SVM을 이용한 효율적인 위암관련 SNP 정보분석

김동회[0*], 김유섭[0], 천세학[1], 천세철[2], 함기백[3], 김진[0]
[0]한림대학교 정보통신공학부
[1]서울산업대학교 경영관리과
[2]건국대학교 생명환경과학대학 분자생명공학과
[3]아주대학교 간소화기 질환센터
e-mail:{kdh,yskim01}@hallym.ac.kr, shchun@snut.ac.kr,
scchun@konkuk.ac.kr, kibaek@ajou.ac.kr, jinkim@hallym.ac.kr

Effective Analysis Of SNP Related Gastric Cancer Using SNP

Dong-Hoi Kim[o], Yu-Seop Kim, Ki-Baek Ham[1], Jin Kim
Dept of Computer Engineering, Hallym University

요약

Single Nucleotide Polymorphism(SNP)는 인간 유전자 서열의 0.1%에 해당하는 부분으로 이는 각 개인의 체질 및 각종 유전질환과 밀접한 관련이 있다고 알려져 있으며 이 SNP 정보를 이용 각종 질환의 유전적 원인규명에 대한 많은 생물학적 연구가 진행되고 있다. 그러나 아직 SNP를 이용한 효율적인 분석방법에 대한 전산학적 연구는 많지 않다. 본 논문에서는 대표적인 패턴인식기 중 하나인 Support Vector Machine(SVM)을 이용 한국인의 대표적인 유전질환으로 알려진 위암에 대한 예측율을 실험하였다. 실험 데이터는 간 및 소화기 질환 유전체 센터에서 얻어진 위 질환 환자를 대상으로 하였으며 실험 결과 예측율은 67.3%로 이는 Case Based Reasoning(CBR)방법의 55% 보다 더 좋은 예측 결과를 보였다.

1. 서론

Single Nucleotide Polymorphism(SNP)는 인간의 유전체를 이루는 전체 30억 염기 중 0.1%에 해당하는 부분으로 집단 또는 개인 간 형질 차이를 결정하는 서열이다. 최근 이 SNP데이터를 이용 각종 질환 예측 및 맞춤의학에 관한 많은 연구가 진행되고 있다. 현재 국내에서는 SNP정보를 이용 한국인의 유전 질환과 SNP와의 관계를 규명하기 위한 많은 연구가 진행되고 있다. 본 논문에서는 Support Vector Machine(SVM)[1]을 이용 한국인의 대표적 유전질환인 위암의 예측율 실험에 대해 논한다. 2장에서는 SNP과 SVM에 대하여 설명하고 3장에서는 실험에 사용된 데이터 및 방법에 대하여 설명한다. 4장에서는 실험에 대한 평가를 하고, 마지막으로 5장에서 결론을 맺는다.

본 연구는 보건복지부 보건의료기반 진흥사업
(01-PJ10-PG6-01GN14-0007)의 지원에 의해 이루어진 것임

2. 관련연구

2-1 SNP

염기서열은 DNA의 경우 {A, T, G, C}의 알파벳으로 이루어진 유한한 길이의 스트링으로 정의할 수 있다. 예를 들면 스트링 "GCCTACCGAGGC"는 DNA의 서열(sequence)이라 할 수 있다. 인간의 서열의 개수는 30억 개이며, 개개인의 서열을 비교하였을 때 99.9%가 동일하다. 하지만 인류 집단 내에서 일부분 0.1%의 차이에 의해 개인 간에 모습이나, 행동 그리고 질환감수성에 차이가 생긴다. 즉 3백만 개 정도의 염기서열 부위에서 서로 다른 염기에 의해 개인 간의 차이 또는 일정 집단이나 인종, 민족 간에 차이가 발생하게 된다. 30억 개 인간 유전체

염기서열 중에서 대략 1.0kb마다 서로 다른 염기가 올 수 있다. 이는 총 3백만 개가 되며, 이를 단일염기 다형성SNP라고 부른다[2]. 즉 전체 염기서열을 분석하지 않아도 다형성을 보이는 이들 3백만 개 염기서열을 분석한다면 전체 염기서열을 분석하지 않아도, 개인 간이나 집단 간의 유전적 차이, 또는 질환군과 정상인의 차이를 알 수 있으며 따라서 질환의 조기 진단 및 개인별 맞춤의학 분야에 널리 사용될 수 있을 것이다.

