Journal of Korean Institute of Intelligent Systems, Vol. 22, No. 2, April 2012, pp. 192-197
http://dx.doi.org/10.5391/JKIIS.2012.22.2.192

유전자 상호작용 정보와 mRMR 필터 기반의 Random Subspace Method를 이용한 질병 진단

Disease Classification using Random Subspace Method based on Gene Interaction Infomation and mRMR Filter

최선욱.이종호†

Sun-Wook Choi and Chong Ho Lee†

인하대학교 정보통신공학과

요 약

DNA 마이크로어레이 기술의 발달과 함께 이를 활용한 질병 진단 및 치료 예후 확인을 목적으로 하는 연구가 활발히 진행 되고 있다. 일반적으로 마이크로어레이 데이터를 이용한 실험에서는 특징들의 수에 비해 적은 샘플의 수, 내재적 측정 노이즈, 서로 다른 샘플들 간의 이질성 등이 분류 성능을 떨어트리는 원인이 된다. 이러한 문제를 극복하기 위해 패스웨이 기반의 기능적 모듈 단위의 마커를 사용하는 방법들이 새롭게 제안 되었다. 이들은 패스웨이의 멤버 유전자들의 발현 값을 요약하여 해당 패스웨이의 활성도로 사용하는데, 기존의 기법들과 비교하여 뛰어난 분류 성능과 재현성을 보여주었다. 그러나 이러한 활성도 계산 방법은 개별 유전자들과 표현형 사이의 상관관계를 무시하거나, 개별 유전자들이 갖는 발현 특성이 제거 되는 단점들이 있다. 본 논문에서는 선택된 기능적 모듈 단위의 유전자들의 부분집합들을 기반으로 약 분류기를 구성하고, 이들의 분류 결과를 결합하여 최종 결과를 추론하는 앙상블 분류 기법을 제안한다. 이 과정에서 유전자 상호작용 정보와 mRMR 필터를 사용하는 필터링과정을 통해 탐색 공간을 최소화하여 분류 성능을 높일 수 있도록 하였다. 제안 된 방법의 성능을 테스트하기 위해 폐암 데이터에 적용한 결과, 기존의 기법들에 비해 신뢰성이 있고 우수한 분류 성능을 보여주었다.

키워드 : 유전자 상호작용, 신호 전달 경로, 앙상블 분류기, mRMR, Random Subspace Method

Abstract

With the advent of DNA microarray technologies, researches for disease diagnosis has been actively in progress. In typical experiments using microarray data, problems such as the large number of genes and the relatively small number of samples, the inherent measurement noise and the heterogeneity across different samples are the cause of the performance decrease. To overcome these problems, a new method using functional modules (e.g. signaling pathways) used as markers was proposed. They use the method using an activity of pathway summarizing values of a member gene's expression values. It showed better classification performance than the existing methods based on individual genes. The activity calculation, however, used in the method has some drawbacks such as a correlation between individual genes and each phenotype is ignored and characteristics of individual genes are removed. In this paper, we propose a method based on the ensemble classifier. It makes weak classifiers based on feature vectors using subsets of genes in selected pathways, and then infers the final classification result by combining the results of each weak classifier. In this process, we improved the performance by minimize the search space through a filtering process using gene-gene interaction information and the mRMR filter. We applied the proposed method to a classifying the lung cancer, it showed competitive classification performance compared to existing methods.

Key Words : Gene Interaction, Signaling Pathway, Ensemble Classifier, mRMR, Random Subspace Method

1. 서 론

DNA 마이크로어레이 기술의 발달로 인해 한 번의 실험으로 대량의 유전자의 발현 양상을 동시에 확인하는 것이 가능하게 되었다. 이를 활용하여 질병 진단, 치료 예후 확인 등을 목적으로 하는 연구가 최근까지 활발하게 진행되어오고 있다.[1] 특히 서로 다른 표현형(예 : 암의 양성, 악성)에 대해 특이하게 발현 되는 유의한 유전자를 찾기 위한 노력들이 주를 이루었다.[2] 유의하게 발현의 차이가 나타나는 유전자들은 진단을 위한 바이오 마커로 사용 될 수 있기 때문에, 신뢰성 있는 유전자 마커를 찾아내는 것이 매우 중요한 일이라 할 수 있다. 그러나 신뢰성 있는 유전자 마커를

접수일자: 2011년 12월 16일
심사(수정)일자: 2012년 3월 23일
게재확정일자 : 2012년 3월 26일
†교신저자
이 논문은 인하대학교의 지원에 의하여 연구되었음.

