한국시뮬레이션학회 논문지
Vol. 22, No. 4, pp. 139-147 (2013. 12)
http://dx.doi.org/10.9709/JKSS.2013.22.4.139
ISSN 1225-5904

단백체 스펙트럼 데이터의 분류를 위한 랜덤 포리스트 기반 특성 선택 알고리즘

온승엽[1†] · 지승도[1] · 한미영[2]

Feature Selection for Classification of Mass Spectrometric Proteomic Data Using Random Forest

Syng-Yup Ohn · Seung-Do Chi · Mi-Young Han

ABSTRACT

This paper proposes a novel method for feature selection for mass spectrometric proteomic data based on Random Forest. The method includes an effective preprocessing step to filter a large amount of redundant features with high correlation and applies a tournament strategy to get an optimal feature subset. Experiments on three public datasets, Ovarian 4-3-02, Ovarian 7-8-02 and Prostate shows that the new method achieves high performance comparing with widely used methods and balanced rate of specificity and sensitivity.

Key words : Feature Selection, Bioinformatics, Pattern Recognition, SELDI-TOF, Proteome, Spectrum, Random Forest, Pearson Correlation

요 약

본 논문에서는 질량 분석 방법에 의하여 산출된 단백체 데이터(mass spectrometric proteomic data)의 분류 분석(classification analysis)을 위한 새로운 특성 선택(feature selection) 방법을 제안한다. 이 방법은 i)높은 상관관계를 가지는 중복된 특성을 효과적으로 제거하는 전처리 단계와 ii)토너먼트(tournament) 전략을 사용하여 최적 특성 부분집합(optimal feature subset)을 탐색해 내는 단계로 구성되어 있다. 제안 되는 방법을 실제 암진단에 사용되는 공개된 혈액 단백체 데이터에 적용하였으며 널리 사용되는 타 방법과 비교할 때 우수한 성능과 균형된 특이도와 민감도를 달성함을 실증하였다.

주요어 : 특성 선택, 생물정보학, 패턴인식, 표면 강화한 레이저 탈착과 이온화 시간 의 비행 질량 분석, 단백체, 스펙트럼, 랜덤 포리스트, 피어슨 상관 계수

1. 서 론

특성 선택(feature selection)은 고차원의 데이터 세트를 분석하기 위해서 반드시 필요한 과정이며 특히, 고차원의 복합 데이터를 다루는 빅데이터 분석(big data analysis) 분야에서는 분석 과정의 효율성 및 분석결과의 정확도등을 향상시키기 위해서 매우 중요한 문제이다. 분류 분석을 위한 특성 선택은 선택된 특성 조합에 의하여 학습된 분류기의 예측 성능을 최대화하는 조합을 탐색하는 최적 조합을 선택하는 문제로 정의할 수 있다[1].

특성 선택은 지도학습과 데이터 마이닝 분야에서 주요 연구 주제 중의 하나이며 많은 연구가 수행되어 다양한 방법이 제시 되었다. 특성 선택 방법은 접근 방법에 따라 i) 필터방법(filter approach), ii) 포함방법(wrapper approach) iii) 내재방법(embedded approach)의 세 가지로 분류할 수 있다. 필터방법에서는 특성 선택을 학습 알고리즘과는 별개의 전단계로 간주하며 다양한 통계적 방법을 이용하여 각 특성의 성능을 예측, 비교하여 특성을 선택하게 된

접수일(2013년 8월 22일), 심사일(2013년 11월 8일),
게재 확정일(2013년 11월 27일)
[1)] 한국항공대학교 컴퓨터 공학과
[2)] 한국과학창의재단
주 저 자 : 온승엽
교신저자 : 온승엽
E-mail; syohn@kau.ac.kr

다[2, 3]. 이 방법의 단점으로는 특성 선택 단계와 학습알고리즘의 성능이 연동되어 수행되지 않는다는 점을 들 수 있다. 포함방법에서는 기계 학습 알고리즘으로 하여금 선택된 특성 집합의 성능을 측정하도록 한다. 다양한 분류 집합을 생성하고 각 집합에 기반 한 분류기를 생성하고 분류 정확도를 산출하여 특성 집합의 성능을 측정 및 비교 하고 휴티스틱 전략에 의거하여 반복적으로 개선된 특성 조합을 탐색해 나간다. 이 방법의 단점은 특성의 조합 수의 증가에 따른 계산 복잡도의 증가이다. 따라서, 계산 복잡도를 줄이기 위한 노력으로 준최적 조합 탐색 방법, 필터와 포함방법을 결합하여 사용하는 혼합방법(hybrid method) 등이 제안되었다. 마지막으로 내재방법은 특성 부분집합의 생성과 성능측정이 학습알고리즘에 포함되어 있는 경우이다. 일반적으로 내재 방법은 포함방법보다 수행 속도가 빠르며 필터 방법보다 우수한 편이다.

