한국시뮬레이션학회 논문지
Vol. 20, No. 2, pp. 77-85 (2011. 6)

Genetic Algorithm과 다중부스팅 Classifier를 이용한 암진단 시스템

온승엽[1†] · 지승도[1]

Cancer Diagnosis System using Genetic Algorithm and Multi-boosting Classifier

Syng-Yup Ohn · Seung-Do Chi

ABSTRACT

It is believed that the anomalies or diseases of human organs are identified by the analysis of the patterns. This paper proposes a new classification technique for the identification of cancer disease using the proteome patterns obtained from two-dimensional polyacrylamide gel electrophoresis(2-D PAGE). In the new classification method, three different classification methods such as support vector machine(SVM), multi-layer perceptron(MLP) and k-nearest neighbor(k-NN) are extended by multi-boosting method in an array of subclassifiers and the results of each subclassifier are merged by ensemble method. Genetic algorithm was applied to obtain optimal feature set in each subclassifier. We applied our method to empirical data set from cancer research and the method showed the better accuracy and more stable performance than single classifier.

Key words : Bioinformatics, Pattern Recognition, 2-D PAGE, Genetic algorithm, Boosting, Ensemble classifier

요 약

생물 및 의학계에서는 생물정보학(bioinformatics)의 데이터 중 혈청 단백질(proteome)에서 추출한 데이터가 질병의 진단에 관련된 정보를 가지고 있고, 이 데이터를 분류 분석함으로 질병을 조기에 진단 할 수 있다고 믿고 있다. 본 논문에서는 혈청 단백질(2-D PAGE: Two-dimensional polyacrylamide gel electrophoresis)로부터 암과 정상을 판별하는 새로운 복합분류기를 제안한다. 새로운 복합 분류기에서는 support vector machine(SVM)와 다층 퍼셉트론(multi-layer perceptron: MLP)와 k-최근접 이웃(k-nearest neighbor: k-NN)분류기를 앙상블(ensemble) 방법으로 통합하는 동시에 다중 부스팅(boosting) 방법으로 각 분류기를 확장하여 부분류기(subclassifier)의 배열(array)으로서 복합분류기를 구성하였다. 각 부분류기에서는 최적 특성 집합(feature set)을 탐색하기 위하여 유전 알고리즘(genetic algorithm: GA)를 적용하였다. 복합분류기의 성능을 측정하기 위하여 암연구에서 얻어진 임상 데이터를 복합분류기에 적용하였고 결과로서 단일 분류기 보다 높은 분류 정확도와 안정성을 보여주었다.

주요어 : 생물정보학, 패턴인식, 2차원 전기영동법, 유전알고리즘, 부스팅, 복합분류기

1. 서 론

급속한 분자 생물학의 발달은 엄청난 양의 생물학 데이터를 생산하였고, 방대한 생물학 데이터의 정리 및 해석을 주목적으로 하는 생물정보학(Bioinformatics)이 생물학의 한 분야로 자리 잡게 되었다.

생물정보학은 생물학적 문재의 답을 구하기 위해 컴퓨터를 활용하여 데이터를 수집, 관리, 저장, 평가, 분석하는 일을 주 내용으로 하고 있다. 이러한 내용을 실현하기 위하여 기초생물학 및 응용 생물학, 의학, 약학은 물론이고 수학, 통계학, 물리학, 화학과 공학 등이 융합되어 생물정보학을 구성하고 있다.

생물 데이터를 얻는 과정에는 다양한 방식이 도입되고 있다. 특히 최근 대량 고속 분석 기기들이 데이터를 대량

접수일(2010년 -월 -일), 심사일(1차 : 2010년 -월 -일), 게재 확정일(2010년 -월 -일)
[1] 한국항공대학교 컴퓨터 공학과
주 저 자 : 온승엽
교신저자 : 온승엽
E-mail; syohn@kau.ac.kr

으로 생산 해 내고 있으며, Automatic DNA Sequencer, DNA microarray, Image Analyzer, Mass Spectroscopy 등이 이에 속한다. 이 왜에도 분자 생물학의 발달과 관련 고속 탐색기술(High throughput screening) 및 기기의 개발을 유도하여 다량의 생물학적 데이터를 양산하고 있다. 많은 경우에 image, signal 또는 pattern 등 image 형태의 데이터가 얻어지게 되는데, 여기에는 데이터의 전산학적인 처리가 중요하다. 유전 정보의 전달발현(Central Dogma)에의 최종산물인 프로테옴(proteome)[1]의 기능과 구조분석을 위해 단백체학(proteomics) 라는 또 다른 학문이 파생되어 본격적인 연구가 진행되어 그 기능이 하나씩 밝혀지고 있다.

현재 대량의 생물학 정보의 해석을 위한 생물정보학(Bioinformatics)은 데이터 분석용 알고리즘 및 프로그램의 개발과 데이터와 관련 지식들을 정리, 분석하는 부분으로 나눌 수 있다. 데이터베이스를 빠르게 정리하고 정확하게 검색하는 것도 중요하지만 생물학 정보들을 이용하여 기능을 예측할 수 있도록 데이터베이스를 가공하는 것에 좀 더 많은 초점이 주어지고 있다.