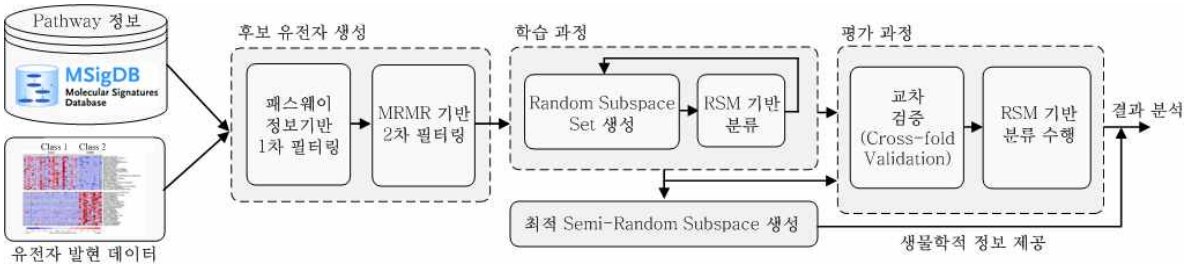

그림 1. 제안 된 기법의 전체 과정

Fig. 1. Overall process of the proposed method

찾는 일은 매우 어려운 문제이다. 최근까지 연구 결과들에 의하면 개별적으로 선별 된 유전자 마커들을 기반으로 한 분류 기법들의 신뢰성에 대한 문제들이 제기 되어오고 있다.[3,4] 의학용 연구 데이터들에서 일반적으로 나타나는 특성과 같이 마이크로어레이 실험과 관련 된 연구 데이터들 또한 많은 특징(유전자)들의 수에 비해 적은 수의 샘플을 가지고 분석을 수행해야 하는 문제가 있다. 이는 분석을 더욱 어렵게 만드는 주요한 원인이라 할 수 있다.

마이크로어레이 데이터의 분류 성능을 높이기 위해 일반적으로 사용되는 기법은 다수의 유전자들로부터 적은 수의 유전자 마커들을 탐색하여 데이터의 차원을 감소시키는 것이다. 그러나 여전히 마이크로어레이를 이용하는 실험 중에 발생하는 내재적인 측정 노이즈, 각 샘플(환자)들 간의 이질성 등이 분류 성능을 감소시키는 원인이 된다. 이러한 문제를 극복하기 위해 제시 된 방법 중 하나는 유전자 발현 데이터를 신호전달 경로 (signaling pathway, 이하 패스웨이)와 같은 기능적 모듈 레벨에서 해석하는 것이다.

일반적인 유전자 기반의 분류 기법이 갖고 있는 한계들 중 하나는 유전자들이 단백질과 같은 그 기능적 산물들이 서로 상호작용하고 있음에도 불구하고 독립적으로 계산되어 선택이 이루어진다는 점이다. 때문에 독립적으로 선택 되어지는 유전자들은 불필요하게 중복되는 정보를 포함하게 되고, 이는 분류 성능 감소의 원인이 될 수 있다. 이러한 한계를 개선하기 위한 방법으로 Gene Ontology나 KEGG database 등으로 부터 얻을 수 있는 유전자들의 상호작용 정보와 유전자들의 발현 정보를 함께 고려하여 분석하는 방법들이 제안 되었다.[5] 이와 관련 된 최근의 연구들에서는 패스웨이 기반의 마커들이 단일 유전자 마커들과 비교해 재현성이 높음을 보였고, 질병과 관련 된 내부 메커니즘의 이해를 도울 수 있는 생물학적 정보 또한 부가적으로 제공해줄 수 있다는 장점이 있다. 이들 연구에서 패스웨이 기반의 분류기들은 기존의 유전자 기반의 분류기들과 비교하여 일반적으로 나은 분류 성능을 보였다.