최근 표면 강화레이저 탈착과 이온화 시간 의 비행 질량 분석(surface-enhanced laser desorption / ionization time-of-flight mass spectrometry, 이하 SELDI-TOF MS로 통칭함.) 방법의 개발로 인간 혈액내의 단백체(proteome) 측정이 용이하게 되었고 단백체의 증감의 변화를 분석함으로서 각종 질병의 진단을 가능하게 하는 성과를 거두었다. Petricoin[4] 등의 결과에 의하면 SELDI-TOF 단백체 스펙트럼에서 건강한 여성군과 초기 난소암까지 포함된 난소암에 걸린 여성군의 혈장 표본을 구분할 수 있는 패턴을 발견하였다. 이 연구 결과는 매우 획기적이고 관련 학계의 많은 주목을 받았다. SELDI-TOF MS 방법에서는 이온화된 펩타이드(ionized peptides)의 질량 대 전하량의 비(m/z)를 x축에 표시하고 이온화된 펩타이드의 상대적인 양을 y축에 표시한 스펙트럼 형식의 그래프를 결과로써 생성한다. m/z 비율은 펩타이드의 질량과 비례하나 서로 다른 펩타이드가 동일한 질량을 가질 수가 있으며 또한, m/z 해상도의 제약 때문에 개별 펩타이드를 확인할 수 없다. 이 그래프는 수만 또는 수십만개의 데이터 점으로 나타내어 질 수 있으므로 SELDI-TOF 데이터 세트는 비교적 적은 표본 수에 비하여 매우 많은 수의 특성을 포함하는 특징을 가진다. 단백체 데이터의 특성 선택 문제를 해결하기 위하여 여러 가지 통계적인 방법 및 기계학습 방법이 시도 되었다[5, 6, 7, 8, 9, 10].

본 논문에서는 랜덤포리스트(random forest)[11]에 기반 한 SELDI-TOF 데이터의 특성 선택을 위한 효과적인 알고리즘을 제안하고자 한다. 이 방법은 두 단계로 구성되어 있으며 첫 번째 단계는 전처리단계로써 중복된 특성 변수를 대량으로 제거하기 위하여 피어슨 상관계수를 적용한다. 두 번째 단계에서는 각 특성의 분류 성능에 대한 등급을 측정하기 위해서 랜덤 포리스트 분류 알고리즘(random forest classification algorithm)으로 생성된 분류기(classifier)의 성능을 토너먼트방식(tournament method)으로 비교한다. 본 논문에서는 제안되는 특성 선택 방법의 성능을 실증하기 위하여 실제의 난소암 및 전립선암 데이터에 방법을 적용하고 결과로 나온 특성 집합의 성능을 측정한 결과를 포함하였다. 기존에 사용되는 다른 방법과 성능을 비교해 보면 균형정확도(balanced accuracy)가 크게 개선되었으며 특이도(specificity)와 민감도(sensitivity)는 약간 개선되는 것으로 나타났다.

본 논문은 다음 과 같이 구성되었다. 2장에서는 랜덤 포리스트와 피어슨 상관 계수에 대하여 간단히 요약되었고 3장에서는 특성 선택 방법이 제안된다. 4장에서는 각각 두 가지의 암 진단을 위한 데이터 세트에 제안된 방법을 적용한 결과를 제시하고와 기존의 방법들과 성능을 비교한다. 마지막으로 5장은 결론이다.

2. 관련 연구

2.1 랜덤 포리스트

랜덤 포리스트는 배깅 방법(bagging method)으로 생성된 복수의 분류회귀 나무(CART: classification and regression tree)들로 구성된 복합 분류기(ensemble classifier)이다. 배깅이란 원래의 데이터 세트에 대하여 대체(replacement)를 허용하는 부트스트래핑(bootstrapping) 샘플추출방법을 적용하여 다수의 표본집합을 구성하고 각 표본 집합을 학습용 데이터 세트로 하여 분류기 집합을 생성하여 복합 분류기를 생성하는 방식이다. 개별의 분류 회귀나무를 생성하는 과정 중, 각 노드(node)에서 가지나누기(split)를 수행할 때 특성 집합 전체를 고려하지 않고 미리 정하여진 수의 무작위로 선택된 작은 부분집합 만을 대상으로 최선의 특성을 탐색하게 된다. 따라서, 랜덤포리스트 알고리즘은 고차원의 데이터에 대해서도 고속으로 분류기를 생성할 수 있게 된다. 가지 나누기의 기준으로는 지니계수(GINI index)가 사용된다. 각 나무는 완전히 성장시키며 가지치기(pruning)는 적용하지 않는다. 복합분류기를 구성하는 분류회귀나무의 개수 및 가지나누기의 대상이 되는 특성부분집합의 크기는 실험에 의하여 최적치를 결정할 수 있다. 랜덤 포리스트의 분류 정확도는 개별 나무 생성을 위하여 부트스트래핑 과정에서 선택되지 않은 샘플(OOB samples: out-of-bag samples)을 테스트 샘플로써 개별나무에 적용하여 측정한다.

