# SNP 데이터의 중요도 평가와 SVM 학습법을 이용한 폐암 감수성 예측

## Prediction of Lung Cancer Susceptibility using an Importance Evaluation of SNP Data and SVM Learning

류명춘*, 김상진*, 박창현**
경운대학교 컴퓨터공학과*, 영남대학교 전자정보공학부 컴퓨터공학전공**

Myung-Chun Ryoo(mcryoo@ikw.ac.kr)*, Sang-Jin Kim(sjkim@ikw.ac.kr)*,
Chang-Hyeon Park(park@yu.ac.kr)**

**요약**

본 논문에서는 폐암의 발생에 관여하는 유전자 데이터인 SNP 데이터의 중요도 평가와 SVM 학습법을 이용하여 폐암 감수성을 예측하는 방법을 제안한다. 학습에 사용될 폐암 관련 양성 데이터에 비하여 음성 데이터의 수가 훨씬 많은 이유로 각 양성 데이터에 대하여 같은 성별과 적은 나이 차를 갖는 음성 데이터를 찾아서 쌍이 되도록 한다. 또한 각 SNP가 발병 예측에 미칠 영향력을 계산하는 수식을 도입하여 각 SNP의 중요도를 평가하고 SNP를 중요도에 따라 서열화 한다. 실험에서는 학습에 사용되는 순위별 SNP 개수에 따라 변화되는 예측률을 관측하였고, LOOCV 테스트 결과 제안된 방법은 실험 데이터에 대하여 최대 65.0%의 예측 정확도를 보였다.

■ 중심어 : | 기계 학습 | 지지벡터머신 | 인간 유전체 | 단일염기다형성 | 폐암 |

**Abstract**

In this paper, we propose a prediction method of lung cancer susceptibility using an importance evaluation of SNP data and the SVM learning, a gene data concerning getting sick with the lung cancer. Since the number of negative data is much larger that of positive data, which are to be used in the SVM learning, for each positive data, a negative data is first searched which has the same sex and the minimum age difference with the positive data. The searched negative data is then coupled with the positive data. For the importance evaluation of each SNP data, an equation which calculates the influence of each SNP data on the prediction of getting sick is adopted. The SNP data are sorted according to the evaluated importance. In experiments, we observed the prediction accuracy which varies according to the number of sorted SNP data used in the learning. LOOCV test results showed that the proposed method yields the prediction accuracy of maximum 65.0% for test data.

■ keyword : | Machine Learning | SVM | Human Genome | SNP | Lung Cancer |

## I. 서 론

게놈(Genome) 프로젝트로 인하여 유전자 데이터가 기하 급수적으로 증가하면서 이들 데이터와 전산학의 한 분야인 기계 학습(machine learning)기법을 이용한 질병의 분류, 예측 및 진단을 위한 다양한 연구가 시도

접수번호 : #080929-001
접수일자 : 2008년 09월 29일
심사완료일 : 2008년 10월 21일
교신저자 : 류명춘, e-mail : mcryoo@ikw.ac.kr

되고 있다[1][2].

질병에 대한 개개인의 감수성은 여러 가지 물리적, 환경적, 유전적 인자의 노출 정도에 따라 달라지지만 최근에는 개인의 유전자(gene)에 대한 관심이 매우 높아지고 있다. 인간의 유전체(human genome)는 약 30억 쌍의 염기로 구성되며, 약 1,000개의 염기쌍마다 염기의 변이가 일어나는데 이를 SNP(Single Nucleotide Polymorphisim, 단일 염기 다형성)라고 하며 인간 유전체의 0.1% 정도가 이에 해당한다[3][4]. 이와 같은 SNP로 인해 개인의 피부, 머리카락 색깔, 체질, 질병 감수성 등과 같은 개인과 인종의 유전적 특성이 다르게 나타나게 된다.

보통 질병 관련 유전자라고하면 하나의 유전자에 의해 관련 질병의 발생 유무를 판단한다고 생각하는 것이 일반적이다. 이런 질환은 환경의 영향이 적고 관련 유전자에 결함이 있으면 해당 질병은 반드시 나타나게 되므로 유전자 검사를 통해서 간단하게 질병의 조기 진단이 가능하다. 대표적인 단일 유전자 질병(monogenic disease)로는 고세병(Gaucher's disease), 겸상적혈구빈혈증(sickle-cell anemia) 등이 있다. 그러나 각종 암, 고혈압 등의 질환은 다수의 유전자와 환경의 복합 작용에 의해 일어나는 질병이기 때문에 단일 유전자 질병에 비해서 유전자 검사를 통한 질병의 조기 진단이 어려운 편이다.

2007년 통계청 발표에 따르면 폐암은 발병률 2위, 사망률 1위인 가장 치명적인 질병이지만 조기 진단 및 발생 위험도의 예측이 가장 어려운 암 중의 하나이다. 최근 폐암에 관련된 연구들을 살펴보면 폐암 발생에 관여하는 SNP를 분석하거나 관련 SNP를 발굴하는 연구가 상당수를 차지하고 있으며, SNP를 이용한 폐암 감수성 예측에 관련된 연구로는 회귀분석법에 의한 연구[5]가 있었으며, 본 논문에서 사용한것과 같은 임상데이터를 대상으로 예측률 65.9%를 얻었다. 이 연구에서는 폐암에 관여하는 7개의 유전적 SNP를 규명하고 폐암의 유전적 위험인자를 검사하기 위해 7개의 SNP에 대한 진단용 키트를 개발한 후 이를 이용하여 폐암 발생 가능성 여부를 조사하였다.

