SVM 앙상블을 이용한 심혈관질환 질환단계 예측

엄재홍[0] 장병탁

바이오지능연구실
서 울 대 학 교 컴 퓨 터 공 학 부
{jheom, btzhang}@bi.snu.ac.kr

Prediction of Cardiovascular Disease Steps using Support Vector Machine Ensemble

Jae-Hong Eom Byoung-Tak Zhang

Biointelligence Laboratory
School of Computer Science and Engineering, Seoul National University

요 약

현재 심혈관 질환은 암 다음으로 높은 사망 원인으로 기록되고 있어 심혈관 질환에 대한 초기 진단은 질환의 치료에 매우 중요한 문제로 대두되고 있다. 본 논문에서는 SVM을 이용하여 심혈관질환 환자의 질환 단계를 예측하였다. 일반적으로 이진분류에 사용되는 SVM을 이용하여 정상 및 질환 1~3기의 총 4가지 분류가 필요한 다분류 분류문제를 처리하기 위해서 논문에서는 독립적 학습된 단일 SVM 분류기들을 결합하여 분류를 수행하는 SVM 앙상블 방법을 사용하였다. 단일 분류기의 결합은 Majority voting, 최소자승에러기반 가중치 부여, 2단계층 결합 등의 방법으로 수행하여 심혈관 질환 분류에 적합한 앙상블의 구성을 시도하였다. 실험 데이터는 ㈜제노프라의 압타머 칩 데이터를 사용하였다. 서로 다른 데이터를 이용하여 학습된 이종의 SVM들을 결합한 결과 질환단계 예측에 있어서 단일 SVM을 이용하여 질환 단계를 예측하는 경우 보다 향상된 질환단계 예측 성능을 관찰할 수 있었으며, 심혈관 질환의 예측에 대해서는 단일 SVM 분류기의 2단 계층 결합법이 가장 좋은 성능을 보임을 확인하였다.

1. 서론

1.1 연구의 배경

심혈관질환은 심부전, 고혈압성 심장질환, 부정맥, 판막질환, 선천성 심장질환, 심근증, 심낭질환과 같은 심장질환과 뇌졸증, 말초혈관질환, 동맥류 등의 혈관질환을 포함하는 질병이다. 심장질환들 중에서 중요한 부분을 차지하는 관상동맥 질환은 대개 동맥경화에 의해 심장에 혈액을 공급하는 관상동맥이 막히거나, 좁아져 발생하는 것으로 심근경색증이나 협심증이 이에 해당한다. 관상동맥질환은 우리나라에서 지난 30여 년 동안 급격히 증가하여왔는데 급속한 경제 발전에 따른 식이의 변화와 주로 관련이 있을 것으로 생각되고 있다. 현재 심혈관질환은 암 다음으로 높은 사망 원인으로 기록되고 있어 질환에 대한 초기 진단은 질환 치료에 매우 중요한 문제로 대두되고 있다.

현재 심혈관 질환의 진단은 심전도 검사, 초음파 검사, 혈액검사, 혈관 조영술(angiography) 등의 방법으로 이루어지고 있으며 이 같은 방법들은 심혈관 질환의 진단 및 관찰을 위해 유용하게 사용되고 있지만, 여러 가지 많은 검사 결과를 종합해야 최종적인 진단을 내릴 수 있다. 때문에 심혈관 질환 검사를 받기 위해서는 검진만을 위해서도 많은 시간과 비용을 지불해야 하는 문제가 있다. 더욱이 현재까지 심혈관 질환 진단에 가장 유용한 방법으로 알려진 혈관 조영술의 경우, 시술 자체에 위험성을 내포하고 있고, 그 비용도 많이 든다. 따라서 보다 용이하게, 그리고 적은 비용으로도 우수한 성능으로 심혈관 질환에 대한 위험도를 예측 할 수 있는 방법에 대한 필요성이 대두되어왔다. 또한 최근 들어 질병 진단 방법에 있어서도, 기존의 임상 진단 방법에만 의존하던 것과는 달리 CDSS(Clinical Decision Support System)와 같은 전문 진단환경 시스템을 이용하는 등의 다양한 방향으로 변화하고 있다. 이에 본 논문에서는 이러한 진단환경 시스템에서 활용 가능한 기계학습 기법들 중에서 현재 다양한 분야의 응용에 대해서 성공적으로 사용되고 있는 SVM(Support Vector Machine)을 이용하여 심혈관 질환의 질환 단계를 예측하였다.

1.2 관련 연구

전체 질환에서 심혈관 질환이 차지하는 비중이 상대적으로 높기 때문에, 초기 진단의 중요성이 대두되어 심혈관질환 진단을 위한 다양한 연구들이 진행되어왔다. 심혈관 질환 진단에 대해서는 일반적으로 각 나라에서 나라별 인구 통계 특징을 반영한 진단 표를 작성하여 실제 의학 진단에 활용하고 있는데[1], 그림 1은 뉴질랜드의 심혈관 질환 판정표이다[2].