2-2 SVM

2-2-1 Support Vector Machine 의 개요

SVM은 최소의 일반화 에러로 나타나게 하는 최적의 분류 평면(Separating Hyperplane)을 결정하는 기법이라고 볼 수 있다. 일반적으로 선형적으로 분류 가능한 문제의 분류식은

$$f_{w\,.\,b} = sign(w \cdot x + b)$$

와 같이 나타낼 수 있다.

SVM에서 최적의 분류 평면은 서로 다른 클래스들을 구분하는 최대 마진(margin) 사이에 존재한다고 본다. 입력벡터 x_i에 대한 클래스 레이블(label)이 y_i라고 할 때, 최적의 분류 평면은 다음의 제약조건 최소화를 만족해야 한다.

$$Min : \frac{1}{2} w^t w \;\; where \; y_i (w \cdot x_i + b) \geqq 1$$

선형적으로 구분이 불가능한 경우, 위의 최소화 조건은 오분류 데이터를 허용하기 위해 수정되어야 한다. 수정된 식에서 soft margin 분류기가 어느 정도의 에러를 허용하는 대신 제약조건의 위반의 측정치로 새로운 변수인 c를 포함한다. 그리고 a_i가 라그랑지(Largrangian) 계수일 때,

$$Min : L(W) = \frac{1}{2} \cdot < w.w > - \sum a_i y_i \left[(< w, \varphi(x_i) > + b) - 1 \right]$$

$$0 \leqq a_i \leqq C$$

$$\frac{\partial L}{\partial b} = 0$$

$$\frac{\partial L}{\partial w} = 0$$

이다. 여기서, C는 ζ_i의 가중치이며, $\phi(\bullet)$는 입력 공간을 보다 고차원의 공간으로 매핑하는 비선형함수이다. 이 때, 위 식의 첫 번째 항을 최소화하는 것은 VC 차원을 최소화 하는 것과 같은 효과이다. 위 식을 풀기 위해서 라그랑지 방법을 이용하여 다음과 같이 변형한다.

$$Max : W(a) = \sum a_i - \frac{1}{2} \cdot \sum a_i a_j y_i y_j K(x_i, x_j)$$

$$0 \leqq a_i \leqq C$$

$$\sum a_i y_i = 0$$

이 때 $K(x_i, x_j) = <\phi(x_i), \phi(x_j)>$ 인 커널 함수이다.

두 클래스(binary class) 분류에서의 임의의 입력 벡터 X에 대한 분류 함수는 다음 식과 같다.

$$f(X) = sign\left(\sum_{i=1}^{l} y_i a_i (x_i \cdot X) + b \right)$$

2-2-2 커널 함수

SVM에서는 고차원의 기본 함수들을 구축하는 것이 매우 중요하다. 이 때 구축되는 고차원의 공간을 자질 공간 (feature space) 이라고 하고 이 전의 공간을 입력 공간 (input space) 이라고 한다[3]. 만일 입력 공간을 그대로 자질 공간으로 사용하여 최적의 hyperplane을 찾고자 한다면, 이 경우의 커널 함수 $K(\bullet)$는 입력 벡터의 단순 내적이 되는데 이 경우의 커널 함수를 선형 커널이라고 한다. 선형 커널은 여러 커널 함수 중에서 가장 단순한 형태의 함수이다.

그러나 만일 큰 집합의 사상 함수가 사용이 된다면 최적화 문제는 기본 함수들로 정의된 자질 공간에서의 새로운 내적이 필요하다. 그리고 내적 커널 $K(\bullet)$는 기본 함수 $\phi_l(\boldsymbol{x})$, $l = 1,\ldots,m$, 를 가지고 다음과 같이 표현할 수 있다. 여기서 m은 무한할 수 있다.

$$K(x_i, x_j) = \sum_{l=1}^{m} \Phi_l(x_i) \Phi_l(x_j)$$

일반적으로 고차원 자질 공간에서의 두 자질 벡터 간의 내적을 계산하는 것은 입력 공간에서의 벡터들의 support vector들 간의 커널 K 의 계산을 통하여 간접적으로 이루어진다. 커널 함수의 일반적인 형식은 다음과 같다.

$$K(x_i, x_j) = (z_i, z_j)$$

여기서 벡터 $\mathbf{z}_i$와 $\mathbf{z}_j$는 입력 벡터들의 자질 공간에서의 사상이다. 그리고 커널 함수는 자신의 비선형적인 형태를 다음 식을 이용하여 자질 공간을 생성하기 위하여 확장시킨다.

$$K(x_i, x_j) = [(x_i \bullet x_j) + 1]^q$$

여기서 q는 다항의 정도를 말한다. 이러한 형식을 가지고 있는 커널 함수를 다항 커널 함수라고 한다.