본 논문에서 사용한 임상 데이터는 폐암과 관련이 있을 것으로 추정되는 50개의 SNP를 포함하고 있으며 환자와 정상인의 SNP값들을 분석해 본 결과 이들 가운데는 기존 연구[5]에서 사용된 7개의 SNP보다 더 의미있는 SNP도 포함되어 있는 것으로 판단되었다. 또한, 복합 유전적 질병인 폐암의 경우에는 다수의 SNP가 복합적인 작용에 의해 발생하는 경우가 많기 때문에 폐암에 관여하는 7개의 SNP를 이용하는 방법 보다는 더 많은 수의 SNP를 이용한 폐암 감수성 예측 방법이 필요할 것으로 보였다. 따라서 본 논문에서는 폐암 발생에 관여하는 것으로 추정되고 있는 50개의 SNP 데이터에 대한 중요도 평가와 더불어 실험 데이터의 이진 분류(양성 혹은 음성)를 위해 기계 학습 기법 중의 하나인 SVM (Support Vector Machine, 지지벡터머신) 학습 기법을 이용하여 폐암 중에서 발생 빈도가 가장 높은 편평 상피암의 감수성을 예측하고자 한다.

본 논문의 구성은 II장에서는 유전자 데이터와 SVM 학습법에 대해서 논하고, III장에서는 SVM 학습 시 예측의 정확도를 높이기 위한 임상 데이터의 매칭 방법과 실험에 사용되는 SNP 데이터들이 폐암 발생에 미치는 중요도를 계산하기 위한 SNP 데이터의 중요도 평가 및 예측 방법에 대해 설명하고, IV장에서는 실험데이터와 SVM 학습 모델을 이용하여 폐암 감수성 예측률을 구한다. V장에서는 결론과 향후 과제를 제시한다.

## II. 관련 연구

### 1. 유전자 데이터

생명체는 세포(cell)들로 이루어져 있으며, 하나의 세포는 세포막(cell membrane), 세포질(cytoplasm), 세포핵(nucleus), 엽록체(chloroplast) 등으로 구성된다. 세포핵 안에는 유전의 근본물질인 유전자가 담겨 있는 염색체(chromosome)가 있으며, 이 염색체는 DNA (DeoxyriboNucleic Acid, 데옥시리보핵산)라는 물질로 구성된다. 유전자는 일정한 수와 순서를 가진 염기쌍의 일부분을 말하는 것으로 세포내 유용한 물질을 만들어 내는 정보를 가진 단위이다. 세포핵 속에는 성을 결정하는 염색체를 포함하여 모두 23쌍의 염색체가 있다.

한 개의 염색체에는 수천 개의 유전자가 들어 있으며, 인간 유전체는 약 30억쌍의 DNA 염기로 구성되어 있다.

대표적인 유전자 데이터로는 A(Adenine, 아데닌), C(Cytosine, 시토신), G(Guanine, 구아닌), T(Thymine, 티민) 등 4종류의 염기로 이루어진 DNA 염기서열(base sequence) 데이터와 A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y 등 20종류의 아미노산 잔기서열(residue sequence)로 이루어진 단백질 서열 데이터가 있다[6].

본 논문에서는 전자의 데이터 형식을 가지며 폐암 발생에 관여하는 것으로 추정되는 50개의 SNP 데이터를 사용한다. 각 SNP의 유전자 타입(genotype)은 이형접합(heterozygous) 1개, 동형접합 (homozygous) 2개를 가진다. 예를 들어 어떤 한 유전자의 특정부위에 존재하는 SNP(C→T)는 CC, CT, TT의 세 가지 타입을 가진다.

## 2. SVM 학습법

기계 학습은 인공지능의 한 영역으로 주어진 데이터의 집합을 이용해서 데이터의 속성에 관한 정보를 추론하는 학습 알고리즘에 관련된 것으로, 대규모의 데이터 세트에서 관심 있는 패턴을 찾기 위해 다양한 분야에서 사용되고 있다.

최근에는 다양한 분야에서 SVM 기계 학습법을 이용한 연구가 이루어지면서 그 성능을 인정받고 있다[7][8]. 예를 들면 질병 관련 연구[9]에서 67.1%, 안면 인식과 필기체 인식 연구[10]에서 89.6%, 불법 침입자 탐지 분야의 연구[11]에서 80.1%의 정확도를 보였다.