2.2 피어슨 상관 계수

피어슨 상관계수[12, 13]는 동일한 객체에 대하여 측정된 두 가지 변수 값의 상관관계를 수치로 나타낸 것이다, 즉, 두 변수가 동반하여 증감하는 경향의 척도를 나타낸다. 피어슨 상관계수는 아래의 공식에 의하여 계산할 수 있다. 아래에서 $X=(x_1,...,x_n)$와 $Y=(y_1,...,y_n)$ 는 각각 n 차원 벡터로서 (x_i, y_i)는 i번째 객체에 대한 두변수의 측정값을 나타낸다.

$$Corr(X, Y) = \frac{Covariance(X, Y)}{\sqrt{Variance(X)\,Variance(Y)}}$$

계수의 값의 범위는 [-1, 1]이며 1의 값은 두 변수의 관계가 완전한 양의 선형관계라는 것을 나타내며 -1의 값은 음의 선형관계를 나타내는 것이다. 0의 값은 두 변수가 선형관계를 가지고 있다고 보기 어려움을 나타낸다. 또한, 양, 음의 1의 값에 가까울수록 강한 양, 음의 선형관계를 가지는 것으로 해석할 수 있다.

상관계수의 특성을 이용하여 중복된 특성을 제거할 수 있다. 상관관계가 변수의 중복성에 어떻게 영향을 미치는가에 대한 연구의 결과가 [14]에 발표되어 있다. 완전한 선형 관계를 보이는 변수들의 집합은 중복된 변수들이라고 볼 수 있고 이러한 변수들은 정보의 손실이 없이 제거가 가능하다. 그러나, 상관관계가 높다할지라도 변수의 상보성(complementarity)이 전혀 없음을 의미하는 것은 아니다. 실제로 상관관계는 필터 방법 중의 하나이며 매우 계산이 간단하여 빠르고 분류기의 종류와는 독립적이므로 유전자 마이크로어레이(microarray) 데이터 분석 같은 초고차원 데이터 등에 자주 적용이 된다.

3. 방 법

본 장에서는 2단계로 구성된 특성 선택 방법이 제안된다. 첫 번째 단계는 전처리 단계로서 피어슨 상관계수를 적용하여 높은 상관계수를 가지는 중복된 특성을 제거되며 제거 되지 않은 특성들은 두 번째 단계의 입력으로 사용된다. 두 번째 단계에서는 랜덤 포리스트을 기본적인 학습 알고리즘으로 하여 특성 조합의 분류 성능을 측정한 결과를 기반으로 분류 성능이 우수한 특성을 토너먼트(tournament) 방식으로 탐색하여 최종적으로 최적 특성 조합을 선택 한다. 각 단계에서 수행하는 내용이 아래에 자세히 나타나 있다.

3.1 피어슨 상관계수를 이용하는 전처리 단계

이 단계에서는 다음 단계에서의 특성 탐색공간을 줄이기 위하여 대량의 중복된 특성을 제거하기 위한 방법으로 특성의 상관관계를 이용한다. 일반적으로 n개의 특성들 간의 상관계수를 계산하기 위해서는 $O(n^2)$의 계산 복잡도가 소요된다. 이러한 계산 복잡도는 특히 방대한 수의 특성을 가지는 스펙트럼 데이터의 경우에는 매우 커진다. 그러나 스펙트럼 데이터의 경우 근접하게 위치한 특성들일수록 높은 상관계수를 가진다는 특징을 이용하면 계산 복잡도를 낮출 수 있다. 즉, SELDI-TOF 스펙트럼 데이터의 경우 협소한 m/z 구간 안에 위치한 특성 들은 거의 동일한 증감 패턴을 가지므로 높은 상관계수를 가진다. 이러한 스펙트럼 데이터의 성질을 이용하여 계산복잡도 $O(n)$으로 대량의 중복 특성들을 제거할 수 있는 알고리즘을 제안한다.

제안된 알고리즘에서는 미리 정하여진 크기의 윈도우를 m/z 축의 왼쪽 끝부터 시작하여 오른쪽으로 스캔하여 나가면서 윈도우 내에 위치한 특성집합에 대하여 계산하여 높은 상관관계를 가지는 중복 특성들에 대하여 한 개의 특성만을 남기고 나머지는 제거하는 방법이다. 아래의 알고리즘 1에 제안된 알고리즘이 자세히 기술되어있다.