많은 연구가들은 생물정보학에서 얻은 단백질 패턴 정보를 데이터 마이닝 기법으로 분석하여 암과 정상을 구별하는 암 진단시스템을 연구 중이다. 이러한 시스템에 사용되는 기술은 분류(classification), 군집(Clustering), 특성추출(Feature Extraction) 등의 데이터마이닝 기법과 회귀분석, 상관분석 등 전통적인 통계학적 기법들을 사용하고 있다[1,2].

본 논문에서는 2D-PAGE로부터 얻은 프로테옴 데이터를 이용하여 암환자와 정상인을 예측 분류하는 암 진단 시스템을 제안하고 그 성능을 평가한다. 향상된 분류 정확도를 달성하기 위하여 잡음, 편이, 편차에 우수하고 특히 자료가 충분하지 않은 경우에 안정적인 결과를 산출하는 앙상블(ensemble)기법과 아다부스팅(Adaboosting) 기법을 복합적으로 응용하여 k-NN(k-Nearest neighbor), SVM(Support vector machine), MLP(Multi-layered perceptron와 같은 서로 다른 종류의 부분류기(subclassifier)를 배열(array) 형식으로 결합한 형태의 복합분류기를 제안한다. 배열 내부의 각 부분류기에서는 분류 모형을 최적화하기 위하여 유전 알고리즘(GA: Genetic algorithm)기법을 적용하여 특성집합(feature set)을 최적화 하였다. 제안된 분류 모형의 성능을 측정하기 위하여 단백체와 암의 관계를 밝히기 위한 의학 연구에서 얻어진 임상 데이터를 분류 모형에 적용하였으며 결과로서 제안된 복합 분류 모형이 단일한 알고리즘을 사용한 분류 모형 보다 정확도와 안정도가 향상되는 것을 보여주었다.

본 논문의 구성은 다음과 같다. 2장에서는 암 진단 시스템의 구성을 살펴보고 이와 관련된 데이터 추출 과정과, 분류기법, 앙상블 기법, 최적의 특성집합을 찾는 기법에 대해서 기술한다. 3장에서는 제안하는 다중 부스팅을 이용한 암 진단 시스템의 구성과 학습 과정에 대하여 설명하고 4장에서는 암진단 시스템의 평가(test) 과정과 평가 결과를 기술하였다. 5장에서는 결론 및 향후 연구 과제를 기술한다.

2. 관련 연구

2.1 단백질 분석 기법

암 진단 시스템에 사용되는 데이터는 단백질(proteome)의 분석을 통해 얻게 된다. 인체 내에서 기능을 가지는 단백질은 대략 3만~10만 개로 추정하고 있지만 변형 · 결합된 단백질까지 고려하면 그 수는 예측할 수 없게 된다. 특정 암의 표지자가 되는 단백질은 암환자와 정상인에서 발현 정도에서 차이를 보이게 되고 이를 분석하여 패턴인식 방법으로 분류해 내는 것이 암 진단 시스템이다. 단백질로부터 데이터를 획득하는 방법에는 전통적인 분석방법인 2D-PAGE(2 Dimensional Polycaylamide Gel Electrophoresis) 기법과 최신 질량분석법인 Seldi(Surface - Enhanced Laser Desorption/Ionization)-Tof, MALDI (Matrix Assisted Laser Desorption/Ionization)-Tof 기법이 있다. 2D-PAGE 기법 후 얻어진 젤(Gel)은 특수한 염색과 스캐닝을 통해 프로테옴 이미지를 얻게 된다. 이를 전용 이미지 처리 소프트웨어를 사용하여 단백질을 분석하게 된다. 많은 과학자들은 2D-PAGE가 MALDI/Ssldi-Tof 보다 더 풍부한 정보를 가지고 있다고 믿는다.

프로테옴 이미지 데이터의 전 처리 과정은 그림 1과 같다[3]. 프로테옴 이미지의 일반적 제작 과정은 먼저 길이 18cm Broad Range IPG Strip을 사용하여 시료 내에 존재하는 단백질을 각각의 등정점(PI)에 따라 2차원 전기영동 법에 의해 분리한다. 그 다음으로 단백질 분자량에 따른 SDS-PAGE를 이용한 분리과정인데 등정점에 의한 단백질의 분리 후 사용자가 요구하는 폴리아크릴 아미드의 농도(통상 8~16%)에 따라 구배 젤을 제조하여 크기에 따라 각각의 단백질을 분리한다. 마지막으로 젤을 발색하는데 전기영동 법으로 분리된 단백질의 이미지는 Colloidal Compassive Blue(CBB G250, R250), Sypro-Ruby, Silver

1) 프로테옴(Proteome)은 단백질(Protein)의 집합체를 말함.

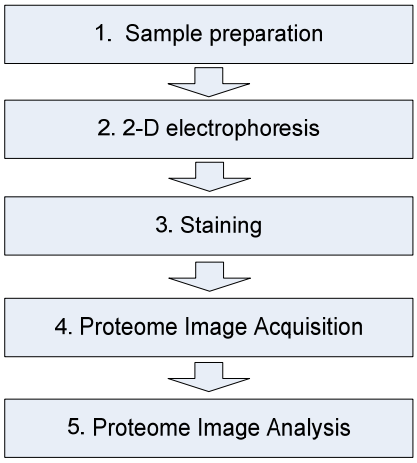

그림 1. 2D-PAGE 기법의 과정

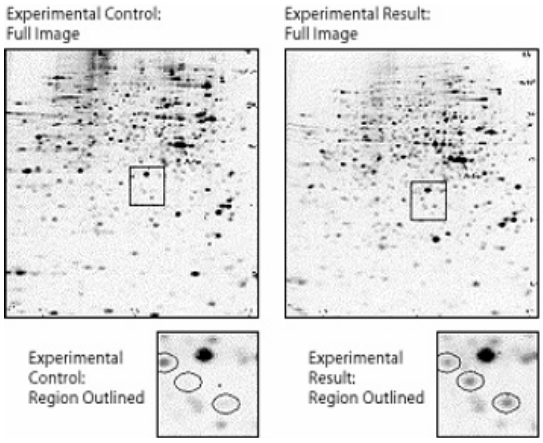

그림 2. 2D-PAGE 기법에 의해 생성된 프로테옴 이미지

Staining 등을 이용해 발색을 한다.