SVM은 이진 분류기법으로 두 그룹의 데이터를 구분시키는 분류경계(hyperplane)와 가장 가까운 거리에 있는 학습 데이터와의 거리를 최대화 시키는 최대 여백 분류 경계(maximal margin hyperplane)를 찾는 학습 방법이다[12]. SVM에서 최대 여백 분류 경계를 찾는 이유는 새로운 테스트 데이터를 분류(양성 혹은 음성)할 때 발생할 수 있는 오류를 줄이기 위해서이다. 분류경계는 커널 함수로 표현되며 SVM에서 가장 널리 사용되는 커널 함수로는 선형 커널(linear kernel), 다항 커널(polynominal kernel), RBF 커널 등이 있다. 이 중에서 선형 커널은 가장 빠르고 간단한 커널 함수로서 SVM 기반 응용에 많이 사용되고 있다. [그림 1]은 SVM을 이용하여 주어진 학습 데이터를 두 개의 그룹으로 분류한 것이다. [그림 1]에서 (a)는 최소 여백 분류 경계에 의한 이진 분류의 예이며, (b)는 최대 여백 분류 경계에 의한 이진 분류의 예이다. 이러한 최대(최소) 여백 분류 경계에 가장 근접한 학습 데이터를 지지 벡터(support vector)라고 한다.

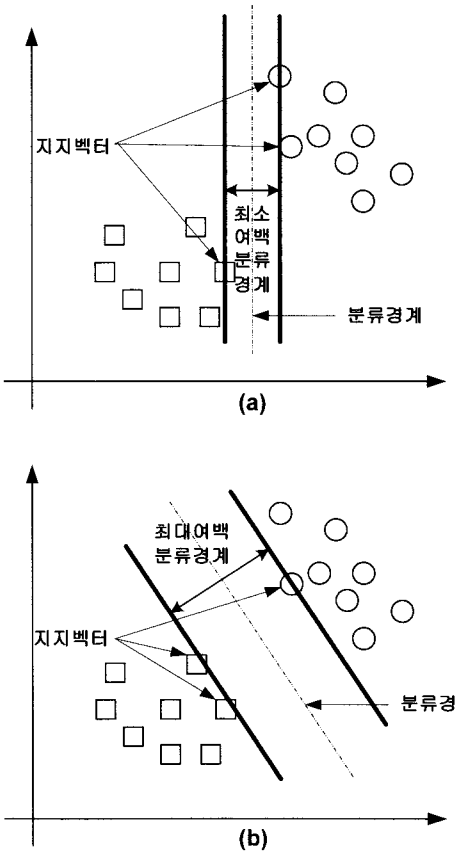



그림 1. SVM 학습법을 이용한 데이터 그룹의 이진분류

## III. SVM 학습을 위한 데이터 전처리 및 예측방법

실험을 위해서 폐암 발생 환자(양성) 320명과 정상인

(음성) 720명으로부터 폐암 발생에 관여하는 것으로 추정되는 50개의 SNP와 관련된 임상 데이터를 (주)디앤피 바이오텍에서 제공받아 사용한다.

일반적으로 특정 질병에 대한 임상 데이터는 양성 데이터가 음성 데이터에 비해 훨씬 적으며 이러한 임상 데이터를 SVM 학습에 그대로 사용할 경우, 생성된 최대 여백 분류 경계는 상대적으로 데이터 개수가 적은 양성 데이터들이 모여 있는 쪽으로 편향되어 위치하게 된다. 이로 인해 양성 데이터가 테스트 데이터로 적용되었을 때 음성으로 잘못 판별할 가능성이 높아진다.

본 논문에서는 SVM의 정확도를 높이기 위해 주어진 임상 데이터를 양성 데이터와 음성 데이터를 일대일 매칭(matching)시킨 후 각 SNP의 중요도를 계산하고 서열화하는 전처리 과정을 수행한 후 중요도가 일정 범위 내에 드는 SNP를 대상으로 SVM을 이용한 학습을 실시하도록 하였다.

## 1. SVM 학습을 위한 임상 데이터 매칭

주어진 임상 데이터로부터 양성 데이터와 음성 데이터를 대상으로 같은 성별 및 나이의 차이를 최소화 할 수 있도록 일대일 매칭을 수행한 후 쌍을 이룬 데이터를 대상으로 학습을 수행하였다.

본 논문에서는 환자 리스트를 $P$, 정상인 리스트를 $N$, 환자와 쌍을 이룬 정상인의 리스트를 $M$, 환자의 수를 $n$이라 하고, $P_i$, $N_i$, $M_i$를 각각 리스트 $P$, $N$, $M$의 $i$번째 원소라 한다. 또, $L_{i,age}$, $L_{i,sex}$, $L_{i,snp}$을 각각 임의의 리스트 $L$에 대한 해당 원소의 나이, 성별, SNP리스트라 하고, $L_{i,snp,g}$를 해당 리스트의 $g$번째 SNP라 정의한다.

이때, 성별이 일치하며 나이차를

$$minimize \sum_{i=1}^{n} |P_{i,age} - M_{i,age}|, (P_{i,sex} = M_{i,sex}) \quad (1)$$

와 같이 최소화하기 위한 알고리즘을 [그림 2]에 나타내었다.