알고리즘 1: 높은 상관관계를 가지는 특성 제거를 위한 알고리즘

1. 미리 정하여진 크기 w의 윈도우를 m/z 구간에 위치하고 윈도우에서 첫 번째 특성을 선택한다.
2. 선택된 특성과 윈도우 내부의 다른 특성들과의 상관계수를 계산하고 높은 상관관계를 가지는 특성을 제거한다.
3. 1, 2 단계를 m/z 구간의 왼쪽 끝부터 시작하여 오른쪽으로 w 만큼씩 이동하여 가며 반복한다.

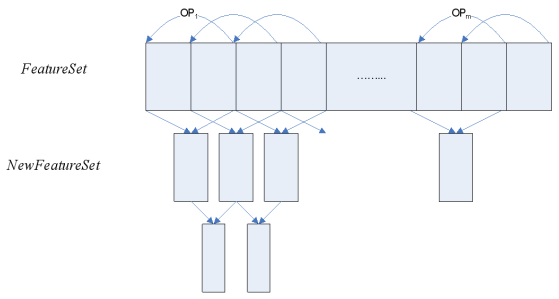

Fig. 1. Tournament strategy for feature selection

알고리즘 1에서 상관계수를 계산하는 횟수는 n이며 스펙트럼 데이터의 경우 특성의 수가 매우 크므로 실제 계산 시간은 원래의 방법에 비하여 크게 감소한다. 또한, 실험을 통하여 이 방법에서 제거된 특성들은 분류성능에 영향을 미치지 못함을 실증하였다.

3.2 랜덤 포리스트를 기반으로 한 토너먼트 특성 선택 단계

본 단계에서는 새로운 특성 선택 전략이 제안된다. 이 알고리즘에서는 전처리 단계에서 선택된 특성 집합으로부터 최적의 특성 조합을 탐색하기 위하여 랜덤 포리스트를 이용하여 특성 조합의 분류 성능을 측정한다. 랜덤 포리스트는 훌륭한 분류 방법일 뿐만 아니라 지니 계수에 기반한 특성의 중요도를 측정하는 좋은 방법이다. 그러나 스펙트럼 데이터의 경우 특성의 수가 수만에서 수십만 개에 이르고 이렇게 특성의 수가 많은 경우에는 랜덤 포리스트가 처리할 수 없거나 좋은 성능을 보여주지 못한다. 이러한 랜덤포리스트의 한계를 극복하기 위하여 분할정복의 전략을 제안한다. 즉, 주어진 특성 집합을 무작위로 m개의 동일한 크기의 부분으로 나누고 랜덤 포리스트 알고리즘을 각 부분에 적용하여 랜덤 포리스트가 산출하는 특성 중요도(feature importance)를 기준으로 각 부분에서 우수한 특성 을 선택한 후 모든 부분에서 산출된 특성을 하나의 집합으로 수집하여 새로운 특성집합을 생성한다. 이러한 과정이 반복 될수록 특성 집합의 크기는 감소되며 특성 집합이 미리 정하여진 크기 및 성능을 가지는 것과 같은 조건이 만족될 때 반복이 종료된다. 이 과정 중에서 무작위의 대체를 허용하는 표본 추출법으로 특성 부분집합을 추출하는 이유는 i)하나의 집합에 우수한 특성이 많을 경우 부분집합간의 중복을 허용하여 이러한 특성이 선택될 기회를 높이기 위한 것과 ii)선택되는 특성이 한정된 구간에 집중되어 국소적인 과적합이 발생하는 것을 막기 위함이다.

알고리즘 2: 랜덤 포리스트에 기반한 특성 선택 전략

1. 주어진 특성 집합으로 부터 대체를 허용하는 무작위 추출 방법으로 p개의 특성을 원소로 가지는 m_i개의 부분 집합을 구성한다. m_i 는 반복횟수 i가 증가할수록 감소한다.
2. 각 부분 집합에 대하여 랜덤 포리스트 모델을 생성하고 높은 중요도를 가지는 q개의 특성을 선택한다.
3. 각 부분집합으로부터 선택된 q개의 특성을 수집하여 한 개의 특성집합을 생성한다.
4. 특성집합의 크기가 미리 정하여진 크기 α가 되거나 미리 정하여진 조건을 만족할 때까지 1 – 3 단계를 반복한다.

최후로 생성된 특성집합이 최적 특성 집합이다.

위의 알고리즘 2에서 4단계의 종료 조건은 특성집합의 크기, 분류성능 등을 단독, 복합적으로 적용할 수 있다.

랜덤 포리스트 알고리즘이 특성의 중요도를 측정하기 위하여 사용되지만 알고리즘 자체가 중복된 특성을 제거하지는 않는다. 일반적으로 중복된 특성들은 랜덤포리스트에서 거의 동일한 중요도를 가진다. 실제로 바이오마커(bio-marker) 탐색의 경우에는 중복된 특성 중에서 한 개만을 선택하도록 하여야한다. 만일, 특성 집합 내에 상관관계가 높은 특성들이 존재할 때에는 단순히 중요도를 임계 하는 방법으로 중요도가 낮은 특성을 제거하는 경우에는 우수한 특성도 제거되는 경우가 발생할 수 있다. 이러한 경우를 줄이기 위하여 본 논문에 제안되는 방법에서는 전처리 단계에서 상관관계가 높은 특성을 미리 제거한다.