패스웨이 기반의 마커들을 분류에 적용하기 위해서는 주어진 패스웨이를 구성하는 유전자들의 발현 값들을 이용하여 해당 패스웨이의 활성도(activity)를 추론 할 필요가 있다. 패스웨이 별로 멤버 유전자들의 발현 값이 통합 되면 개별 유전자들의 발현 값의 편차가 낮아지기 때문에 재현성이 높고, 안정적인 분류 성능을 얻을 수 있게 된다. 이를 위하여 패스웨이의 활성도를 추론하기 위한 다양한 기법들이 제시되었다. Guo 등은 패스웨이의 활성도를 추론하기 위해 해당 멤버 유전자들의 발현 값들의 중앙값(mean)과 중간값(median)을 사용하는 것을 제안하였다.[6] Tomfohr 등과 Bild 등은 멤버 유전자들의 발현 값들의 첫 번째 주성분을 해당 패스웨이의 활성도 값으로 사용하였다. [7,8] Lee 등은 패스웨이의 활성도 예측을 CORGs 라고 부르는 멤버 유전자들의 부분 집합을 추출하여, 해당 CORG들의 발현 값을 결합하는 기법을 제안하였다.[9] 그러나 이들 패스웨이의 활성도에 기반 한 분류 기법들은 여전히 몇 가지 문제점을 가지고 있다. 패스웨이에 속하는 멤버 유전자들은 복합적으로 상호연관 되어있기 때문에 특정 표현형에 대해 어떤 유전자들은 강하게 발현하고, 어떤 유전자들은 발현 하지 않을 수도 있다, 즉 멤버 유전자들은 특정 표현형에 대해 양의 상관관계일 수도 있고, 음의 상관관계를 가질 수도 있다. 그러나 기존의 기법들은 이러한 특성들을 무시한다는 단점들이 있다. 이에 Su 등은 멤버 유전자들이 서로 다른 표현형에서 발현 할 확률에 대한 확률 밀도 함수를 계산 하여 이용하는 방법을 제안하였다.[10] 확률 밀도의 차이를 로그 비(log ratio)로 계산하여 해당 패스웨이에 속한 멤버 유전자의 점수를 구하고, 이들을 합산한 결과를 최종적인 활성도로 사용한다. 그러나 이들 활성도를 이용한 방법의 단점은 최적의 활성도 계산 방법이 알려져 있지 않은 상태이고, 멤버 유전자들 각각의 유전자 발현 상태가 하나의 활성도로 축약되면서, 멤버 유전자들이 각각 가지고 있던 발현 특징들이 사라진다는 점이다.

본 논문에서는 마이크로어레이로부터 얻어지는 유전자 발현 데이터와 유전자 상호작용 정보를 결합하여 분석 할 수 있는 새로운 방식의 분류 기법을 제안하고자 한다. 제안하는 방법은 먼저 유전자 상호작용 정보를 탐색 범위를 줄이기 위한 1차적인 정보로써 활용하여 전체 유전자로부터 상호작용할 가능성이 높은 후보 유전자 집합을 추출해 낸다. 이들 1차 필터링 된 후보 유전자들을 대상으로 상관성(relevance)이 낮거나 중복성(redundancy)이 높은 유전자들을 제거하는 과정을 2차적으로 수행한다. 이렇게 2단계에 걸친 후보 유전자 생성 과정을 거친 후, Random Subspace Method(RSM)를 사용하여 각 표현형에 대해 분류 성능이 높은 유전자 부분 집합들을 탐색해 내고, 이들 유전자 부분 집합들 각각에 대한 분류 결과를 앙상블 기법을 통하여 결합함으로써 분류 성능을 높일 수 있도록 하였다. RSM을 이용하는 과정에서 발생하는 각 부분 집합들에 기반 한 약 분류기(weak classifier)들은 앞서 활성도 계산을 이용한 기법들에서 발생했던 문제점을 개선하기 위한 대안이 될 수 있다.

2. 제안하는 기법

2.1 패스웨이 정보 및 mRMR 기반의 유전자 필터링

기존의 RSM과 달리 본 논문에서 제안하는 기법은 학습을 통하여 최적의 부분 공간(subspace)을 탐색하는

과정을 거치는 Semi-Random Subspace 기법이기 때문에 탐색 공간을 줄이는 것이 학습에 유리하다. 이를 위해 특징 선택(feature selection) 과정이 필요한데, 기존의 일반적인 기법들은 앞서 언급한 것과 같이 각 특징들의 상호작용을 고려하지 않는 단점이 있다. 따라서 본 논문에서는 넓은 탐색 범위를 줄이면서, 상호작용하고 있을 가능성이 높은 후보 유전자들을 추출하기 위해 신호 전달 경로 데이터베이스 정보를 이용한다. 신호 전달 경로란 세포 내 생화학적 반응이 일어나는 일련의 신호 전달 과정을 의미하는 것으로 이를 통해 유전자의 발현을 유도하거나 억제하여 생체 내의 반응을 조절하는 역할을 한다. 따라서 동일한 패스웨이에 속한 유전자들은 서로 상호작용 할 가능성이 매우 높다고 할 수 있다. 이러한 가정을 바탕으로 1차적으로 후보 유전자들을 필터링 해내기 위해, 접근 가능한 패스웨이 정보로부터 소속 유전자들의 유전자 심볼(gene symbol)을 추출하여 나열하고, 실험에 사용 된 마이크로어레이를 구성하는 유전자들의 Probset ID들과 매칭 되는 유전자 정보를 추출 한다. 이렇게 추출 된 유전자들은 서로 상호작용 할 가능성이 있는 후보 유전자들이라고 할 수 있다.