매칭 알고리즘의 앞부분에서는 greedy방법에 의하여 각 양성 데이터가 순서에 따라 자신과 가장 나이차가 작은 데이터를 음성 데이터 가운데서 찾아 자신의 짝으로 삼도록 한다. 알고리즘의 후반부 최적화 단계에서는 모든 두 양성 데이터에 대하여 서로의 짝을 교환했을 때, 전체적인 나이차가 줄어든다면 서로의 짝을 교환하게 되는데, 모든 두 쌍에 대해 이러한 시도를 수행한 후 변화가 일어난다면 이를 다시 반복하는 방법으로 전체 나이차의 합을 최소화한다.

**for** $i=1$ **to** $n$ **do**
  $M_i = N_j$ , where $P_{i,sex} = N_{j,sex}$ **and**
    minimize $|P_{i,age} - N_{j,age}|$
  delete $N_j$ from $N$
**end for**
**do**
 **for** $\forall 0 < (i,j) \le n$ **and** $i < j$ **do**
  **if** $P_{i,sex} = M_{j,sex}$ **and**
    $|P_{i,age} - M_{i,age}| + |P_{j,age} - M_{j,age}| >$
    $|P_{i,age} - M_{j,age}| + |P_{j,age} - M_{i,age}|$ **then**
   swap $M_i$, $M_j$
  **end if**
 **end for**
**while** some change occured

**그림 2. 일대일 매칭 알고리즘**

**표 1. 매칭 알고리즘에 의한 최적화의 예**

| Phase | 1 | | 2 | | 3 | |
|---|---|---|---|---|---|---|
| | $P$ | $N$ | $P$ | $M$ | $P$ | $M$ |
| i=1 | 10 | 5 | 10 | 12 | 10 | 5 |
| j=2 | 11 | 12 | 11 | 5 | 11 | 12 |
| 합 | | | 8 | | 6 | |

간단한 예를 들어, [표 1]과 같이 초기 단계에서 음성 데이터의 나이 $P_{1,age}, P_{2,age}$가 각각 10과 11로 구성되고 양성데이터 $N_{1,age}, N_{2,age}$이 각각 5와 12로 구성될

경우, 알고리즘의 초반부는 greedy 방법에 의해 표의 2단계와 같이 $P_1$은 $N_2$를 선택하고 $P_2$는 남은 $N_1$을 선택하게되어 각 쌍의 나이차의 합은 8이 된다. 다음으로 최적화를 위한 3 단계에서는 각 쌍의 차의 합 $|P_{1,age}-M_{1,age}|+|P_{2,age}-M_{2,age}|=8$보다, 서로의 쌍을 교환 할 경우 각 쌍의 차의 합이 $|P_{1,age}-M_{2,age}|+|P_{2,age}-M_{1,age}|=6$으로 더 작게 되어 $M_1$과 $M_2$를 교환하게 되며 최종적으로 각 쌍의 차의 합은 6으로 최적화 된다.

## 2. SNP의 중요도 계산

본 논문에 사용된 50개 SNP의 유전자 타입 3가지에 대해 양성과 음성 데이터의 차이비가 클수록 그 유전자가 예측률에 미치는 영향이 큰 것으로 판단하고 3가지 타입에 대한 차이비의 합을 이들 개별 SNP의 중요도로 설정할 수 있다. 조사된 SNP 중 $g\,(1 \le g \le 50)$ 번째 SNP를 $SNP_g$라하고 유전자 타입의 검사 값이 $v$ $(v=1,2,3)$인 환자의 수를 $C_{g,v}$라 하며 이들과 쌍을 이룬 정상인의 수를 $D_{g,v}$라 할 때, $SNP_g$의 중요도 $W_g$를 다음과 같은 식(weighting expression)으로 정의한다.

$$W_g = \sum_{v=1}^{3} \frac{|C_{g,v} - D_{g,v}|^e}{C_{g,v} + D_{g,v}}, \quad e \ge 1 \qquad (2)$$

이때, 지수 $e$는 각 SNP 타입별 양성과 음성의 차이가 두드러진 부분에 더 많은 가중치를 두고자 함이다. 예를 들어, $e$가 2일 때, 세 타입에 대한 차가 각각 1, -1, 1인 SNP 보다 3, 0, 0 인 SNP가 9배로 큰 가중치를 갖게 된다. 또, SNP의 유전자 타입에 해당하는 인원의 수에 대한 로가리즘의 곱 $\log(C_{g,v}+D_{g,v})$을 적용하여 보다 자연스러운 가중치를 사용할 수도 있다. 이는 각 SNP의 유전자 타입별 인원수에 대한 로그값을 곱함으로써 많은 인원수를 가진 항에 대해 더 높은 비중을 두어 예측에 대한 객관성을 높이는 역할을 한다.