지금까지 많은 연구자들이 다양한 형태의 커널 함수를 그들의 응용에 적용시키기 위하여 개발하여 왔다. 그러나 선형 커널과 다항 커널이 그 중에서 가장 광범위하고 다양한 응용에 사용된 커널이다. 따라서 본 연구에서는 선형 커널 및 다항 커널을 각각 적용시켜 문제를 해결하고자 하였다.

3. 데이터 및 방법

3.1 데이터

SVM을 이용한 예측율 실험에 사용된 데이터는 아주대학교 간 소화기 질환 유전체 센터에서 얻어진 소화기 질환 환자 중 SNP가 확보된 위암환자 85명과 일반환자 175명 총260명의 데이터를 이용하였으며, SNP데이터는 위암과 관련성을 가진다고 알려진 MMP2유전자의 SNP 3개를 사용하였다. MMP2 유전자는 Matric Metallo Protainase(MMPs)의 한 부류이다. 표1은 MMP2유전자의 SNP들과 위치 그리고 해당 SNP가 가질 수 있는 염기를 나타낸다.[4] 그림1은 MMP2 유전자 구조를 나타낸다.[5]

표 1 MMP2 Gene

No	SNP Name	Position	Value
1	MMP2-F602F	10243470	C,T
2	MMP2-D383D	54081206	C,T
3	MMP2-T460T	54084614	G,C

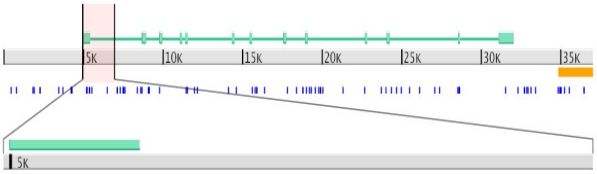

그림 1 MMP2 Structure

SVM을 이용한 실험을 위해 SNP값을 표2와 같이 A,C,G,T를 1,2,3,4 형태의 정수로 변환하였다. 이때 환자가 위암이면 0, 정상이면 1로 결과를 변환하였다.

표 2 실험을 위한 변환된 데이터

	실제 값			변환 값			
환자레코드	SNP1	SNP2	SNP3	SNP1	SNP2	SNP3	위암여부
환자 1	CT	CC	GG	24	22	33	0
환자 2	CC	CT	GA	22	24	31	1
환자 3	CC	TT	AA	22	44	11	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
환자 n	CT	CC	GG	24	22	33	정상

실험에 사용한 입력데이터는 표3과 같다. 표3에서의 Class는 암의 유무를 나타내며 각 factor는 factor번호 : factor값으로 표현한다.

표 3 SVM 입력데이터 (Class 1:normal 2:cancer)

Class	Factor1	Factor2	Factor3
1	1:22	2:24	3:31
2	1:24	2:22	3:33
2	1:24	2:22	3:33
2	1:24	2:24	3:33
1	1:22	2:24	3:31
. . .	. . .	. . .	. . .
1	1:22	2:22	3:31

3.2 실험방법

본 논문에서는 SVM Multiclass[6]를 사용하였으며 학습데이터와 테스트데이터는 전체 환자데이터를 이용하였다. 전체 260개의 환자정보 가운데 첫 번째 환자정보를 위암여부를 판정하기 위한 테스트 데이터로써 사용하며, 나머지 256개를 학습데이터로 하였다. 이러한 방식을 모든 환자정보에 라운드 로빈 방식으로 적용하여, 총 260개의 테스트 데이터를 만들어 이들을 테스트를 함으로서 어느정도의 정확도를 가지고 질환과 정상을 구별할 수 있는가를 산출하였다. 이러한 방식을 적용한 이유는 환자의 SNP 정보를 얻는 것은 매우 어렵기 때문에 기존 정보를 최대한 활용해야 하기 때문이다.

이 실험의 전체 흐름은 그림2와 같다.