그림 2는 위와 같은 과정을 통해 얻은 프로테옴 이미지의 샘플이다. 각각의 프로테옴은 특정 위치의 스팟으로 발현 되며 각 스팟의 농도는 시료에 포함되어 있는 스팟과 관련된 프로테옴의 양을 의미한다. 그림 2에서 좌, 우측 이미지는 특정 위치 스팟의 농도가 시료에 따라 다르게 나오는 모습을 보여준다.

그림 3은 프로테옴 이미지 처리 과정이다. 획득된 프로테옴 이미지는 GS-710 CalibratedImaging Densitometer, Mole- cularImagerFX 등과 같은 프로테옴 이미지 전용 스캐너로 디지털화 하여 컴퓨터에 저장한다. 저장된 프로테옴 이미지는 프로테옴 이미지 분석 소프트웨어인 PDQuest를 이용하여 전 처리 작업, 스팟 검출(Spot Detection), 특

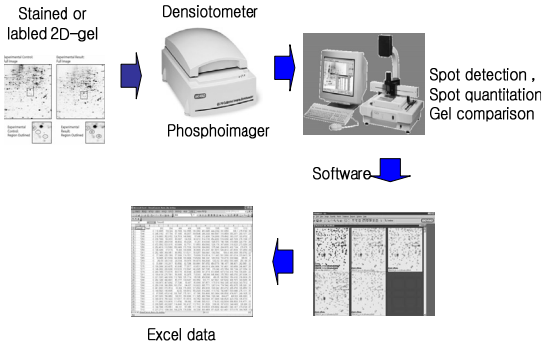

그림 3. 프로테옴 이미지 처리 과정

징 추출(Feature Extraction)작업을 수행한다. 이미지의 전 처리 작업은 Image Sizing과 Orientation, Matching, Filtering 등을 통해 조정한다.

하나의 기준 이미지를 선택한 후, 기준 이미지에서 사용자가 직접 스팟을 선택하거나 PDQuest가 제공하는 자동 검출 기능을 사용하여 스팟을 검출한다. 검출된 스팟은 일련번호를 주어서 라벨링(Labelling) 하고 마스터 이미지로 저장한다. 그림 2의 왼쪽은 정상인의 프로테옴 이미지이고 오른쪽은 암환자의 프로테옴 이미지이다. 아래의 작은 그림은 동일한 스팟에서 그 단백질의 발현정도가 차이를 보이는 'Key Spot'[2]을 비교하고 있다.

입력 이미지는 매치 셋(Matchset)이라고 하는 복사본으로 마스터 이미지와 비교해서 마스터에서 미리 라벨링 된 스팟에 가장 근접한 위치에 존재하는 매치 셋의 스팟을 찾아내고 이를 마스터와 같은 스팟으로 간주한다. 이런 스팟들은 마스터의 스팟과 같은 번호로 라벨링 하고, 라벨링된 스팟들은 PDQuest의 스팟 분석 툴을 이용하여 Center X, Center Y, Size X, Size Y, Peak Value, Quantity, Norm_Qty, Quality 필드의 이름을 갖는 고유의 값으로 엑셀파일에 저장한다. 각 위치에 존재하는 스팟은 해당 위치와 관련된 특정 프로테옴의 발현이며 각 스팟의 Peak Value는 혈액 중에 포함 되어 있는 해당 프로테옴의 양을 나타낸다. 생물학, 의학계에서는 다양한 종류의 프로테옴의 양 등이 생체의 상태 및 암과 같은 특정 질병의 유무를 나타낸다고 알려져 있다. 본 논문에서 제안 되는 진단 시스템에서는 선정된 다수의 스팟들로 구성된 집합의 Peak Value를 특성 값으로 삼아 패턴 인식 기법을 기반으로 암의 발병여부를 판단한다.

2) 종양표지자라고 추정되어지는 스팟.

2.2 바이오인포매틱스(BT)에서 사용되는 패턴 인식 기법

2.2.1 k-NN(k-Nearest Neighbor)

k-NN 분류 방식은 어떤 클래스에 속해 있는지 모르는 패턴 p를 가장 가까운 거리에 있는 클래스로 분류한다. 패턴분포에 관한 정보로 판별함수를 계산하는 대신에 미리 저장해 놓은 기준패턴과의 거리를 계산하여 가장 가까운 기준패턴의 클래스를 미지패턴의 클래스로 결정하는 방법으로 이와 같은 기본원리에 따른 최소거리 분류규칙을 k-최근접 이웃(k-NN: k-Nearest Neighbor) 분류규칙이라고 하며 이를 위하여 사전에 집단(Class)별 기준이 되는 표준패턴을 선정하여야 한다.