Wilson 등은 90년대 후반 영국에서 수행된 건강조사 자료를 기초로 심장질환의 중요 요인이 되는 콜레스테롤 측정을 통한 질환 예측 연구를 수행하였다[3]. 또한, Quaglini 등은 각 나라별로 구성되어있는 심혈관 질환 계산표에 대한 분석과 함께 새로운 계산표의 구성에서 고려해야 할 사항들을 제시하였다[4]. 이 외에도 국가별 인구조사를 통해 측정된 데이터를 기반으로 다양한 연구가 수행되고 있다.

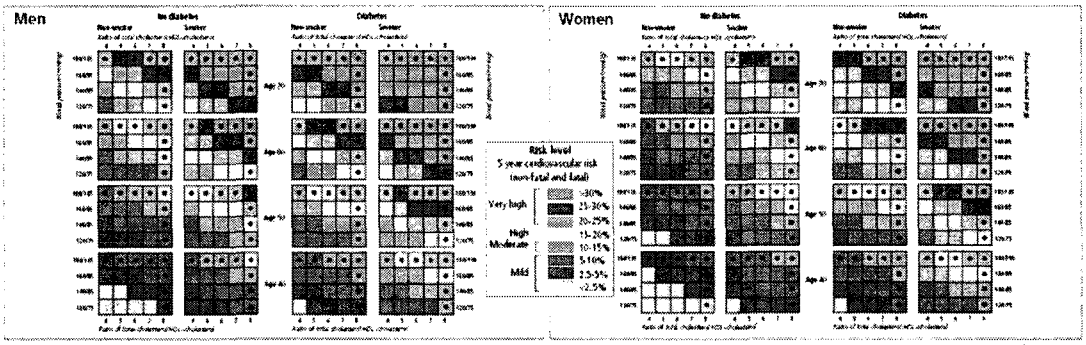

그림 1. 주요 위험요인을 고려한 뉴질랜드의 임상 진단용 심혈관질환 위험도 계산표의 예[1]. 일반적으로 심혈관 질환의 위험도는 나이 및 성별, 흡연 여부, 혈중 콜레스테롤 및 당뇨질환 여부에 따라 그 분포가 상당히 다르게 나타난다.

2. SVM

SVM은 Vapnik과 그의 연구그룹이 발표한 이진분류를 수행하는 기계학습방법이다. 모든 학습데이터는 특징으로 표현되는 벡터공간에 표현된다[5]. SVM에서의 학습은 벡터공간 내에서 분류경계(hyperplane)와 가장 가까운 거리에 있는 학습데이터(지지벡터)와의 최소거리를 최대화시키는 것, 즉 즉 최대여백 분류경계(maximal margin hyperplane)를 찾는 것을 목표로 한다. SVM에서는 지지벡터만이 분류경계를 구성하는데 사용된다. 분류경계는 커널함수(kernel function)로 표현되며, 커널함수는 벡터공간에서의 벡터간의 스칼라 곱(dot product)으로 표현된다. SVM에서 가장 널리 사용되는 커널함수로는 선형커널(linear kernel), 다항커널(polynomial kernel), RBF커널(RBF kernel) 등이 있다. 이 중 선형커널은 가장 빠르고 간단한 커널함수로서 많은 SVM기반 응용에서 강력한 성능을 나타내었다. 본 논문에서는 Joachims의 SVM light에 구현된 선형커널함수를 이용하여 SVM을 학습하고 환자의 분류를 수행하였다.

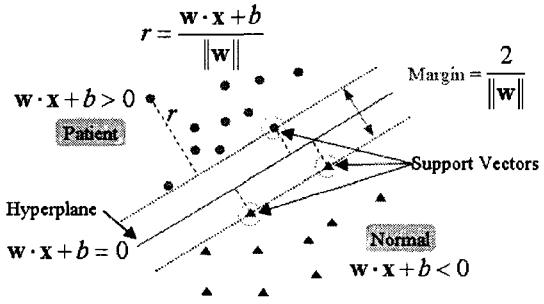

그림 2. SVM을 이용한 데이터 이진분류와 지지벡터 표현

$$h(\mathbf{x}) = sign\left[w\cdot x + b\right],\ w = \sum \alpha_i y_i x_i,\ \ \alpha_i \geq 0$$

주어진 샘플 s_i의 특징 벡터를 $x_i = \langle f_1(s_i),\ldots,f_n(s_i)\rangle$로 정의하면 선형커널함수에 기반을 둔 샘플의 이진 분류는 위의 식과 같이 표현된다. 일반적으로 SVM의 학습데이터 S는 $S = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$과 같이 표현되는데, 여기에서 $y_i \in Y = \{+1,-1\}$는 분류기의 출력공간(샘플의 유형: +1은 질환 환자 샘플, -1은 정상 샘플)을 나타낸다. 이때, $\mathbf{x}_i$의 가중치 α_i가 $\alpha_i \neq 0$ 인 경우 $\mathbf{x}_i$를 지지벡터(support vector)라 정의한다. 이와 같은 SVM을 이용한 샘플의 이진 분류의 개념을 도식화하면 그림 2와 같다.