[그림 3]은 각 SNP의 중요도를 결정하기 위한 알고리즘을 나타낸 것이다. 알고리즘의 전반부에서는 주어진 임상 데이터로부터 각 SNP와 SNP의 타입별로 해당되는 환자와 정상인의 수를 구한다. 그리고, 후반부에서는 제안된 식(2)를 이용하여 각 SNP의 중요도 리스트 $W$를 구하게된다.

```
initialize C,D
for i = 1 to n do
 for g = 1 to 50 do
  u = P_{i,snp,g}
  v = M_{i,snp,g}
  increment C_{g,u}
  increment D_{g,v}
 end for
end for
for g = 1 to 50 do
  W_g = weighting expression
end for
```

**그림 3. SNP 중요도 계산을 위한 가중치 할당 알고리즘**

## 3. 학습에 사용된 SNP 개수에 따른 예측률

본 논문에서는 실험 결과의 성능을 판단하기 위해 [표 2]에 정리된 TP(True Positive), FN(False Negative), FP(False Positive), TN(True Negative)의 값과 식(3)을 이용하여 민감도(sensitivity), 특이도(specificity), 정확도(accuracy)를 구하였다.

**표 2. 민감도, 특이도, 정확도를 위한 평가 항목**

| | | 발병사실 | |
|---|---|---|---|
| | | + | − |
| 예측결과 | Positive | TP | FP |
| | Negative | FN | TN |

$$\begin{aligned} &\text{민감도} = TP/(TP+FN) \\ &\text{특이도} = TN/(TN+FP) \\ &\text{정확도} = (TP+TN)/(TP+TN+FP+FN) \end{aligned} \qquad (3)$$

TP는 실제 환자를 환자로 식별한 횟수, FN은 환자를 정상인으로 식별한 횟수, TN은 정상인을 정상인으로 식별한 횟수, FP는 정상인을 환자로 식별한 횟수를 의미 한다. 민감도는 환자중 환자로 올바르게 식별된 비율, 특이도는 정상인중 정상인으로 올바르게 식별된 비율, 그리고 정확도는 전체적으로 올바르게 식별한 비율을 말하며 특이도와 민감도의 비율이 높으면서 그 차이가 작을수록 이상적인 예측이라 판단된다. 예를 들면 환자 100명, 정상인 100명이 있다고 가정할 때, 실험 결과 TP=80, FN=20, FP=25, TN=75의 값을 얻었다면 식(3)에 의해 민감도는 80%, 특이도는 75%, 정확도는 77.5%의 값을 얻게 된다.

폐암 감수성 예측률을 구하기 위해서 매칭을 통해 얻어진 각 쌍에 대해 50개의 SNP 데이터를 실험에 사용할 수 있으나 일부 SNP는 발병 예측에 도움을 주지 못하는 것으로 판단되어 앞 절에서 정의한 각 SNP의 중요도에 따라 가중치가 높은 선두 m개의 SNP만을 사용하기로 하며, m의 값에 따라 예측률은 다른 결과를 보였다.

```
matching()
weighting()
for m = 7 to 50 do
 for t = 1 to n do
  learn SVM by P+M-(P_t+M_t)
        on {SNP_l | 1 ≤ l ≤ m}
  classify and test P_t, M_t
  accumulate TP, FP, FN, TN
 end for
 sensitivity = TP/(TP+FN)
 specificity = TN/(TN+FP)
 accuracy    = (TP+TN)/(TP+TN+FP+FN
 print sensitivity, specificity, accuracy
end for
```

그림 4. SNP 개수에 따른 예측률 평가 알고리즘

[그림 4]는 SNP개수에 따른 예측률 평가 알고리즘으로 먼저, [그림 2]와 [그림 3]에 의해 앞서 설명된 전처리과정인 양성 데이터와 음성 데이터의 매칭과 각 SNP에 대한 가중치 계산을 수행한다. 이후 예측에 사용할 SNP의 각 수 $m$에 대해 모든 쌍의 데이터에 대해 예측의 정확성을 실험한다. 데이터 각 쌍은 자신이 제외된 다른 데이터에 의해 학습된 분류기에 의해 예측이 시험되고 그 결과는 발병사실과 예측결과에 따라 4개의 변수 $TP, FP, FN, TN$에 각각 누적된다. 7에서 50개까지의 SNP범위에 대해 모든 쌍의 데이터가 실험된 후 4개의 변수에 기록된 결과에 따라 식(3)에 의해 민감도, 특이도, 정확도가 각각 계산된다.

## IV. 실험 및 고찰

본 논문에서는 폐암 감수성 예측률을 구하는 실험을 위해 매칭 알고리즘을 이용하여 320명의 폐암 발생 환자에 대한 임상 데이터와 대조군으로 사용할 정상인 320명의 임상 데이터를 매칭한 후 사용하였으며, SVM 학습을 위해 선형커널을 사용하였다.

실험을 위해 펜티엄4 듀얼코어 CPU와 2Gb의 메모리를 내장한 40대의 PC를 네트워크로 연결하여 병렬 연산을 수행하였다. [그림 5]는 전송된 학습과 시험에 대한 각 PC의 수행 결과를 수신하여 통계 작업을 수행하는 프로그램의 실행 화면이다.