사용될 군집화 기준(clustering criterion)은 휴리스틱(heuristic) 기법으로 표현하거나 특정한 성능 지표의 최소(또는 최대)화에 바탕을 둘 수 있다. 휴리스틱기법은 경험과 직관을 바탕으로 한 방법으로, 패턴을 군집 영역에 배정하기 위해 선택한 유사도를 이용하는 규칙들로 구성되며 유클리드 거리 척도가 근접 척도로서 쉽게 해석되므로 이 접근 방식에 적합하다.

2.2.2 SVM(Support Vector Machine)

Support Vector Machine(SVM)은 1995년 Vapnik에 의해 제안된 커널을 이용한 기계학습 방법으로서 통계적 학습 이론에 기반 하여 최적 분류를 행함으로써 뛰어난 일반화 성능을 보여준다[6]. SVM은 다양한 응용분야를 가지며, 각 분야에서 SVM을 적용한 논문들이 성능 향상을 보고 함으로서 이론적 근거를 공고히 하고 있다[7].

SVM은 기존의 신경망 등에서 이용된 경험적 위험(Empirical Risk)을 최소화하는 원리와 달리 구조적 위험(Structural Risk)을 최소화 하는 근사적 구현이다. 또한 VC-dimension을 최소화하여 에러의 확률분포도 작게 하며 일반화 특성을 좋게 한다. 이는 패턴을 커널(Kernel)을 통하여 고차원의 특성공간으로 사상하여 대역적으로 최적의 식별하게 하기 때문이다.

SVM의 목적은 학습 자료로 주어지는 두 개의 부류를 구분하는 함수를 추정하는 것이다. 이 함수는 n 차원의 벡터 공간의 n - 1 차원의 초평면(hyperplane)의 형태로 나타난다. 이러한 평면은 무수히 많이 존재 할 수 있지만, 두 부류 간에 모든 점들 사이의 거리를 최대화하도록 제한을 두면 하나의 유일한 평면만이 해로 나타난다. 이 선형 평면 분류 경계를 OHP(optimal hyperplane)이라 한다. SVM은 이러한 OHP를 찾는 과정이라 할 수 있다.

2.2.3 MLP(Multilayer Perceptron)

신경망 또는 인공신경망(artificial neural networks)에 관한 연구는 뇌 신경생리학(neurophysiology)으로부터 영감을 얻어 시작되었다[8]. 자료 분석 분야에서 신경망은 복잡한 구조를 가진 자료에서의 예측(prediction) 문제를 해결하기 위해서 사용되는 유연한 비선형모형(nonlinear models)의 하나로 분류될 수 있다. 그러나 신경생리학과의 유연성 때문에 일반적으로 다른 통계적 예측 모형에 비해 보다 흥미롭게 받아들여지고 있다. 신경망은 은닉마다(hidden units)라고 불리는 독특한 구성요소에 의해서 일반적인 통계모형과 구별되어 진다. 은닉마디는 인간의 신경세포를 모형화한 것으로써 각 은닉마디는 입력변수들의 결합(combination)을 수신하여 목표변수에 전달한다. 이때 결합에 사용되는 계수(coefficient)들을 연결강도(synaptic weights)라고 부르며, 활성함수는 입력 값을 변환하고 이를 입력으로 사용하는 다른 마디로 출력하게 된다.

신경망에는 여러 가지 다양한 모형이 있으나, 그 중에서도 자료 분석을 위해 가장 널리 사용되는 모형은 MLP (Multilayer Perceptron, 다층인식자) 신경망이다. MLP는 입력층(input layer), 은닉마디로 구성된 은닉층(hidden layer), 그리고 출력층(output layer)으로 구성된 전방향(feed-forward) 신경망이다.

2.3 Ensemble Classification 기법

Ensemble Classifier는 다양한 분류기 또는 분류작업을 통해 나온 예측치 들을 투표형식(Voting)으로 통합하여, 최다수 투표를 얻은 클래스가 최종 결과로 결정하는 방법이다. 이러한 Ensemble 방법으로는 Breiman(1996)의 Bagging(Bootstrap AGGregatING), Fruend와 Schapire (1996)의 Adaboosting(Adaptive Boosting), Quinlan(1998)의 다중 boosting 방법 등을 들 수 있다. 이중 대표적인 Bagging과 AdaBoosting에 대하여 소개하기로 한다.

2.3.1 Bagging

Bagging 기법은 Bootstrap 샘플에 기반 한 것으로 각 분류기의 학습샘플 집합은 N개의 예제로 구성되는 원래의 학습 집합으로부터 무작위로 N개의 예제를 추출함으로써 생성된다.

Bagging 기법을 사용한 분류 과정은 동일하게 복원 추출(Replacement Sampling)을 통해서 크기가 N인 Bootstrap의 샘플링과 같은 방법으로 생성하여 각각의 Subclassifier를 학습시킨다. Bagging 기법의 최종판정은 subclassifier의 분류 결과를 다수결 투표방식으로 결합하여 결정하게

된다. 이러한 Bagging 기법은 결정트리(Decision Tree) 기반 분류기와 같은 불안정한 분류과정의 편차를 현저히 줄여주며, 결과적으로 정확성 향상을 기할 수 있게 해 준다.