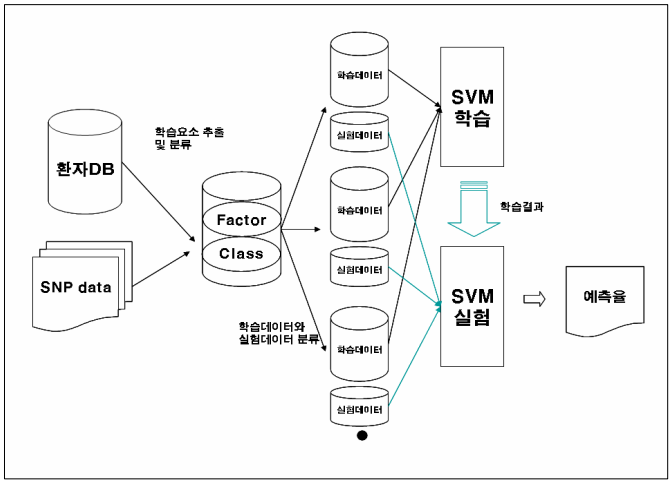

그림 2 전체 실험 구조

우선 환자데이터베이스의 검진결과와 해당 환자의 SNP정보를 이용 3.1과 같이 데이터를 변환하고 변환된 데이터 가운데 테스트데이터로 사용할 1개를 제외한 나머지 데이터를 학습데이터로 하여 총 260번의 학습과 테스트를 하였다. 그림3은 실험결과의 일부분이다.

```
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
```

그림 3 실험 결과 출력

4. 평가

실험 결과 SVM을 이용 선형 커널 및 다항 커널을 각 각 적용한 결과 선형 커널을 이용 했을 때 67.3%, 다항 커널을 사용 했을 때 49.2%로 나타났으며 선형커널을 사용했을 때 동일한 데이터를 사용하여 최근 연구된 Case Based Reasoning(CBR) 기법[7]을 이용한 방법(55%) 보다 더 높은 예측율을 보였다. 표 4는 CBR 기법을 사용하였을 때와 SVM을 사용하였을 때의 결과이다.

표 4 CBR과 SVM의 예측율비교

CBR	SVM	
	Linear	Polynomial
평균 55.00%	67.31	49.23%

위의 표는 본 데이터를 예측하기 위해서는 CBR이나 다항 커널 SVM보다 선형 커널 SVM이 더 적합함을 알 수 있다. 본 논문에서는 세 개의 SNP정보만을 factor로 사용하였으나, 보다 효율적인 예측을 위해서 더 많은 SNP정보나 기타 환자 관련 정보를 추가해야 할 것으로 생각된다.

5. 결론

본 논문에서는 한국인의 대표적인 유전질환으로 알려진 위암에 대한 진단예측을 위해 대표적인 패턴인식기 중 하나인 SVM을 적용하였다. SNP 정보를 이용 각종 질환의 유전적 원인규명에 대한 많은 생물학적 연구가 진행되고 있으나 아직 SNP를 이용한 효율적인 분석방법에 대한 전산학적 연구는 많지 않다. 따라서 본 연구는 다양한 유전자 분석에 적용할 수 있다. 더 많은 위암 관련 SNP와 충분한 학습데이터를 사용한다면 더 효율적인 예측성능을 가져올 수 있을 것이다. 향후연구로는 더 높은 예측율을 위하여 더 많은 SNP정보와 환자임상정보, 환자의 생활 패턴 등의 학습요소를 추가할 예정이다. 또한 위암이외에 간질환 같은 질병에 대해서도 본 논문에서의 방법을 적용할 예정이다.

참고문헌

[1] Vapnik, V. N., "The Nature of Statistical Learning Theory," *Springer*, 1995.

[2] Anthony J. Brookes "The essence of SNPs" GENE 1999.

[3] Cherkassky, V., and F. Mulier, "Learning from Data - Concepts, Theory, and Methods," John Wiley & Sons, Inc., 1998.

[4] http://www.ncbi.nlm.nih.gov/SNP

[5] http://www.mutationdiscovery.org

[6] http://svmlight.joachims.org/svm_multiclass.htm

[7] Se-Hak Chun, JinKim , Yoon-joo Park, Ki=Baek Ham, Se-Chul Chun "Data Mining Techniques for Medical Informatics:Application to SNP Analysis" 한국지능정보시스템학회 2005