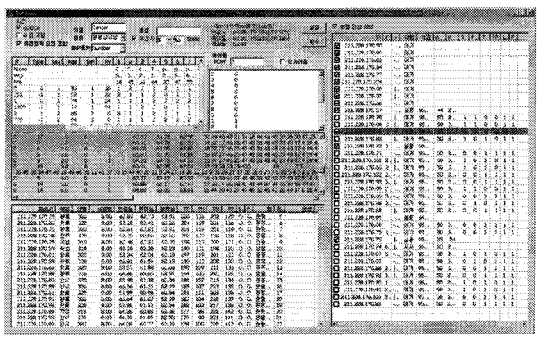


그림 5. 예측률 평가를 위한 병렬 연산 관리자

본 논문에서는 보다 객관적인 자료의 평가를 위해 LOOCV(Leave One-pair Out Cross Validation)[13] 검

사를 수행하였다. LOOCV평가는 전체 데이터에서 한 쌍(양성, 음성데이터)만을 제외한 나머지 데이터를 이용하여 학습한 후, 제외된 한 쌍의 데이터로 테스트하는 방법을 말하며, 이러한 과정을 전체 데이터 쌍의 개수만큼 반복하여 전체 결과에 대한 평균을 평가치로 삼는 것이다. 이러한 검사방법은 자신을 제외한 데이터를 바탕으로 예측모델을 구성하고 이 모델에서 자신을 검증받는 방법으로써, 검사받기를 원하는 대상이 검사모델의 구성에 참여할 수 없게 함으로써 예측모델의 객관성을 높이고자 하는 검사방법이다.

7~50개의 SNP 갯수에 대해 LOOCV 테스트를 수행을 위해 총 43*320번의 학습과 테스트가 이루어졌으며 PC 40대로 병렬 연산을 수행하는데 392초의 시간이 걸렸다.

연산 수행 결과 [표 3]과 [그림 6]과 같이 19개의 SNP에 의한 학습에 대해 64.22%로 가장 높은 정확도를 보였으며, 이때의 민감도와 특이도는 각각 64.92와 63.58로 나타났다.

**표 3. 서열화에 의한 각 SNP개수별 예측률**

| SNP수 | 민감도 | 특이도 | 정확도 | SNP수 | 민감도 | 특이도 | 정확도 |
|---|---|---|---|---|---|---|---|
| 7 | 63.32 | 60.97 | 62.03 | 29 | 60.40 | 59.35 | 59.84 |
| 8 | 62.28 | 60.11 | 61.09 | 30 | 58.86 | 57.77 | 58.28 |
| 9 | 59.86 | 58.38 | 59.06 | 31 | 59.33 | 58.24 | 58.75 |
| 10 | 62.58 | 62.42 | 62.50 | 32 | 58.88 | 58.04 | 58.44 |
| 11 | 62.08 | 62.62 | 62.34 | 33 | 57.74 | 57.27 | 57.50 |
| 12 | 63.01 | 62.93 | 62.97 | 34 | 58.28 | 57.40 | 57.81 |
| 13 | 63.01 | 62.93 | 62.97 | 35 | 59.03 | 58.48 | 58.75 |
| 14 | 63.07 | 61.98 | 62.50 | 36 | 58.10 | 57.85 | 57.97 |
| 15 | 62.05 | 60.83 | 61.41 | 37 | 58.60 | 58.28 | 58.44 |
| 16 | 64.77 | 62.87 | 63.75 | 38 | 60.13 | 59.28 | 59.69 |
| 17 | 63.42 | 61.70 | 62.50 | 39 | 58.71 | 58.18 | 58.44 |
| 18 | 64.33 | 62.65 | 63.44 | 40 | 59.34 | 58.51 | 58.91 |
| 19 | 64.92 | 63.58 | 64.22 | 41 | 58.17 | 57.49 | 57.81 |
| 20 | 63.73 | 62.57 | 63.13 | 42 | 58.14 | 57.23 | 57.66 |
| 21 | 64.80 | 63.39 | 64.06 | 43 | 58.42 | 57.02 | 57.66 |
| 22 | 62.75 | 61.11 | 61.88 | 44 | 57.43 | 56.40 | 56.88 |
| 23 | 63.09 | 61.40 | 62.19 | 45 | 60.00 | 58.55 | 59.22 |
| 24 | 63.96 | 62.95 | 63.44 | 46 | 58.28 | 57.40 | 57.81 |
| 25 | 63.25 | 61.83 | 62.50 | 47 | 58.72 | 57.60 | 58.13 |
| 26 | 60.98 | 60.00 | 60.47 | 48 | 60.42 | 58.52 | 59.38 |
| 27 | 61.59 | 60.36 | 60.94 | 49 | 59.57 | 57.54 | 58.44 |
| 28 | 60.33 | 59.40 | 59.84 | 50 | 60.28 | 58.10 | 59.06 |