2.3.2 AdaBoosting

아다부스팅(AdaBoosting: Adaptive Boosting)은 Freund and Schapire가 제안하였으며[9] 알고리즘의 원리는 "Weak" 또는 "Base" Learning Algorithm을 반복해 가면서 오분류 되는 자료들의 가중치(sample weight)를 증가시켜서 학습 데이터에 포함될 확률을 높여 분류 알고리즘이 오분류 되는 자료들에 더욱 집중하여 학습하도록 하고 학습시키고, weak classifier의 결과를 분류 정확도에 따른 가중치에 의하여 가중투표(Weighted Voting) 방식으로 통합하는 방법이다.

2.4 유전 알고리즘(Genetic Algorithm)

1970년대 초 John Holland 의해 본격적으로 연구되기 시작한 유전 알고리즘은(GA)는 자연 생태계의 진화과정에서 관찰된 몇 가지 처리 과정 중에서 '적자생존(survival of the fittest)'의 원리를 컴퓨터 알고리즘과 결합시켜 정립된 최적화(optimization) 알고리즘이다[8].

유전 알고리즘은 자영생태계의 진화 메커니즘을 모방하였는데 실제로 자연계의 진화과정이 모두 밝혀져 있지는 않지만 중요한 몇 가지는 알려져 있어 유전 알고리즘은 이러한 진화과정에서 일부 관찰된 것을 사용하였다. 자연생태계의 진화과정을 간단히 살펴보면 다음과 같다.

진화는 염색체(chromosome)-생명체의 구조를 부호화하기 위한 기관소자에서 일어나며 생명체는 염색체를 복호화 하는 과정을 통해 부수적으로 창조된다. 염색체의 부호와-복호화 과정의 자세한 것은 알 수 없어도 일반적으로 인정되고 있는 몇 가지 과정은 다음과 같다.

진화는 부호화되는 생명체가 아닌 염색체에 대해 작용하는 과정이다.

자연도태(natural selection)는 염색체와 복호화 된 구조의 수행과의 관계이며, 훌륭한 구조의 염색체를 보다 자주 부호화 해주는 과정이다.

재생산(reproduction)과정은 진화가 일어나는 시점이다. 돌연변이는 자손의 염색체와 어버이의 염색체를 다르게 할 수 있고 재 조합처리는 어버이 염색체들을 결합함으로써 매우 다른 염색체의 자손을 만들어 준다.

3. 다중 부스팅을 이용한 암진단 시스템

3.1 다중 부스팅을 이용한 복합 classifier의 구현

학습 데이터의 수는 분류기의 학습 강도에 큰 영향을 미친다. 그러나 생물학 분야에서 학습하기 충분한 데이터의 수집하기에는 엄청난 비용이 들어 표본의 수집도 용이하지 않다. 이러한 이유로 자료의 재현성(reproducibility)과 안정성이 떨어지는 문제가 되기도 한다. 이러한 특성에서 좋은 성능을 보인다고 알려진 방법이 앙상블 기법이다.

고전적인 앙상블 알고리즘은 각각의 Weak-learned 부분류기들을 'Ensemble' 개념을 도입하여 Strong-learned Classifier로 투표와 가중치를 적용 하여 통합하는 방법을 말한다.

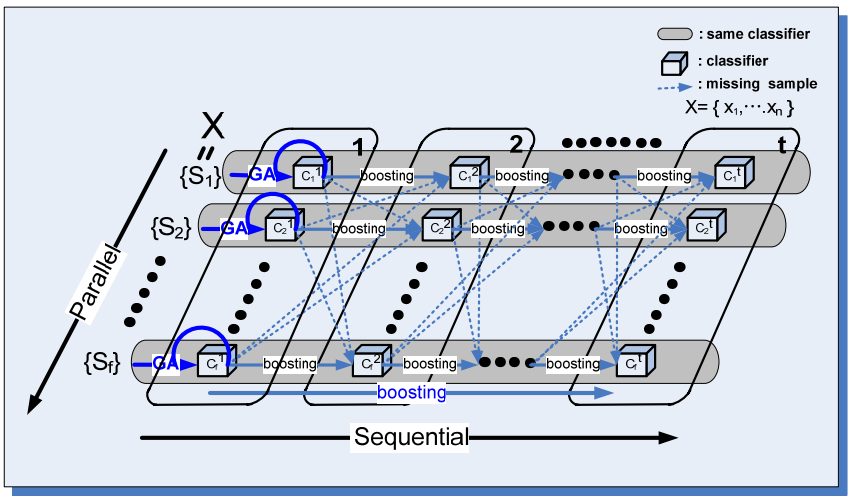

그림 4. 다중-부스팅 분류 모형의 구성

이는 결합을 통해 그 성능을 강화하지만 그 분류기가 갖는 구조적 약점을 반영하지는 못한다. 이러한 문제를 해결하기 위하여 본 논문에서는 서로 다른 방법론 또는 특징을 갖는 분류기 들이 서로 상호 보완적으로 작용하도록 함으로써 분류기의 성능을 향상 시키고자 하는 것이다.

그림 4는 본 논문이 제안하는 다중 부스팅 기반의 복합 분류기를 나타낸 것이다. 복합분류기 내의 배열의 각 행은 동일한 구조를 가지는 분류기로 구성되어 있으나 각각의 분류기는 다중 부스팅에 따라 생성된 서로 다른 학습 데이터 세트에 의하여 학습 되고 유전알고리즘에 의하여 최적화된 특성 집합을 가진다. 초기화 단계에서는 전체 학습 데이터 세트로부터 부트스트래핑(bootstrapping) 방법에 의하여 생성된 학습용 집합(training dataset)에 의하여 각 부분류기가 학습 되어 진다. 이후 단계부터는 다중 부스팅 방법에 의하여 각각의 subclassifier에 대하여 오분류 되는 경향이 높은 샘플들을 학습 데이터로 하여 새로운 분류 모형을 만드는 sequential 단계가 병렬로 진행된다.