패스웨이 정보를 이용하여 필터링 된 후보 유전자들에 대해 2차적으로 mRMR(minimum Redundancy Maximum Relevance) 특징 선택 기법에 기반을 두어 상관성이 낮으면서 중복성이 높은 유전자들을 제거하여 보다 효율적으로 유전자 집합을 탐색 할 수 있도록 한다. 일반적으로 고차원 데이터에서는 연관성이 없거나, 중복 된 변수가 다수 존재한다. 이들은 불필요하게 복잡도(complexity)를 높이고, 모델의 일반화 성능을 저하시키며, 과적합(over-fitting)의 가능성을 높일 뿐만 아니라 해석력과 설명력을 떨어트린다. 따라서 이들 문제를 최소화 할 수 있는 최적의 집합을 효율적으로 찾아내는 것은 매우 중요한 일이다.

mRMR 기법은 변수 간 상호정보량에 기반 하여 높은 상관성(maximum relevancy)을 가지는 동시에 낮은 중복성(minimum redundancy)을 가지는 특징들을 추출할 수 있도록 해준다.[11]

본 논문에서 우리는 이진 분류 문제를 가정하고 있으므로, 샘플 k의 클래스 라벨(label)이 $y_k = l \in \{+1, -1\}$와 같다고 한다면 클래스 라벨 l과 유전자 i사이의 상호정보량을 이용하여 분류를 위한 유전자 i의 중복성을 수치화 할 수 있다. 이를 이용하여 유전자 전체(G)에 대한 유전자 부분 집합(S)들의 상관성 R_S는 다음과 같이 구할 수 있다.

$$I(l,i) = \sum_{x_i} p(l,x_i) \log \frac{p(l,x_i)}{p(l)p(x_i)} \quad (1)$$

$$R_S = \frac{1}{|S|} \sum_{l} \sum_{i \in S} I(l,i) \quad (2)$$

여기서 $I(l,i)$는 클래스 라벨 l과 유전자 i 사이의 상호정보량을 의미하고, $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,k})$는 k개의 샘플들 간의 i번째 유전자들을 의미한다. 연속형 변수의 경우 상호정보량의 계산하는데 어려움이 있을 수 있는데, 이 경우 이산화를 수행하거나 비모수적 기법을 이용하여 확률 밀도를 계산하는 방법이 있다.

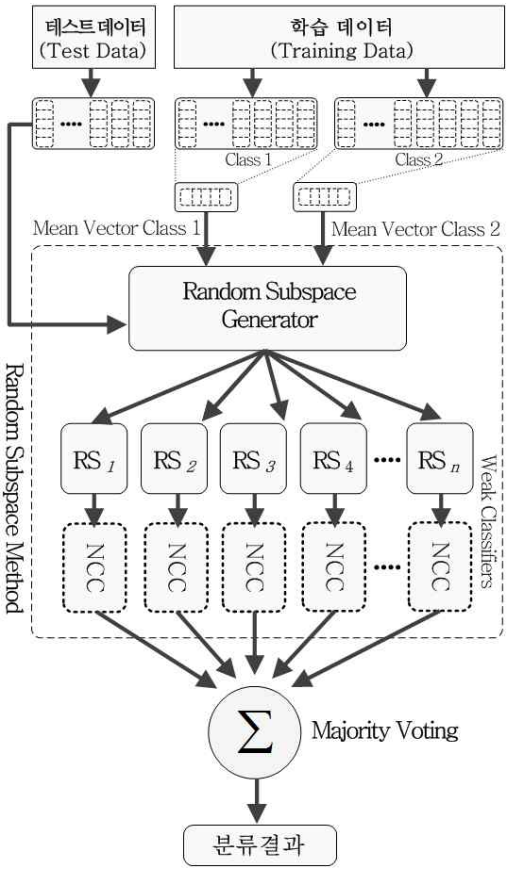

그림 2. 랜덤 부분 공간 기법
Fig. 2. Random Subspace Method

유전자 부분 집합의 중복성은 유전자들 간의 상호정보량을 이용하여 구할 수 있다. 유전자 i와 유전자 부분 집합 S 내의 다른 유전자들과의 중복성 Q_{Si}는 다음과 같이 구할 수 있다.

$$Q_{Si} = \frac{1}{|S|^2} \sum_{i' \in Si' \neq i} I(i,i') \quad (3)$$

mRMR 기법에서, 유전자들의 순위는 유전자 부분 집합 S 에 속한 유전자들의 상관성과 중복성의 비(ratio)를 이용하여 구할 수 있다. 따라서 높은 상관성을 갖으면서 낮은 중복성을 갖는 유전자들의 순위는 다음과 같이 구할 수 있다.