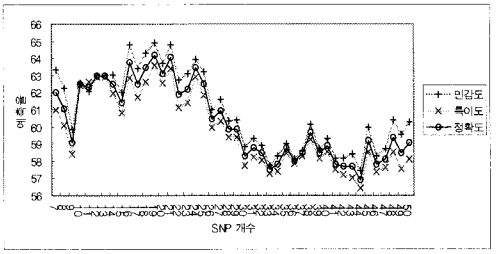



그림 6. 서열화에 의한 각 SNP개수별 예측률

**표 4. 로그화된 서열에 의한 예측률**

| SNP수 | 민감도 | 특이도 | 정확도 | SNP수 | 민감도 | 특이도 | 정확도 |
|---|---|---|---|---|---|---|---|
| 7 | 63.16 | 63.41 | 63.28 | 29 | 63.16 | 60.56 | 61.72 |
| 8 | 63.00 | 63.58 | 63.28 | 30 | 64.38 | 62.07 | 63.13 |
| 9 | 62.65 | 62.97 | 62.81 | 31 | 64.89 | 61.80 | 63.17 |
| 10 | 63.44 | 63.44 | 63.44 | 32 | 63.24 | 59.78 | 61.25 |
| 11 | 63.24 | 63.32 | 63.28 | 33 | 64.73 | 61.10 | 62.66 |
| 12 | 64.19 | 63.33 | 63.75 | 34 | 64.49 | 60.99 | 62.50 |
| 13 | 63.16 | 61.90 | 62.50 | 35 | 63.54 | 60.33 | 61.72 |
| 14 | 65.16 | 64.24 | 64.69 | 36 | 64.58 | 60.70 | 62.34 |
| 15 | 63.49 | 62.20 | 62.81 | 37 | 62.67 | 60.63 | 61.56 |
| 16 | 65.48 | 64.55 | 65.00 | 38 | 63.38 | 60.67 | 61.88 |
| 17 | 63.85 | 61.99 | 62.85 | 39 | 64.06 | 61.00 | 62.34 |
| 18 | 64.63 | 62.43 | 63.44 | 40 | 62.11 | 59.72 | 60.78 |
| 19 | 64.11 | 61.47 | 62.66 | 41 | 62.15 | 59.94 | 60.94 |
| 20 | 63.77 | 60.44 | 61.88 | 42 | 63.03 | 60.39 | 61.56 |
| 21 | 61.35 | 58.94 | 60.00 | 43 | 61.92 | 59.33 | 60.47 |
| 22 | 62.63 | 59.89 | 61.09 | 44 | 62.37 | 60.06 | 61.09 |
| 23 | 62.59 | 60.17 | 61.25 | 45 | 61.87 | 59.12 | 60.31 |
| 24 | 63.96 | 61.06 | 62.34 | 46 | 60.49 | 58.47 | 59.38 |
| 25 | 63.48 | 60.61 | 61.88 | 47 | 60.84 | 58.76 | 59.69 |
| 26 | 63.16 | 60.56 | 61.72 | 48 | 60.78 | 58.54 | 59.53 |
| 27 | 64.31 | 61.34 | 62.66 | 49 | 61.97 | 59.55 | 60.62 |
| 28 | 63.57 | 61.32 | 62.34 | 50 | 60.43 | 58.01 | 59.06 |

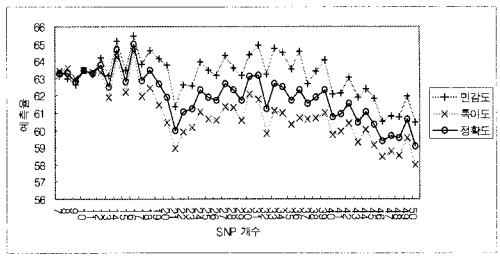



그림 7. 로그화된 서열에 의한 예측률

각 SNP의 유전자 타입별 인원수에 대한 로그값을 곱하여 SVM을 학습한 경우, 연산 수행 결과 [표 4]와 [그림 7]과 같이 16개의 SNP에 의한 학습에 대해 65.00%로 가장 높은 정확도를 보였으며, 이때의 민감도와 특이도는 각각 65.48와 64.55로 나타났다. 실험에서는 SNP 서열을 위한 가중치를 위해 수식 (2)($e = 10$)를 사용하였다.

7개의 SNP를 사용하여 만들어진 예측 기법[5]에서는 정확도 65.9%를 기록하였으나 이는 LOOCV방법을 적용하지 않은 실험 결과이며, 본 논문에서 제시된 예측 모델을 LOOCV방법을 적용하지 않고 평가할 경우, 다항 커널을 사용하여 예측률 75.0%까지의 결과를 얻을 수 있었으나 LOOCV를 사용하지 않은 상태에서의 예측률은 큰 의미가 없는 것으로 판단되어 여기서는 이에 대한 언급을 생략한다. 또, 간염 감수성 예측에 관한 연구[9]의 경우, 본 연구와는 달리 민감도와 특이도의 균형을 맞추기 위한 임상데이터의 전처리와 각 SNP 데이터의 중요도에 대한 고려가 포함되어 있지 않으며 실험 결과, 민감도와 특이도의 차이 값이 22로 나타났으나, 본 연구에서는 그 차이 값이 5이하로 나타나고 있다.

## V. 결 론

SVM 학습에 있어서 정상인의 수와 환자의 수를 일치시킬 경우 민감도 및 특이도가 좋은 균형을 이루는 결과를 보였고, 발병환자에 대해 가능한 의미있는 대조그룹을 형성하기 위하여 각 환자에 대해 동일한 성별을 가진 정상인 중 평균 나이차를 최소화 할 수 있도록 짝을 지었다. 또, 폐암 발생에 관여 하는 것으로 추정되는 50개의 SNP 중 어떤 SNP를 선택하는가에 따라 발병 예측률이 달라짐에 착안하여 통계적 방법에 의해 예측률에 미치는 각 SNP의 영향력에 대한 가중치를 계산하고 이를 서열로 정하여 이 가운데 발병 예측률을 가장 높일 수 있는 유전자의 개수를 실험적으로 구하였다. SNP 개수에 따라 예측률 평가를 수행하였고 높은 중요도를 갖는 16개의 SNP를 사용한 실험에서 가장 우수한 결과를 보였다.

