은 모델 선택 기법에 의해 정할 초인자(hyper parameter)가 된다.

3. 데이터 및 분석방법

3.1. 데이터

본 논문에서 분석할 데이터는 50대 남성 심혈관 질환 관련 샘플로부터 (주)제노프라의 3K 압타머칩(ABB)으로 생성한 3000×135 크기의 2차원 행렬 형태 데이터이다. 샘플에는 [표 1]과 같이 심혈관 질환의 진행 정도에 따라 4가지의 클래스 레이블이 부여되어 있다.

최초 데이터의 형태는 cDNA 마이크로어레이와 동일하며, 각 샘플별로 GenePix 스캐너에서 생성한 'gpr'파일 형태이다.

표 1. 분석 대상 데이터의 클래스별 샘플 수. Normal은 정상인을 의미하며, 심혈관 질환이 진행됨에 따라 SA(안정형 협심증), UA(비안정형 협심증), MI(심근경색)로 표지가 되어 있다.

클래스 구 분	Normal	SA (stable angina)	UA (unstable angina)	MI (myocardial infarction)	합계
샘플수	40	38	29	28	135

3.2. 데이터 전처리 및 학습데이터 구성

데이터는 다음의 프로토콜에 따라 전처리를 하였다.

(가) 'Ratio of Medians' 필드 추출 및 품질 조정(quality control) : 각 샘플의 gpr 파일에서 Cy5/Cy3의 median 값을 추출하되, 각 단백질별(즉, spot)로 스캐너가 인지하지 못하였거나, background값이 foreground값보다 큰 경우 missing value로 처리한다.

(나) imputation: 전단계에서 생성한 행렬 데이터에서 missing value는 각 샘플의 중앙값(median)으로 채운다.

(다) 선형 정규화: [18]의 권장사항을 반영하여 각 샘플의 평균값을 1로 조정하여 Cy5와 Cy3채널의 영향을 동일하게 고려한다.

(라) 로그변환: 행렬 전체에 대해 밑을 2로 하는 로그를 취하여 Cy5/Cy3를 log(Cy5)-log(Cy3)와 같이 변환한다.

위와 같이 전처리한 데이터에 대해 PSVM 및 비교 대상 자질 선택기법을 적용하기 위하여 [그림 1] 및 [표 2]와 같이 학습/테스트 데이터셋을 구성한다.

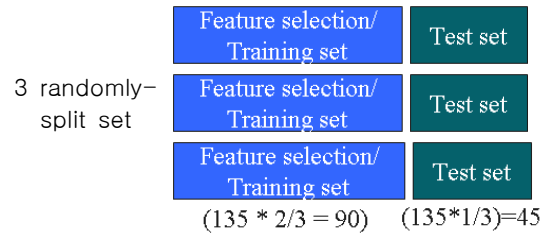

그림 1. 실험데이터 구성. 135개 샘플을 랜덤하게 분리하여 2/3을 학습데이터로 구성하여 자질 선택 알고리즘 적용 및 선택된 자질군을 이용한 분류기 학습을 수행하며, 1/3을 테스트데이터로 구성하여, 분류기의 성능을 측정한다.

표 2. 자질 선택 및 분류 알고리즘을 적용하기 위한 데이터셋 구성. 랜덤하게 샘플 중 2/3=90개를 학습 데이터(TR)로, 1/3=40개를 테스트 데이터(TE)로 구성하며, 세 개의 랜덤셋을 생성하여 1~3까지 일련번호를 부여한다.

데이터셋 구 분	Normal	SA	UA	MI	합계
TR1/TE1	25/15	26/12	19/10	20/8	135
TR2/TE2	25/15	30/8	20/9	15/13	135
TR3/TE3	25/15	24/14	21/8	20/8	135

3.3. 분석 방법

학습데이터에 대해 자질 선택 방법을 적용하여 자질군을 선택한 후, 해당 자질군만을 이용하여 학습데이터로 분류기를 학습하며, 테스트데이터로 분류기의 성능을 측정한다. 3가지의 데이터 셋에 동일한 과정을 적용하여, 테스트데이터에서의 분

류정확도를 기준으로 자질선택 알고리즘의 평가한다. 분류 알고리즘으로는 나이브베이즈(naïve Bayes), *k*NN(*k*-nearest-neighbor)을 적용한다. *k*NN에서 *k*의 값은 초인자(hyperparameter)로서 LOOCV (leave-one-out cross validation) 방식으로 결정한다.