$$i^* = \arg max_{i \in S} \frac{R_S}{Q_{Si}} \quad (4)$$

상관성과 중복성에 기반 한 유전자들의 순위 측정은 다양한 방식으로 이루어 질 수 있는데, 식 (4)와 같은 방식이 일반적으로 우수한 성능을 보이는 것으로 알려져 있다. 식(4)에 의하여 가장 높은 순위의 유전자를 선택한 후, 다음 순위에 해당하는 유전자들도 전진 선택(forward selection) 기법에 의하여 식(4)를 최대화 하는 유전자들의 순서로 구할 수 있다. (이에 대한 보다 상세한 정보는 [11]에서 확인 할 수 있다.)

2.2 Random Suspace Method 및 학습 과정

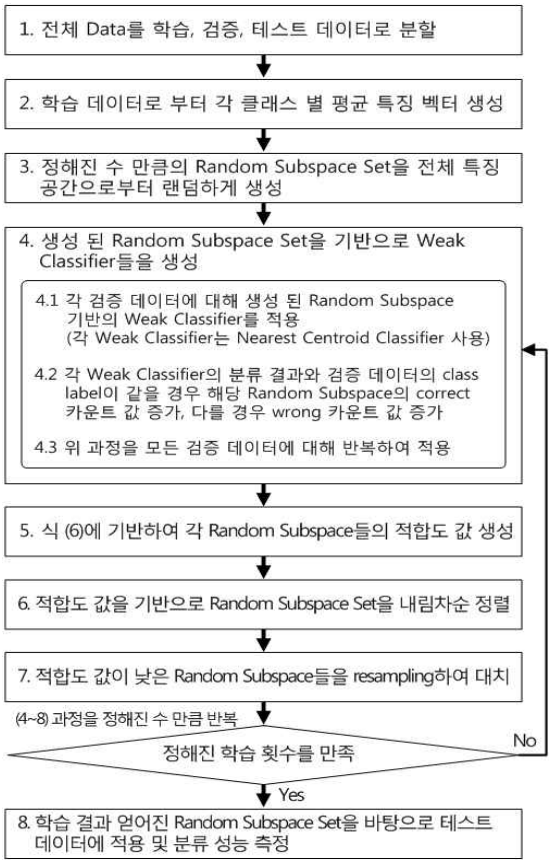

그림 3. RSM의 학습 과정

Fig. 3. Learning Process of the RSM

RSM은 Ho에 의해 제안 된 앙상블 기법 중에 하나이다.[12] 다른 앙상블 기법과 차별되는 RSM의 특징은 원래의 특징 벡터로부터 랜덤하게 생성 된 부분 공간(random subspace)을 사용하여 앙상블을 구성한다는 점이다. 이러한 특성을 기반으로 RSM은 충분한 양의 데이터를 수집하기 어려운 상황에서도 만족할 만한 성능을 얻을 수 있도록 한다. RSM에서 각각의 약 분류기들은 전체 특징 벡터의 random subspace를 사용하여 구성 된다. 본 논문에서는 약 분류기로써 최근접 중심 분류기(nearest centroid classifier, NCC)를 사용하였다. 최근접 중심 분류기는 그 사용이 단순하여 빠른 처리에 적합하기 때문에 많은 수의 분류기를 생성하여 분석해야 하는 앙상블 기법에 적합하다. 또한 패스웨이의 각 멤버 유전자들의 발현 양상을 하나로 통합하여 분석하기에 적합한 분류기라고 할 수 있다. 학습 데이터 $T=(T_1, T_2,..., T_n)$가 주어졌을 때, 각각의 샘플 벡터 T_i는 d차원의 특징벡터 $g_i=(g_{i1},g_{i2},...,g_{id})$로 구성 된다. RSM은 m개의 새로운 무작위로 선택 된 부분 공간인 random subspace $RS=(RS_1,RS_2,...,RS_m)$를 각각의 학습 데이터로부터 생성 할 수 있다. 생성 된 각 random subspace RS_m은 랜덤하게 생성 된 p개의 특징들 $r_m=(r_{m1},r_{m2},...,r_{mp})$로 구성 된다. 이때 $p<d$이며, 경우에 따라 고정 된 크기 혹은 가변적인 크기의 p값을 사용 할 수 있다. 이들 random subspace들을 사용하여 구성 된 약한 분류기들 $wc(r_i)$은 일반적으로 식(5)와 같이 다수 투표(majority voting) 방식을 통하여 결합 된다. 이렇게 결합 된 결과들이 강 분류기(strong classifier)를 구성하는데 사용 된다.