4. 실험 결과

3장에서 제안된 특성 선택 방법의 성능을 시험하기 위하여 NCI/CCR 및 FDA/CBER의 Clinical Proteomics Program Databank[15]에 저장되어 있는 Ovarian 4-3-02, Ovarian 7-8-02, Prostate의 세 가지 데이터 세트에 대하여 제안된 알고리즘을 적용하였다. 실험 시 사용한 파라미터 변수의 값은 다음과 같다. 종료조건으로써 구하고자 하는 최적 특성 집합의 크기 $\alpha = 100$, 각각의 특성 부분집합의 크기 $p = 100$, 각 특성 부분집합으로부터 선택할 특성의 수 $q = 50$으로 정하였다. 파라미터의 변수의 값은 여러 번의 실험으로 가장 좋은 결과를 보여주는 값을 선택하였다. 교차검증을 위하여 각 데이터 샘플에 대하여 10회의 무작위 추출에 의한 교차검증을 하였으며 산출된 정확도 척도는 민감도(sensitivity), 특이도(specificity), 양성예측값(positive predictive value), 균형정확도(balaced accuracy)의 평균값들이다. 알고리즘의 성능 평가 결과를 [5, 6]의 결과와 비교하였다. 다음에서 각 데이터세트의 결과를 자세히 기술하였다.

4.1 Ovarian 4-3-02

본 난소암 데이터 세트는 WCX2 단백체 배열(protein array)로 생성되었다. 이 데이터 세트는 216개의 샘플으로 구성되었고, 이중 100개 는 정상, 16개는 양성, 100개

는 암질환의 샘플이다. 각 샘플은 15,154개의 특성을 포함한다. 제안된 특성선택 알고리즘의 전처리 단계에서 6,784개의 특성이 선택되었으며 약 2/3 정도의 특성이 전처리 단계에서 제거된 것이다. 2 단계에서 학습, 시험, 검증 데이트 세트의 수는 [5]에서와 같은 각각 108, 54, 53 이다. Table 1에 [5, 6]에서 제안된 다양한 방법과 제안된 방법에 의하여 구해진 최적 특성 조합의 분류성능 평가 결과가 비교되어 표시되어있다. 이 데이터에 대해서는 다른 방법과 비교하여 우수한 결과를 보이지 않았다.

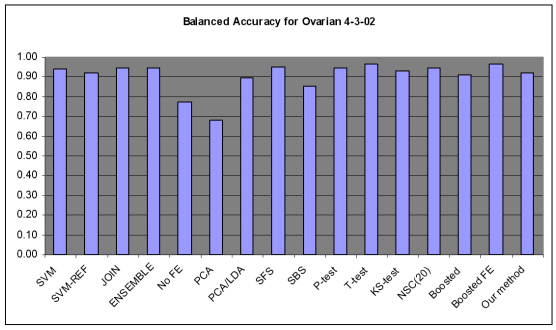

Fig. 2. Balanced accuracy for data set Ovarian 4-3-02

Table 1. Performance on data set Ovarian 4-3-02

	Accuracy (Standard deviation)	Balaced accuracy (Standard deviation)	Specificity (Standard deviation)	Sensitivity (Standard deviation)	Positive predictive value (Standard deviation)
SVM[5]	NA	0.9390	0.9500 (0.0020)	0.9280 (0.0028)	NA
SVM-REF[5]	NA	0.9225	0.9200 (0.0018)	0.9250 (0.0032)	NA
JOIN[5]	NA	0.9440	0.9200 (0.0018)	0.9679 (0.0010)	NA
ENSEMBLE[5]	NA	0.9475	0.9200 (0.0011)	0.9750 (0.0010)	NA
No FE[6]	0.773 (0.09)	0.773 (0.09)	0.828 (0.02)	0.717 (0.18)	0.800 (0.05)
PCA[6]	0.682 (0.18)	0.682 (0.18)	0.687 (0.14)	0.677 (0.25)	0.671 (0.18)
PCA/LDA[6]	0.899 (0.02)	0.899 (0.02)	0.889 (0.10)	0.909 (0.06)	0.900 (0.09)
SFS[6]	0.949 (0.03)	0.949 (0.03)	0.980 (0.03)	0.919 (0.05)	0.979 (0.04)
SBS[6]	0.854 (0.15)	0.854 (0.15)	0.929 (0.08)	0.778 (0.23)	0.903 (0.12)
P-test[6]	0.944 (0.03)	0.944 (0.03)	0.970 (0.03)	0.919 (0.06)	0.969 (0.03)
T-test[6]	0.965 (0.02)	0.965 (0.02)	0.949 (0.05)	0.980 (0.02)	0.953 (0.04)
KS-test[6]	0.929 (0.02)	0.929 (0.02)	0.970 (0.03)	0.889 (0.05)	0.968 (0.03)
NSC(20)[6]	0.944 (0.04)	0.944 (0.04)	0.990 (0.02)	0.899 (0.08)	0.989 (0.02)
Boosted[6]	0.914 (0.06)	0.914 (0.06)	1.000 (0.00)	0.828 (0.12)	1.000 (0.00)
Boosted FE[6]	0.965 (0.01)	0.965 (0.01)	1.000 (0.00)	0.929 (0.02)	1.000 (0.00)
Suggested method	0.922 (0.04)	0.924 (0.04)	0.923 (0.06)	0.925 (0.06)	0.939 (0.04)