본 논문의 암 진단 시스템 구현에 사용된 부분류기들은 각각 polynomial, RBF, sigmoid 함수를 kernel 함수로 가지는 세 종류의 SVM, 역전파 알고리즘에 기반 한 MLP, 각각 Euclidean, City-block, Mahalanobis 거리를 유사도측정방법으로 하는 세 종류의 k-최근접이웃 분류기를 부분류기로 사용하였다.

3.2 부분류기 배열 생성 알고리즘

부분류기 배열을 생성해나가는 과정은 다음과 같다.

우선 배열의 첫 단계에 부분류기들을 배치시킨다. 부분류기들은 서로 다른 알고리즘을 가지거나 다른 파라미터를 가지는 분류기들이다. 각각의 부분류기는 전체 학습 데이터 세트로부터 부트스트래핑 방법으로 추출된 데이터세트 S_i 을 학습 하게 되는데, 각각의 부분류기 D_i 에 대해서 유전알고리즘을 통하여 최적의 특성집합을 찾게 된다.

둘째, 각각의 부분류기 D_i 는 부스팅 기법을 통하여 오분류 된 표본들에 가중치가 적용 되어 Strong-learned 부분류기로 학습하게 된다. 가중치 적용 기법은 전통적인 아다부스팅 기법을 기초로 전 단계에서 오 분류된 학습 표본의 가중치를 합하는 방법을 제안한다.

다음은 subclassifier array 생성 알고리즘이다.

Step 1.(초기화과정) 학습용 데이터세트를 부트스트래핑방법으로 표본화(sampling)하여 F개의 부분집합(subset), S_1^1, S_2^1, ...,S_F^1을 만들고 $t=1$로 놓는다.

Step 2.(단계 t의 부분류기생성) 각각의 부분집합 S_i^t를 학습 데이터로 사용하여 부분류기 D_i^t를 생성한다. 각 부분류기는 유전알고리즘을 통하여 최적화된 특성집합을 가지게 된다.

Step 3.(단계 t + 1의 데이터 세트 준비) 다음 단계에서 생성될 부분류기 D_i^{t+1}의 학습 데이터로 학습 데이터 세트 S_i^{t+1}를 생성한다. S_i^{t+1}는 S_i^t로부터 아다부스팅 방법으로 추출한 샘플로 구성된 집합에 단계 t에서 D_i^t 이외의 부분류기에서 오분류된 샘플들을 추가한 데이터세트이다.

Step 4. $t=t+1$로 놓고 Step 2, 3를 부분류기 배열의 오분류율이 원하는 수치 이하로 감소하거나 미리 정하여진 횟수만큼 반복한다.

부분류기 배열에서 i 행 t 열에 위치한 부분류기 D_i^t에 대한 중요도 w_i^t는 다음 식에 의하여 정의된다.

$$w_i^t = \frac{1}{Z^t}(\log(\frac{1-\epsilon_i^t}{\epsilon_i^t}) + \alpha|S_i^t|)$$

위 식에서 ϵ_i^t는 D_i^t의 오분류율이며 α는 D_i^t에 주어진 학습용 데이터 세트 S_i^t의 샘플 개수에 따라 중요도를 증가시켜 주기 위한 상수이다. 또한, Z는 배열 내의 모든 부분류기의 중요도의 합이 1이 되도록 해주는 정규화 요소(normalization factor) 이며 T개의 행과 F개의 열을 가지는 부분류기 배열의 경우 다음과 같이 계산될 수 있다.

$$Z = \sum_{t=1}^{T}\sum_{i=1}^{F}(\log(\frac{1-\epsilon_i^t}{\epsilon_i^t}) + \alpha|S_i^t|)$$

표본(sample) x에 대한 T개의 행과 F개의 열을 가지는 복합 분류기의 예측 D^Γ은 중요도를 가중치로 하는 투표방식에 의하여 결정되며 다음 식과 같이 계산된다. 이 식에서는 각 샘플을 -1과 +1의 두 범주로 분류하는 이진분류를 고려하였으며 $D_i^t(\mathrm{x})$는 샘플 x에 대한 부분류기 D_i^t의 예측 이며 -1 또는 +1의 값을 가진다.

$$D^\Gamma = \mathrm{sgn}\sum_{t=1}^{T}\sum_{i=1}^{F} w_i^t D_i^t$$

3.3 성능 평가

3.3.1 데이터의 구성과 실험

실험 데이터는 표 1과 같이 유방암(Breast Cancer)환자와 폐암(Lung Cancer)의 혈액샘플로부터 혈장(serum)

을 추출하여 2D-PAGE 기법으로 추출하였다. 실험에 사용된 유방암 데이터의 수는 샘플 301개(암 151, 정상 150)이며, 특성은 151개로 구성 되어 있으며, 폐암 환자의 수는 표본의 총수는 120개(암 66, 정상58)이고 특성은 23개로 구성 되어 있다.