$$class\ label\ y^*=\underset{y\in\{+1,-1\}}{\arg max}\sum_{i=1}^{m} wc(r(x)_i)_y \quad (5)$$

본 논문에서는 1,2차 필터링을 통해 얻어진 후보 유전자들을 대상으로, 상호작용 할 가능성이 높은 유전자 집합을 찾아내기 위해 학습 알고리즘 기반의 RSM 기법을 제안한다. 일반적인 RSM 에서는 분류 시마다 새롭게 부 공간을 생성하여 분류에 적용한다. 그러나 학습 알고리즘 기반의 RSM에서는 분류 성능을 최대로 하는 부분 공간 집합(semi-random subspace)을 찾아내는 것을 목표로 한다. 이를 위해 비교적 단순한 진화 연산 기반의 학습 알고리즘을 사용한다. 먼저 학습 데이터를 내부적으로 다시 분할하여 검증 데이터를 얻고, 이를 학습에 사용한다. 일반적인 RSM과 같은 방식으로 random subspace들을 생성하여 이를 초기 해로 사용하고, 매 회 학습 과정에서 각 random subspace들의 적합도(fitness)값을 아래와 같이 계산한다. 여기서 α와 β는 상수로써, 옳은 매칭(correct matching)과 잘못된 매칭(wrong matching)의 중요도를 결정 할 수 있으며, 실험적으로 정하여 사용한다. 본 실험에서 기본적으로 사용한 값은 $\alpha=1$, $\beta=0.001$이다.

$$fitness\ of\ RS_i=\frac{\alpha}{wrong\_cnt_i}+\beta\times correct\_cnt_i \quad (6)$$

각 random subspace의 적합도 값을 바탕으로 내림차순 정렬 한 후, 하위 순위에 있는 random subspace들을 일정한 양 만큼 새로운 random subspace들로 교체한다. 이러한 학습 과정을 정해 놓은 회수만큼 반복한 뒤 얻어지는 상위 순위에 있는 random subspace들을 구성하는 유전자들은 서로 상호작용 할 가능성이 높은 집합으로 생각 할 수 있다. 또한 이렇게 학습 된 random subspace들을 기반으로 최고의 분류 성능을 보이는 분류기를 구할 수 있다. RSM 분류기를 위한 진화 연산 기반의 학습과 분류에 대한 진행 과정은 그림 3과 같다.

3. 실험 및 결과

3.1 데이터 및 전 처리

제안한 방법의 성능을 평가하기 위해 본 연구에서는 Beer DG 등(이하 Michigan 데이터)과 Bhattachargee 등(이하 Boston 데이터)이 이전의 연구 결과 발표에서 사용한 두 개의 독립적인 폐암(lung cancer) 데이터를 사용하였다.[13,14] 두 데이터 각각 Affmetrix사의 HU6800, HGU95Av2 마이크로어레이를 사용한 실험으로부터 획득 되었다. Michigan 데이터는 환자 86명의 유전자 발현 데이터(폐암:24명, 정상:62명, 유전자:7,219개)로 구성 되어 있다. Boston 데이터는 환자 62명의 유전자 발현 데이터 (폐암:31명, 정상:31명, 유전

자:12,600)로 구성 되어 있다. 서로 다른 환경에서 얻어진 두 가지 데이터를 서로 교차하여 실험함으로써 제시한 방법의 강건함과 재현성을 확인 할 수 있다.

또한 생물학적 신호 전달 경로 데이터를 얻기 위해, MSigDB v2.5의 C2 curated geneset의 canonical pathway 데이터를 다운로드하여 사용하였다. 해당 데이터는 639개의 패스웨이 정보로 구성 되어 있는데, 이들은 KEGG 데이터베이스와 GenMAPP 등과 같은 다양한 패스웨이 데이터베이스로부터 얻어진 데이터들을 취합하여 만들어졌다.

Michigan 데이터와 Boston 데이터의 취득에 사용 된 마이크로어레이의 유전자 리스트에 공통적으로 포함 되어 있는 유전자는 5,217개 이었고, 이들 중 전체 패스웨이 내에 1개 이상 포함 되어 있는 유전자를 필터링한 결과 2,718개의 유전자를 얻을 수 있었다. 따라서 이들 2,718개의 유전자가 1차 필터링 된 후보 유전자라고 할 수 있다.