4.2 Ovarian 8-7-02

이 스펙트럼 데이터는 WCX2 칩에 의하여 생성되었다. 데이터세트는 253개의 샘플로 구성되어 있으며 이중 91개는 정상, 162개는 난소암 샘플이다. 이 샘플은 처리과정에서 로봇장치에 의하여 처리 되었다. 각 샘플은 15,156개의 특성으로 구성되어 있다. 전처리단계에서 선택된 특성은 659개 이며 Ovarian 4-3-02 데이터 세트에 비하여 특성들 간의 상관관계가 크다는 결론을 얻을 수 있다. 다음 단계에서는 [5]과 동일하게 학습, 시험, 검증용 데이터

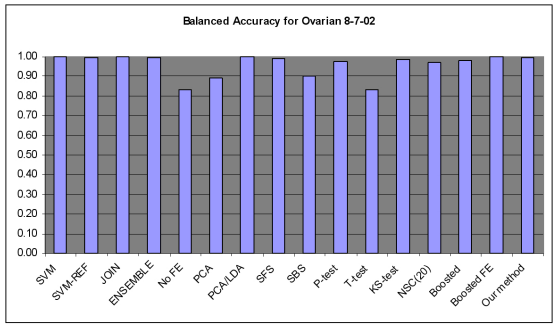

Fig. 3. Balanced accuracy for data set Ovarian 8-7-02

Table 2. Performance on data set Ovarian 8-7-02

	Accuracy (Standard deviation)	Balaced accuracy (Standard deviation)	Specificity (Standard deviation)	Sensitivity (Standard deviation)	Positive predictive value (Standard deviation)
SVM[5]	NA	0.9978	0.9955 (0.0002)	1.0000 (0.0000)	NA
SVM-REF[5]	NA	0.9932	0.9864 (0.0005)	1.0000 (0.0000)	NA
JOIN[5]	NA	1.0000	1.0000 (0.0000)	1.0000 (0.0000)	NA
ENSEMBLE[5]	NA	0.9955	0.9909 (0.0004)	1.0000 (0.0000)	NA
No FE[6]	0.837 (0.14)	0.834 (0.12)	0.822 (0.07)	0.846 (0.20)	0.891 (0.05)
PCA[6]	0.901 (0.05)	0.893 (0.03)	0.867 (0.03)	0.920 (0.07)	0.926 (0.02)
PCA/LDA[6]	1.000 (0.00)	1.000 (0.00)	1.000 (0.00)	1.000 (0.00)	1.000 (0.00)
SFS[6]	0.992 (0.01)	0.991 (0.01)	0.989 (0.02)	0.994 (0.01)	0.994 (0.01)
SBS[6]	0.901 (0.14)	0.903 (0.13)	0.911 (0.10)	0.895 (0.17)	0.942 (0.07)
P-test[6]	0.980 (0.02)	0.975 (0.03)	0.956 (0.05)	0.994 (0.01)	0.976 (0.03)
T-test[6]	0.837 (0.07)	0.834 (0.04)	0.822 (0.05)	0.846 (0.13)	0.897 (0.01)
KS-test[6]	0.984 (0.02)	0.983 (0.02)	0.978 (0.04)	0.988 (0.01)	0.988 (0.02)
NSC (20)[6]	0.972 (0.02)	0.973 (0.03)	0.978 (0.04)	0.969 (0.03)	0.988 (0.02)
Boosted[6]	0.980 (0.01)	0.982 (0.00)	0.989 (0.02)	0.975 (0.02)	0.994 (0.01)
Boosted FE[6]	1.000 (0.00)	1.000 (0.00)	1.000 (0.00)	1.000 (0.00)	1.000 (0.00)
Suggested method	0.994 (0.01)	0.995 (0.01)	0.992 (0.01)	0.999 (0.00)	0.982 (0.04)

세트를 각각 127개, 54개, 62개의 샘플로 구성하였다. Table 2에 표시된 바와 같이 이 샘플에 대해서 제안되는 알고리즘은 기존 방법에 비하여 개선된 성능을 나타내었다. 민감도와 특이도가 모두 99%에 도달하였다. 100%의 성능을 나타내는 JOINT, PCA/LDA, Boosted FE 등을 제외하고는 다른 방법에 비하여 개선된 결과를 얻었다.