본 논문을 통해 얻은 결과에 따라 폐암 예측 시스템을 구성할 경우, 폐암에 대한 조기 진단에 많은 도움이 될 것으로 기대한다.

다양한 서열화 연산과 다양한 SVM 옵션의 사용 그리고 시뮬레이티드 어닐링에 의한 정련 작업이 이루어진다면 보다 나은 결과를 얻어낼 수 있을 것이며 나이, 성별, 흡연력등의 정보를 추가하여 학습에 적용한다면 보다 정확한 예측이 가능해 질것으로 보인다.

## 참 고 문 헌


[1] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," Cancer Informatics, pp.59-78, 2006.

[2] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, J. Mackey, A. Wishart, R. Greiner, and B. Zanke, "Predictive models for Breast Cancer susceptibility for multiple single nucleotide polymorphism," Clinical Cancel Research, Vol.10, pp.2725-2737, 2004.

[3] J. I. Bell, "Single nucleotide polymorphisms and disease gene mapping," Arthritis Research, Vol.4, pp.S273-S278, 2002.

[4] Z. Wang and J. Moult, "SNPs, Protein Structure, and Disease," Human Mutation, Vol.17, pp.263-270, 2001.

[5] 박재용, *폐암 감수성 진단용 마커 및 이를 이용한 폐암 감수성 예측 및 판단방법*, 대한민국 특허출원 제 10-2006-0100277호, 2006.

[6] 박현석, 정철희, *자바로 배우는 바이오인포매틱스*, 사이텍미디어, 2006.

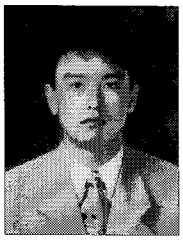
[7] T. G. Dietterich, "Machine Learning Research: Four Current Directions," The AI Magazine, Vol.18, No.4. pp.97-136, 1997.

[8] S. Mukkamala, G. Janowski, and A. H. Sung, "Intrusion Detection Using Support Vector Machines," Proceedings of High Performance

Computing Symposium-HPC, pp.178-183, 2002.
[9] 김동회, 엄상용, 함기백, 김진, "Single Nucleotide Polymorphism(SNP) 데이터와 Support Vector Machine(SVM)을 이용한 만성 간염 감수성 예측", 정보과학회논문지: 시스템 및 이론, 제37권, 제7호, 2007.
[10] 제홍모, 방승양, "앙상블 구성을 이용한 SVM 분류성능의 향상", 정보과학회논문지: 소프트웨어 및 응용, 제30권, 제3호, pp.251-258, 2003.
[11] 김한성, 권영희, 차성덕, "SVM 기반의 효율적인 신분위장기법 탐지", 정보보호학회논문지, 제13권, 제5호, pp.91-104, 2003.
[12] L. Wang, *Support Vector Machines: Theory and Applications,* Springer, 2005.
[13] http://www.slcmsr.net/boulesteix/papers/wilcoxon.pdf

## 저 자 소 개

**류 명 춘(Myung-Chun Ryoo)** 정회원



- 1989년 2월 : 영남대학교 컴퓨터학과(공학사)
- 1991년 2월 : 영남대학교 컴퓨터공학과(공학석사)
- 1995년 8월 : 영남대학교 컴퓨터공학과(박사과정수료)
- 1997년 3월 ~ 현재 : 경운대학교 컴퓨터공학과 교수

<관심분야> : 인공지능, 바이오인포매틱스

**김 상 진(Sang-Jin Kim)** 정회원

- 1994년 2월 : 계명대학교 컴퓨터공학과(공학사)
- 1996년 2월 : 경북대학교 컴퓨터공학과(공학석사)
- 2000년 8월 : 경북대학교 컴퓨터공학과(공학박사)
- 1999년 9월 ~ 현재 : 경운대학교 컴퓨터공학과 교수

<관심분야> : 알고리즘, 바이오인포매틱스

**박 창 현(Chang-Hyeon Park)** 정회원

- 1986년 2월 : 경북대학교 전자공학과 전산학전공(공학사)
- 1988년 2월 : 서울대학교 전산과학과 인공지능전공(이학석사)
- 1992년 8월 : 서울대학교 전산과학과 인공지능전공(이학박사)
- 1992년 9월 ~ 1993년 8월 : 서울대학교 컴퓨터신기술공동연구소 특별연구원
- 1998년 2월 ~ 1999년 2월 : Univ. of Maryland, UMIACS 연구소 방문연구원
- 1993년 9월 ~ 현재 : 영남대학교 전자정보공학부 컴퓨터공학전공 교수

<관심분야> : 인공지능(지식기반시스템, 기계학습)