3.2 실험 방법

제안 된 방법의 성능을 평가하기 위해 2가지 방법으로 실험을 수행 하였으며, 공통적으로 교차 검증 방식(cross-fold validation)을 사용하였다. 먼저 단일 데이터 실험에서는 Michigan 데이터와 Boston 데이터 각각에 대한 분류 성능을 평가하기 위해 5-fold 교차 검증 실험을 수행하였고, 이를 각각 100회 반복 실험하였기 때문에, 총 500회의 실험을 수행하였다.

제안 된 기법의 분류 성능을 측정 하기위한 성능지표로써 AUC(Area Under ROC Curve)값과 분류 정확도 값을 사용한다.[15] AUC 값은 민감도와 특이도를 모두 고려 할 수 있는 평가 수치로써 ROC curve 아래의 면적(본 논문에서는 0~100 사이의 값으로 환산)이 넓은 수록 좋은 성능을 보인다고 할 수 있다. AUC와 정확도 각각의 값은 전체 500회의 실험에서 얻어진 값들의 평균값을 구하여 최종 결과 값으로 사용하였다.

또한 제안 한 방법의 재현성(reproducibility)을 평가하기 위해 서로 다른 환경으로부터 얻어진 데이터를 교차 하여 실험을 수행하였다. 교차 데이터 실험에서는 한 데이터를 테스트 데이터로, 나머지 한 데이터를 학습 데이터로 사용한다. 단일 데이터 실험에서와 마찬가지로 테스트 데이터에 대해 5-fold 교차 검증 실험을 수행하였고, 이를 각각 100회 반복하여 실험하여, 역시 전체 500회의 실험을 수행하였다.

3.3 결과 분석

제안 된 방법의 성능을 기존의 기법들과 공평하고 효과적으로 비교하기 위해, Lee 등과 Su 등이 사용하였던 실험 방법과 동일하게 수행하였고, 이들이 LDA(Linear Discriminant Analysis)와 LR(Logistic Regression)을 기반으로 제안 하는 방법을 검증하는 실험을 수행하였기 때문에 본 논문에서도 이들 기법의 성능과 제안 하는 방법의 성능을 비교 실험 하였다. 실험 결과 기존의 다른 기법들에 비해 제안 하는 방법이 AUC와 분류 정확도 모두에서 전반적으로 우수한 성능을 보임을 확인 할 수 있었다. (표 1, 2 참조) 그러나 Boston 데이터를 이용한 실험에서는 제안하는 기법이 기존의 기법들과 유사한 성능을 보이고 있는데, 이는

표 1. 단일 데이터 실험 결과
Table 1. Experiment results of the single data

	Michigan Data		Boston Data	
	정확도	AUC	정확도	AUC
LDA-CORG	66.94%	67.63	54.79%	55.95
LDA-LLR	68.98%	63.04	56.48%	58.59
LR-CORG	70.36%	64.60	53.83%	55.48
LR-LLR	61.06%	68.05	54.72%	57.08
MRMR-RS	**73.00%**	**68.50**	**54.82%**	**58.61**

표 2. 교차 데이터 실험 결과
Table 2. Experiment results of the cross data

	Michigan to Boston		Boston to Michigan	
	정확도	AUC	정확도	AUC
LDA-CORG	53.52%	56.49	63.25%	67.70
LDA-LLR	56.63%	59.64	69.28%	64.65
LR-CORG	56.40%	58.61	68.76%	62.52
LR-LLR	56.13%	59.12	60.77%	66.59
MRMR-RS	**57.72%**	**61.51**	**73.72%**	**69.02**

표 3. 선택된 유전자 바이오 마커
Table 3. Selected gene bio-markers

순위	Gene	P-value	t-score	발생 빈도
1	**SPRR1B**	**5.299e-01**	**6.318e-01**	**17,027**
2	SKP1A	5.012e-01	6.766e-01	7,761
3	CAT	9.500e-03	2.679e-00	4,598
4	CCND1	2.401e-01	1.186e-00	1,805
5	CBX3	8.910e-02	1.728e-00	1,656

Michigan 데이터와 Boston 데이터 내의 양성과 음성 데이터의 클래스 비율이 서로 상이하기 때문인 것으로 생각 된다. 제안 된 방법의 신뢰성 있는 바이오마커 탐색 능력을 평가하기 위해, 교차 데이터(Boston to Michigan) 실험 중의 학습 과정에서 얻어진 semi-random subspace들 중 적합도를 기준으로, 상위 100개에 속한 유전자들의 발생 빈도를 전체 500번의 반복 실험을 통해 누적 계산하였다. 선택된 상위 5개의 유전자 마커들은 표3과 같다. 이들을 폐암과 관련 된 연구 문헌들을 통하여 검증한 결과 모두 폐암과 밀접하게 관련된 유전자들인 것으로 확인 되었다. 가장 빈번하게 발견 된 'SPRR1B' 유전자의 경우 [16]에서와 같이 전체 폐암 중에서 약 25~30%에 해당하는 형태인 편평상피세포 암종의 진단 시 민감도와 특이도가 모두 우수한 바이오 마커로 알려져 있다. 이는 제안 된 방법이 폐암을 진단하기 위한 바이오 마커를 탐색하기 위한 방법으로도 사용 될 수 있음을 의미한다.