4.3 Prostate

이 데이터 세트는 H4 단백질 칩을 이용하여 생성되었으며 전체 322개 샘플은 69명의 암환자, 253명의 정상 또

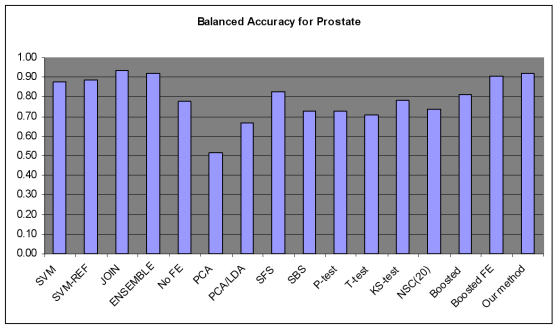

Fig. 4. Balanced accuracy for data set

Table 3. Performance on data set Prostate

	Accuracy (Standard deviation)	Balaced accuracy (Standard deviation)	Specificity (Standard deviation)	Sensitivity (Standard deviation)	Positive predictive value (Standard deviation)
SVM[5]	NA	0.8761	0.9698 (0.0011)	0.7824 (0.0062)	NA
SVM-REF[5]	NA	0.8871	0.9270 (0.0028)	0.8471 (0.0094)	NA
JOIN[5]	NA	0.9332	0.9016 (0.0022)	0.9647 (0.0009)	NA
ENSEMBLE[5]	NA	0.9217	0.8905 (0.0055)	0.9529 (0.0029)	NA
No FE[6]	0.732 (0.05)	0.777 (0.06)	0.698 (0.05)	0.855 (0.09)	0.439 (0.06)
PCA[6]	0.530 (0.20)	0.516 (0.18)	0.540 (0.24)	0.493 (0.21)	0.248 (0.11)
PCA/LDA[6]	0.692 (0.15)	0.667 (0.14)	0.710 (0.22)	0.623 (0.33)	0.431 (0.17)
SFS[6]	0.885 (0.05)	0.827 (0.17)	0.929 (0.03)	0.725 (0.36)	0.728 (0.03)
SBS[6]	0.773 (0.03)	0.729 (0.09)	0.806 (0.11)	0.652 (0.27)	0.498 (0.07)
P-test[6]	0.813 (0.02)	0.728 (0.11)	0.877 (0.08)	0.580 (0.28)	0.572 (0.07)
T-test[6]	0.816 (0.04)	0.709 (0.14)	0.897 (0.05)	0.522 (0.31)	0.575 (0.07)
KS-test[6]	0.826 (0.04)	0.784 (0.14)	0.857 (0.08)	0.710 (0.35)	0.579 (0.05)
NSC (20)[6]	0.791 (0.04)	0.736 (0.10)	0.833 (0.12)	0.638 (0.31)	0.529 (0.07)
Boosted [6]	0.850 (0.06)	0.810 (0.11)	0.881 (0.04)	0.739 (0.22)	0.627 (0.10)
Boosted FE[6]	0.960 (0.01)	0.906 (0.03)	1.000 (0.00)	0.812 (0.07)	1.000 (0.00)
Suggested method	0.918 (0.02)	0.921 (0.04)	0.924 (0.08)	0.918 (0.03)	0.984 (0.02)

는 양성 샘플로 구성되어 있다. 전처리 단계에서 6,784 개의 특성이 선택되었다. 다음 단계에서 학습, 시험, 검증용 데이터세트는 각각 162개, 80개, 80개의 샘플로 구성되었다. 이 데이터 세트에서 제안된 방법이 Table 3에서와 같이 균형정확도 91.2%, 민감도 91.8%, 특이도 92.4%의 가장 좋은 결과를 보여주었다. JOINT, ENSEMBLE과 제안된 방법이 최상의 성능을 보여 주었으며 민감도와 특이도의 차이가 줄어드는 효과를 나타내었다.

5. 결 론

본 논문에서는 피어슨 상관계수와 랜덤 포리스트를 기반으로 하는 SELDI-TOF 데이터에 적용되는 새로운 특성 선택 방법이 제안되었다. 전처리 단계에서는 피어슨 상관 계수를 이용하여 중복된 특성을 제거하여 탐색공간을 효율적으로 대폭 감소시키며 두 번째 단계에서는 랜덤 포리스트에서 산출된 중요도를 기반으로 토너먼트 방식의 전략으로 최적 특성조합을 탐색하는 방법이 적용되었다. 공개된 단백체 스텍트럼 데이터 세트에 제안된 알고리즘을 적용하여 기존의 방법들과 비교하였을 때 본 방법은 널리 사용되는 다른 방법과 비교할 만한 성능을 보여주었으며 일부 데이터에 대해서는 최상의 성능을 보여 주었다.