3.4. 자질선택 알고리즘 적용

3.4.1. PSVM

PSVM의 최적화 문제 설정에서 클래스 레이블 값이 이진이 아닌 실수값을 가지는 경우, 자연스럽게 PSVM최적화 문제는 최적의 회귀함수 탐색 문제가 된다[14]. 회귀모드의 PSVM에서는 실수값을 가지는 예측대상 변수는 표준화(standardization)를 적용후 입력하여야 한다. 현재 데이터의 클래스 레이블은 이진이 아니며, 각 단계별로 선후관계가 명백하기 때문에, PSVM을 회귀분석 모드로 실행하여 자질 선택을 수행한다. 4단계의 클래스 변수를 실수로 매핑하는 방법은 다음의 두 가지를 적용한다:

Encoding 1 - Normal부터 MI까지 순서대로 0,1,2,3 부여
Encoding 2 - Normal부터 MI까지 순서대로 1,2,4,8 부여

PSVM을 이용한 자질 선택에는 [표 3]과 같이 초인자 ε을 최적화하기 위한 LOOCV 및 최적인자를 적용하여 자질평가를 하는 두 수준의 LOOCV를 적용한다. 최종 선택된 자질군의 각 자질에는 LOOCV 수행 중 해당 자질이 선택된 빈도(현재 데이터에 대해 최대값은 학습데이터 수에 해당하는 90)가 부여되며, 빈도가 높을수록 보다 정보가 많은 자질이라 판단할 수 있다.

표 3. PSVM을 이용한 자질 선택에 적용한 모델 선택기법

외부 LOOCV
내부 LOOCV
hyperparameter 선택
ε : 1.0~3.0, 0.1 단위로 시험
test point에 대해 오류(MSE)를 최소화하는 ε 선택
최적의 ε을 적용하여 test point에 테스트 수행
자질이 선택되는 빈도를 기준으로 순위 부여

3.4.2. Gain Ratio

Information Gain 및 Gain Ratio는 각 자질별로 클래스 레이블에 대한 정보량(information)을 측정하는 척도이며, 정의는 다음과 같다(H(X)는 변수 X의 정보 엔트로피(entropy)이다):

$$\text{InfoGain(Class, Feature)} = H(\text{Class}) - H(\text{Class}|\text{Feature}). \quad (8)$$

$$\text{GainRatio(Class, Feature)} = \text{InfoGain(Class, Feature)} / H(\text{Feature}) \quad (9)$$

각 자질의 gain ratio 값을 10-fold CV 결과 평균값을 기준으로 상위 자질을 선택한다.

4. 결과 및 분석

4.1. 자질 선택

세 랜덤분할 데이터셋 각각에 대해 PSVM은 3000개의 자질(단백질) 중 53~63개의 자질을 선택하였다(표 5). 각 분할셋별 선택된 자질 중 빈도가 10%(9회) 미만인 자질을 제외하고 중복 여부를 살펴본 결과, [표 4]와 같이 중복된 자질은 많지 않다(TR1-TR2 중복자질은 5개, TR1-TR3 중복자질은 8개).

표 4. PSVM을 통해 선택된 자질군. 클래수 변수에는 Encoding 1을 적용하고(3.4.1절 참조), 빈도를 기준으로 10% 미만(9 미만)인 자질은 제외한 목록이다. [그림 1]에서와 생성한 세 가지의 학습 데이터 각각에서 선택된 자질(단백질)의 ID 및 빈도를 표기하였다. 단백질ID는 칩 상에서 spot의 위치에 해당하며, 단백질 상세 정보는 수록하지 않는다.

TR1		TR2		TR3	
FeatureID	빈도	FeatureID	빈도	FeatureID	빈도
P61	20	P131	69	P146	16
P71	33	P146	31	P147	89
P259	74	P480	26	P259	90
P394	90	P696	43	P305	90
P497	20	P757	32	P394	69
P498	54	P860	76	P463	12
P505	32	P895	40	P516	63
P516	90	P968	90	P709	68
P548	30	P1122	9	P731	77
P649	20	P1146	13	P753	87
P731	21	P1216	90	P795	11
P838	44	P1256	71	P906	9
P860	16	P1299	47	P926	89
P926	90	P1417	24	P1216	90
P1122	37	P1595	88	P1257	12
P1216	25	P1809	29	P1299	24
P1299	31	P1858	85	P1430	30
P1430	75	P1875	29	P1431	40
P1634	83	P1905	25	P1546	88
P1663	90	P2221	53	P1595	18
P1828	38	P2336	29	P1604	90
P2057	56	P2653	11	P1651	13
P2091	11	P2749	14	P2413	90
P2099	21	P2837	32	P2417	86
P2132	72			P2418	17
P2221	90			P2535	72
P2453	13				
P2553	14				
P2719	54				
P2944	83				

4.2. 선택된 자질군을 적용한 분류

각 데이터셋 별로 학습데이터를 통해 추출한 자질만을 이용하여 학습데이터에서 분류기(classifier)를 학습한 후, 테스트데이터에서 일반화성능을 측정한 결과는 [표 5], [표 6]과 같다.