얻어진 학습 데이터는 복원 추출(replacement sampling)을 통해서 7개의 데이터세트로 나누어 kernel function을 다르게 한 SVM 3종, MLP 1종, KNN의 유사도 측정 방법을 다르게 한 3종, 모두 7종의 부분류기를 학습시켰고 3장에서 소계 한 것처럼 유전알고리즘를 통하여 20%의 최적해(optimal feature)를 찾아 학습 시켰다.

본 논문에서 사용한 부분류기의 파라미터는 표 2, 표 3, 표 4와 같다.

표 1. 데이터 구성

종 류	암	정상	Feature 수
유방암 (Breast Cancer)	151	150	151
폐암 (Lung Cancer)	66	58	23

표 2. SVM 파라미터

Parameter	유방암 (Breast Cancer)	폐암 (Lung Cancer)
C	316	10085
Class cost	NONE	NONE
Gamma	0.0001	0.00014905
Optimal feature 수	30	4

표 3. MLP 파라미터

Parameter	유방암 (Breast Cancer)	폐암 (Lung Cancer)
Hidden layer 수	2	1
Node 수	5,3	3
Initial weights	1.00	1.00
Iterations	1000	1000
Optimal feature 수	30	4

표 4. KNN 파라미터

Parameter	유방암 (Breast Cancer)	폐암 (Lung Cancer)
Neighbors 수	5	3
Optimal feature 수	30	4

본 논문에서 제안한 다중 부스팅 분류기의 성능을 평가하기 위하여, 성능 평가에 우수하다고 알려진 k-겹 검증(k-fold cross validation) 기법으로 10개의 표본집합으로 나누고 1개의 부분집합을 제외한 나머지 9개를 학습시키고 제외된 1개 집합으로 평가하는 방법으로 반복하여 총 100회를 측정하였다. 또한 단일 분류기와 그 성능을 비교 평가 하였다.

3.3.2 결과 및 분석

3.2.1절에 소계 한 방법으로 그 결과를 정리한 결과를 표 5, 표 6에 정리 하였다. 그림 5에서 유방암 Error rate를 비교해 보면 다중 부스팅 기법의 성능이 23.8%로 오분류 율을 보임으로써 가장 우수함을 보이고 있다. 그림 6의 분산 분석에도 두 번째 좋은 성능을 보인 SVM(RBF) 보다 좋은 편차를 보였다. 반면 폐암 결과에서는 그림 7에서 보듯이 전반적으로 SVM, MLP, KNN순의 성능차이를 보였으나, SVM과 다중 부스팅과의 성능차이는 근소한 것으로 나타났다. 그림 8에서 보듯이 분산도 SVM(RBF)와 다중 부스팅 기법이 다르지 않아 큰 성능의 차이를 보이지 않았다. 이는 base 분류기가 안정적인 성능

표 5. 유방암 데이터에 대한 오분류율 비교 결과

분류 알고리즘	평균	편차	중앙값
다중 boosting	23.80	3.60	24.57
SVM(RBF)	28.02	8.60	26.67
SVM(sigmoid)	33.16	8.83	33.33
SVM(polynomial)	39.38	9.25	38.10
KNN(Euclidean)	40.30	8.543	40.97
KNN(city-block)	45.98	4.32	46.66
KNN(Mahalanobis)	41.68	8.28	43.33
MLP	33.64	8.44	33.33

표 6. 폐암 데이터에 대한 오분류율 비교 결과

분류 알고리즘	평균	편차	중앙값
다중 boosting	12.92	7.49	15.38
SVM(RBF)	13.50	8.67	15.38
SVM(sigmoid)	14.25	9.14	15.38
SVM(polynomial)	14.25	9.14	15.38
KNN(Euclidean)	18.73	10.5	18.18
KNN(city-block)	53.22	1.64	53.85
KNN(Mahalanobis)	17.19	9.74	15.38
MLP	15.61	9.57	15.38