References

1. S. Das, Filters, wrappers and a boosting-based hybrid for feature selection, Proceedings of the 18th ICML, pp. 74-81, 2001.
2. A.Y. Ng, “On feature selection: learning with exponentially many irrelevant features as training examples”, Proceedings of the Fifteenth International Conference on Machine Learning, 1998.
3. E. Xing, M. Jordan and R. Carp, “Feature selection for highdimensional genomic microarray data”, Proc. of the 18th ICML, 2001.
4. E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn and L.A. Liotta, “Use of proteomic patterns in serum to identify ovarian cancer”, Lancet. Vol. 359, No. 9306, pp. 572-577, 2002.
5. K. Jong, E. Marchiori, M. Sebagy and A. Vaart, Feature Selection in Proteomic Pattern Data with Support Vector Machines, pp. 41-48, 2004.
6. I. Levner, Feature selection and nearest centroid classification for protein mass spectrometry, BMC Bioinformatics, 2005, available from http://www.biomedcentral.com/1471-2105/6/68.
7. R.H. Lilien, H. Farid and B.R. Donald, Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. Computational Biology, Vol. 10, No. 6, pp. 925-946, 2003.
8. R. Tibshirani, T. Hastiey, B. Narasimhanz, S. Soltys, G. Shi, A. Koong and Q. Le, Sample classifcation from protein mass spectrometry by ‘peak probability contrasts’. BioInformatics, Vol. 7, No. 17, pp. 3034-3044, 2004.
9. W. Michael, D.N. Naik, S. Kasukurti, A. Pothen, R.R. Devineni, B.L. Adam, O.J. Semmes and G.L. Wright, Computational protein biomarker prediction: a case study for prostate cancer. BMC Bioinformatics, 2004, available from http://www.biomedcentral.com/1471-2105/5/26.
10. B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams and H. Zhao, Comparison of statistical methods for classifcation of ovarian cancer using mass spectrometry data. BioInformatics, Vol. 19, No. 13, pp. 1636-1643, 2003.
11. L. Breiman, Random forest, Machine Learning, Vol. 45, pp. 5-32, 2001.
12. R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification; 2nd Edition, John Wiley & Sons Inc, 2001.
13. P.N. Tan, M. Steinbach and V.S Kumar, Introduction to Data mining, Addison-Wesley, 2006.
14. I, Guyon and A. Elisseeff, An introduction to variable and feature selection, Machine learning, Vol. 3, Special Issue on variable and feature selection, pp. 1157-1182, 2003.
15. http://clinicalproteomics.steem.com/

온 승 엽 (syohn@kau.ac.kr)

1984 서울대학교 전기 공학과 학사
1988 미국 뉴욕 폴리테크닉 대학교 컴퓨터 공학과 석사
1995 미국 뉴욕 폴리테크닉 대학교 컴퓨터 공학과 박사
1996～1997 한국 통신 멀티미디어 연구소 선임 연구원
～현재 한국항공대학교 한국항공대학교 항공전자 및 정보통신공학부 교수

관심분야 : 데이터 마이닝, 멀티미디어, 패턴인식, 바이오인포머틱스, 컴퓨터비젼

지 승 도 (sdchi@kau.ac.kr)

1982 연세대학교 전기공학과 학사
1984 연세대학교 전기공학과 석사
1985～1986 두산 컴퓨터 (현 한국 디지털) 근무
1991 미국 아리조나대학교 전기전산공학과 박사
1991～1992 미국 SIMEX Systems and S/W 회사 S/W 담당자로 근무
1992～현재 한국항공대학교 항공전자 및 정보통신공학부 교수

관심분야 : 이산사건 시스템 모델링 및 시뮬레이션, 컴퓨터 보안, 지능시스템 디자인 방법론, 시뮬레이션 기반 인공생명, 교통 모델링

한 미 영 (myhan@kofac.re.kr)

1974～1978 연세대학교 이과대학 생화학과 학사
1978～1980 연세대학교 이과대학 생화학과 석사
1980～1985 연세대학교 이과대학 생화학과 박사
1991～1992 UCSF Dep.of Medicine Fellow
1987～2001 한국생명공학연구원 선임/책임연구원
1995～2004 인천대학교 자연대학 생물학과 객원교수
2001～2006 녹십자의료재단 분자의학연구소 연구소장
2005～2006 ㈜바이오인프라 연구소장
2007～현재 한국과학창의재단 위원

관심분야 : 단백질생화학, 프로테오믹스, 바이오인포마틱스, 진단법 개발