[표 4]에서, 나이브베이즈 분류기의 경우 PSVM을 통해 선택된 자질군이 Gain Ratio(GR)기법으로 뽑은 자질의 수와 유사하거나 적은 수로, 분류성능이 GR의 경우 이상으로 나타나고 있다. kNN 분류기의 경우는 'Encoding2-TR/TE2'의 경우를 제외하면 PSVM이 선택한 자질군이 GR보다 낫다고 보기 힘들다. 그러나, PSVM이 선택한 자질군에서 빈도가 낮은 자질을 제외하고 살펴본 결과 [표 5]에서와 같이, 'TR/TE2, TR/TE3'의 경우는 GR의 경우에 비해 훨씬 적은 수의 자질로 비슷하거나 더 높은 분류 정확도를 보이고 있다. 'TR/TE1' 데이터셋의 경우는 PSVM으로는 좋은 품질의 자질군을 선택하지 못한다고 판단되며, PSVM의 특성과 이 데이터셋의 특성에 대한 상세분석은 추후 연구 대상이다.

클래스 레이블에 대한 실수매핑방식을 분류 정확도를 기준으로 비교해보면, 대체로 'Encoding 2' 방식이 'Encoding 1'방식보다 좋은 것으로 판단된다. 심혈관 질환의 각 단계 간의 '차이'가 단순 증가가 아닌 최적화의 대상이 될 수 있음을 살펴볼 수 있다.

표 5. PSVM 및 비교 대상 기법으로 자질선택 후 테스트데이터를 이용하여 분류한 결과. 나이브베이즈, kNN의 분류결과는 분류 정확도(accuracy)의 백분율값이다.

Feature Selector	Data Set	# features	Naïve Bayes	kNN (LOOCV for k)
PSVM_all (Encoding 1)	TR/TE1	54	75.56	66.67
	TR/TE2	60	75.56	75.56
	TR/TE3	63	75.56	71.11
PSVM_all (Encoding 2)	TR/TE1	61	84.44	71.11
	TR/TE2	53	71.11	88.89
	TR/TE3	63	75.56	75.56
GR50 (Gain Ratio, 50 features)	TR/TE1	50	71.11	80
	TR/TE2	50	71.11	75.56
	TR/TE3	50	66.67	75.56
GR100	TR/TE1	100	73.33	86.67
	TR/TE2	100	73.33	84.44
	TR/TE3	100	73.33	84.44

표 6. PSVM으로 선택한 자질군에서 각 자질에 부여된 빈도가 지정된 값(10% cut의 경우 9, 5% cut의 경우 5)보다 작은 경우 제외한 새로운 자질군에 대한 분류결과 비교.

Feature Selector	Data Set	# features	Naïve Bayes	kNN (LOOCV for k)
PSVM_10%cut (Encoding 1)	TR/TE1	30	64.44	55.56
	TR/TE2	24	73.33	75.56
	TR/TE3	26	68.89	64.44
PSVM_5%cut (Encoding 2)	TR/TE1	34	77.78	60
	TR/TE2	25	68.89	73.33
	TR/TE3	25	73.33	73.33
GR50 (Gain Ratio, 50 features)	TR/TE1	50	71.11	80
	TR/TE2	50	71.11	75.56
	TR/TE3	50	66.67	75.56

4.3. 심혈관 질환 진단 바이오마커 후보

4.2절의 결과를 바탕으로 PSVM을 이용한 자질 선택이 GR과 비교해볼 때 상대적으로 작고 우수한 자질군 선택을 가능케 할 수 있다는 한 가지 증거를 얻었다. 앞의 결과를 바탕으로, 클래스 변수에는 'Encoding 2'를 적용한 후 135개 샘플 전체를 이용하여 PSVM 자질 선택을 수행하고, 빈도 기준 10% 미만인 자질을 제외하여 얻은 다음의 단백질군(37개)을 심혈관 질환 진단을 위한 바이오마커 후보로 제시한다.

표 7. PSVM으로 선택한 심혈관 질환 진단 바이오마커 후보

P71, P131, P259, P394, P505, P516, P649, P712, P731, P832, P846, P926, P1110, P1125, P1146, P1216, P1257, P1299, P1373, P1503, P1634, P1852, P1858, P1882, P1915, P2004, P2095, P2099, P2221, P2389, P2453, P2540, P2590, P2763, P2812, P2857, P2944

5. 맺음말

본 논문에서는 마이크로어레이 데이터에서의 자질 선택 도구로서의 Potential SVM(PSVM)에 대해 소개하고, 압타머칩 데이터로부터 심혈관 질환의 바이오마커 후보 단백질 선정에 PSVM을 적용한 결과를 정리하였다. PSVM은 자질간 관계를 고려하여 성기고(sparse) 중복을 최소화하는 자질 집합을 제시하기 때문에, 중복된 정보를 고려하지 못하는 많은 자질 선택 기법과 비교하여 마이크로어레이 데이터를 이용한 바이오마커 선택에 적절한 기법이다. 3K압타머칩 데이터에 PSVM을 적용한 결과 PSVM이 다른 자질 선택 기법에 비해 적은 수의 자질로도 유사하거나 더 나은 심혈관질환 분류 성능을 보임을 확인하였으며, PSVM이 선택한 자질 단백질 심혈관질환 진단을 위한 바이오마커로서 제시하였다.