4. 결론

본 논문에서는 유전자 발현 정보와 유전자 상호작용 정보를 결합하여 분석 할 수 있는 앙상블 분류기 기반의 방법을 제안하였다. 제안 한 방법은 기존의 패스웨이 기반의 방법들에서 사용하였던 활성도를 계산하며 발생하는 단점들을 개선하기위해 최근접 중심 분류기 기반의 약 분류기를 사용하였다. 분류 성능을 높이기 위해 유전자 상호작용 정보와 mRMR 필터를 사용하여 상호작용 할 가능성이 높은 후보 유전자들이 남도록 탐색 공간을 최소화 한 후, 학습 기반의 RSM 기법을 적용하여 최적의 분류 성능을 얻을 수 있도록 하였다. 학습 과정 중에 얻어지는 정보는 해당 질병의 생물학적 메커니즘을 이해하는데 도움을 줄 수 있고, 신뢰성 있는 바이오 마커를 탐색하는데도 적용할 수 있음을 알 수 있었다. 실험결과 기존의 기법들과 비교하여 경쟁력 있는 분류 성능을 보여주는 것을 확인 할 수 있었다. 차후 학습 중에 생성 되는 유전자 부분 집합들이 실제 생체 내에서 상호작용 하는지에 대한 검증 작업을 수행할 예정이다.

참 고 문 헌

[1] Golub, T., et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, 1999.

[2] 이선아, 이건명, 류근호, "퍼지 시그너쳐 집합을 이용한 마이크로어레이 데이터 검색," *한국지능시스템학회 논문지*, 제19권, 제4호, pp. 545-549, 2009.

[3] Braga-Neto UM, D.E., "Is cross-validation valid for smallsample microarray classification?," *Bioinformatics*, vol. 20, pp. 374-380, 2004.

[4] ER, D., "Small sample issues for microarray-based classification," *Comparative and functional genomics*, vol. 2, pp. 28-34, 2001.

[5] Subramanian A., et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl Acad Sci USA*, vol. 102, pp. 15545-15550, 2001.

[6] Guo Z., et al., "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinformatics,* vol. 6, pp. 58, 2005.

[7] Tomfohr J, Lu J, K.T., "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, vol. 6, pp. 225, 2005.

[8] Bild AH, et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, pp. 353-357, 2006.

[9] Lee E., et al., "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, e1000217, 2008.

[10] Su J., et al., "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," PLoS ONE, vol. 4, no. 12, e8161, 2009.

[11] Peng, H., et al., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[12] Ho, T.K., "The random subspace method for constructing decision forests," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.

[13] Beer DG., et al., "Gene expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, vol. 8, pp. 816-824, 2002

[14] Bhattacharjee A., et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Natl Acad Sci USA*, vol. 98, pp. 13790-13795, 2001.

[15] T, Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861-874, 2006.

[16] Ruide Hua, et al., "A small proline-rich protein, spr1: Specific marker for squamous lung carcinoma," *Lung Cancer*, vol. 20, Issue 1, pp. 25-30, 1998.

저 자 소 개

최선욱(Sun-Wook Choi)

2007년 : 인하대학교 전자공학 학사
2009년 : 인하대학교 정보통신공학 석사
2010년~현재 : 인하대 정보통신공학과 박사과정 재학중
관심분야 : 로봇지능, SLAM, 패턴인식

Phone : 032-860-7396
E-mail : swchoi@inhaian.net

이종호(Chong Ho Lee)

1976년 : 서울대학교 전기공학 학사
1978년 : 서울대학교 전기공학 석사
1986년 : 아이오와 주립대학교 전기 및 컴퓨터공학 박사
1986년~1989년 : 노틀담 대학교 조교수
1989년~현재 : 인하대학교 정보통신공학과 교수
1997년~1998년 : 인하대 직접회로설계센터 소장
2000년~2010년 : 인하대 슈퍼지능기술연구소 소장
2004년~2005년 : 브라운 대학교 두뇌 및 신경회로망 연구소 방문 교수

관심분야 : 지능형시스템, VLSI 설계
Phone : 032-860-7396
E-mail : chlee@inha.ac.kr