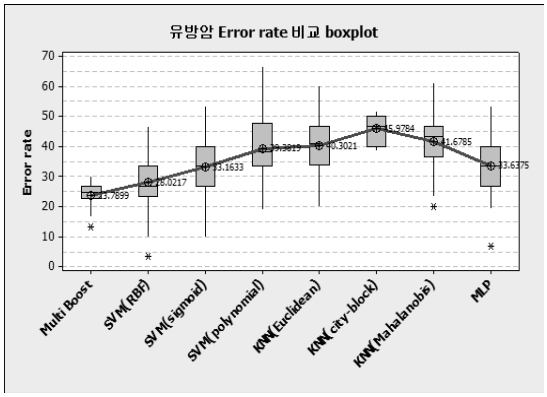

그림 5. 유방암 데이터에 대한 오분류율 비교 boxplot

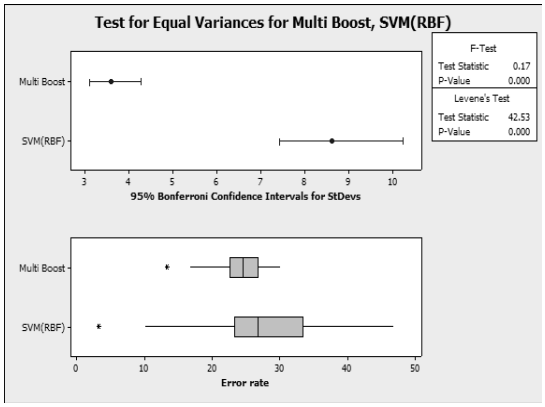

그림 6. 유방암 분산 검정

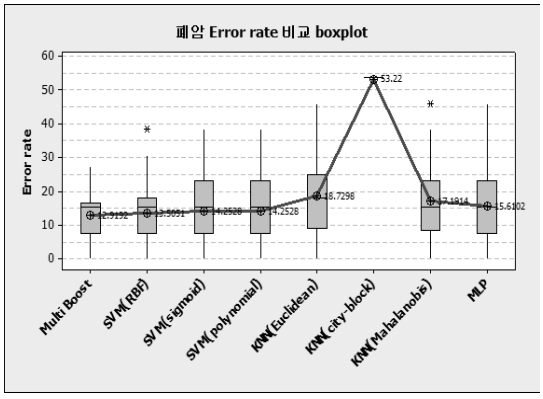

그림 7. 폐암 데이터에 대한 오분류율 비교 boxplot

을 보이므로 다중 부스팅 기법에도 그 영향을 주는 것으로 판단된다.

결과에서도 나타 났듯이 다중 부스팅 기법은 base 분류기의 안정성에 따라 차이가 있으나 결과적으로 모집단

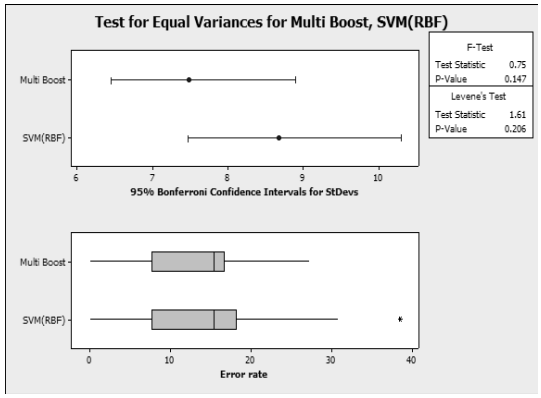

그림 8. 폐암 분산 검정

과 표본 집단 사이의 편차와 편이를 줄이는 역할을 하게 됨을 알 수 있었다.

4. 결론 및 향후 연구

본 논문에서는 여러 종류의 부분류기에 대하여 유전자 알고리즘(GA : Genetic Algorithm)를 통하여 최적해(optimal feature)를 찾고 부분류기 집합을 다중 부스팅기법을 통하여 통합한 복합 분류기를 제안하였다. 또한, 모형의 성능 평가 과정에서는 임상데이터를 사용하였으며 객관적인 평가를 위하여 검증(cross validation)기법을 사용하였으며, 단일 분류기와 그 성능을 비교 평가 하였다. 평가 결과 단일 분류기보다 4.2%까지의 평균적인 성능 향상을 보여주었으며, 정확도의 분산 측면에서도 단일 분류 모형보다 안정적임을 보여 주었다.

향후 분류 모델의 정확도 향상 기법을 최적화하기 위하여, 각 분류기에서 최적의 파라미터를 찾는 방법, 배깅(bagging) 기법과의 병합 방법과 차원 축소(dimensionality reduction) 방법의 비교 평가 연구가 추가로 연구 해 보아야 할 필요성이 있다. 나아가 조기 암 진단을 위한 마이크로 어레이(Microarray) 분석과, 임상소견 및 병리소견(Clinico Pathlogic data)를 이용하여 생존율과 재발율을 예측하는 시스템 개발의 필요성도 대두 되고 있다.

참 고 문 헌

1. B. Krishnapuram, L. Carin, and A. Hartemink, “Joint Classifier and feature optimization for cancer diagnosis using gene expression data”, Proceedings of the seventh annual international conference on computational molecular

biology, pp. 167-175, 2003.
2. S. Ando, and H. Iba, "Classification of Gene Expression Profile Using Combinatory Method of Evolutionary Computation and Machine Learning", Genetic Programming and Evolvable Machines, Volume 5 Issue 2, 2004.
3. "PDQuest User Manual", http://www.bio-rad.com
4. Ha-Nam Nguyen, Syng-yup Ohn and Ohn Woo-Jin, Combined Kernel Function for Support Vector Machine and Learning Method Based on Evolutionary Algorithm." 1273-1278.
5. R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd Ed., Wiley Interscience, New York, 2001.
6. V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, Berlin Heidelberg, New York, 1995.
7. J. P. Anderson, "Computer Security Threat Monitoring and Surveillance", James P Anderson Co., Technical report, Fort Washington, Pennsylvania, April 980.
8. Adaptation in Natural and Artificial Systems, Ann Arbor: The University of Michigan Press, 1975. (Second edition printed in 1992 by MIT Press, Cambridge, MA.)
9. Y. Freund, and R. Schapire, "Experments with new boosting algorithm" Proc. Of the 13th International Conference on Machine Learning, pp. 148-156, Bari, Italy, 1996.
10. Ljubomir J.Buturovic PCP-Pattern Classification Program, version1.2 http://pcp.sourceforge.net
11. C.-C. Chang, C.-J Lin, "LIBSVM : a Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf, August 12,2004.
12. C.-W. Hus, C.-C. Chang, C.-J. Lin, "A Paractical Guide to Support Vector Classification," http://www.csie.ntu.edu.tw/~cjlin/paper/guide/guide.pdf
13. Dong Seong Kim, Ha-Nam Nguyen, Jong Sou Park "Genetic Algorithm to Impove SVM Based Network Intrusion Detection System" IEE Computer Society Press.

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