본 논문에서는 PSVM과 다른 state-of-the-art 마이크로어레이 자질 선택기법과의 비교가 충분히 이루어지지 않았으며, PSVM을 회귀모드로 실행하는 과정에서 클래스 레이블의 실수로의 매핑 방법에 대한 보다 다양한 경우를 살펴보지 못하였다. 추후 연구 과제로서 SNR(signal-to-noise ratio), RFE(recursive feature elimination)-SVM과 같은 state-of-the-art 기법과의 비교를 보다 엄밀히 수행할 것이며, PSVM을 통해 선택된 바이오마커 후보에 대한 생물학적 검증도 차후의 과제로 남겨둔다.

감사의 글

PSVM 실행 코드를 제공하고, 분석관련 조언을 해준 Technische Universität Berlin, Neural Information Processing Group의 Johannes Mohr와 Prof. Klaus Obermayer에게 감사드립니다.

이 논문은 과학기술부 국가지정연구실사업(NRL), Brain Korea 21 사업에 의하여 지원되었음을 밝힙니다. 이 연구를 위해 연구장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소(ICT)에 감사드립니다.

참고문헌

[1] Southern E., United Kingdom patent application GB8810400, 1988.

[2] Lysov, Y., Florent'ev, V., Khorlin A., Khrapko K., Shik V., and Mirabekov A., DNA sequencing by hybridization with oligonucleotides, *Doklady Academy Nauk USSR*, **303**:1508-1511, 1988.

[3] Bains W. and Smith G., A novel method for nucleic acid sequence determination., *Journal of Theoretical Biology*, **135**:303-307, 1988.

[4] Liu H. and Yu L., Toward integrating feature selection Algorithms for classification and clustering, *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491-502, 2005.

[5] Smith M. W., *et al.*, Identification of novel tumor markers in hepatitis C virus-associated hepatocellular carcinoma, *Cancer Res.*, **63**(4):859-64, 2003.

[6] Jiang H. et al., Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes, *BMC Bioinformatics*, 5:81, 2004.

[7] Aburatani H., *et al.*, Discovery of a new biomarker for gastroenterological cancers, *J Gastroenterol*, 40 Suppl 16:1-6, 2005.

[8] Ren, B., et al., Genome-wide location and function of DNA binding proteins. *Science*, **290**:2306-2309, 2000.

[9] Liang, R.-Q., et al., An oligonucleotide microarray for microRNA expression analysis based on labeling RNA with quantum dot and nanogold probe, *Nucleic Acids Res*, **33**(2):e17, 2005.

[10] Ramachandran, N. et al., Self-assembling protein microarrays, *Science*, **305**:86-90, 2004.

[11] 김병희, 김성천, 장병탁, 기계학습에 의한 압타머칩 데이터 기반 심혈관 질환 단계의 예측, *한국컴퓨터종합학술대회 2006 논문집*,

제33권 1(A), pp. 85-87, 2006.

[12] Hochreiter S. and Obermayer K., Kernel Methods in Computational Biology, chapter Gene Selection for Microarray Data, pp. 319-356. Eds.: Schölkopf B., Tsuda K. and Vert J.-P., MIT Press, Cambridge, Massachusetts, 2004.

[13] Guyon I. and Elisseeff A., An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, 3:1157-1182, 2003.

[14] Hochreiter J. and Obermayer K., Support vector machines for dyadic data, *Neural Comput.*, 18:1472-1510, 2006.

[15] Schölkopf B. and Smola A. J., Learning with kernels-Support Vector Machines, Regluarization, Optimization, and Beyond. MIT Press, Cambridge, 2002.

[16] Hochreiter J. and Obermayer K., "Nonlinear feature selection with the potential support vector machine," in *Feature extraction, Foundations and Applications*, I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds. Springer, 2005.

[17] Mohr J., Puls I., Wrase J., Hochreiter S., Heinz A., and Obermayer K., P-svm variable selection for discovering dependencies between genetic and brain imaging data, In *IJCNN 2006 Conference Proceedings*, 2006. In press.

[18] Damian V., The Axon guide to microarray analysis, Application Note, Axon Instruments, 2004. (http://www.moleculardevices.com)