3. SVM 앙상블

3.1 SVM 앙상블의 구성

SVM은 기본적으로 이진분류를 위한 모델이라는 점과 SVM의 학습이 대용량 데이터에 대해서는 상당히 계산 집약적 모델이기 때문에 실제 구현에서는 SMO와 같은 근사법을 사용하여 분류 성능이 저하될 수 있다는 문제를 갖는다. 논문에서는 일반적으로 개별 분류기의 앙상블 구성이 전체 분류 성능을 향상시킬 수 있다는 점에 착안하여 SVM 앙상블을 구성하여 총 4단계의 질환 단계의 질환 단계를 예측하였다.

분류기의 앙상블을 구성함에 있어 전체 성능의 향상을 위해 각각의 분류기를 가능한 한 다르게 구성하도록 하였다. 이를 위해 각 SVM 분류기는 Bagging과 Boosting 방법을 통해 각각 다른 데이터집합으로 학습하였다.

3.2 개별 SVM 분류기 결과의 선형 취합

Majority voting — 개별 SVM 분류기의 결과를 취합하는 가장 간단한 방법이다. $f_k(k=1,\ldots,K)$를 SVM 앙상블의 k번째 SVM의 결정함수라 하고 $C_j(j=1,\ldots,C)$가 j번째 클래스 레이블이라 할 때, j전째 클래스로 판별하는 SVM 분류기의 개수를 $N_j = \#\{k|f_k(\mathbf{x}) = C_j\}$와 같이 정한다. 이 경우 주어진 데이터 벡터 $\mathbf{x}$에 대한 SVM 앙상블의 majority voting 분류 결과 $f_{mv}(\mathbf{x})$는 다음과 같이 정해진다.

$$f_{mv}(\mathbf{x}) = \arg\max_j N_j.$$

LSE weighting — 이 방법에서는 개별 SVM 분류기에 독립적인 가중치를 부여한다. 각 분류기의 가중치는 훈련 데이터에 대한 예측 정확도로 부여하였다. k번째 분류기를 위한 훈련데

이터 $\tau_k = \{(\mathbf{x}_i';y_i')|i=1,...,L\}$로 훈련된 분류기$f_k$라 하면, 가중치 벡터 $\mathbf{w}$는 $\mathbf{A}=(f_i(\mathbf{x}_j))_{K\times L}$과 $\mathbf{y}=(y_i)_{1\times L}$에 대해 $\mathbf{w}_E=\mathbf{A}^{-1}\mathbf{y}$로 구할 수 있다. 이 경우 가중치를 이용한 주어진 데이터벡터 $\mathbf{x}$에 대한 SVM 앙상블의 분류 결과 $f_{mv}(\mathbf{x})$는 다음과 같이 정해진다.

$$f_{LSE}(\mathbf{x}) = sign(\mathbf{w}_E \bullet [(f_i(\mathbf{x}))_{K\times 1}]).$$

Double-layer Hierarchical Combination — 앞의 방법과 달리 비선형 결과 취합 방법으로 하단의 SVM 분류기의 결과를 상단의 SVM 분류기의 입력으로 사용하여 최종 분류를 수행하는 방법이다. $f_k(k=1,...,K)$를 SVM 앙상블의 k번째 SVM의 결정함수라 하고 상단의 SVM의 결정 함수를 F 라 하면 SVM 앙상블의 최종 결정함수 $f_{SVM}(\mathbf{x})$는 다음과 같이 정해진다.

$$f_{SVM}(\mathbf{x}) = F((f_1(\mathbf{x}),...,f_K(\mathbf{x}))).$$

4. 실험 및 결과분석

4.1 데이터 및 전처리

논문에서는 ㈜제노프라의 암타머 칩 데이터를 사용하였다. 아래의 표 1은 논문에서 심혈관 질환을 분류에 사용된 전체 데이터에 대한 정보를 나타낸다.

표 1. 실험에 사용한 데이터의 각 질환 단계별 샘플 정보

데이터베이스	데이터 개수	전체 단백질 수 (사용된 수)
정상	20	3000 (400)
질환 1기	20	3000 (400)
질환 2기	20	3000 (400)
질환 3기	20	3000 (400)
계	80	3000/샘플

샘플별로는 칩 스캔 데이터 중에서 median 값을 칩 데이터로 사용하였고 분산분석(ANOVA)을 이용하여 각 단백질 별로 p-value 측정 한 후 샘플별로 전체 3000개의 단백질 들 중에서 p-value가 낮은 순으로 상위 400개의 단백질들만을 선택하여 분석 데이터로 사용하였다. 또한, 정상~질환3기의 다분류 문제를 처리하기 위하여 정상vs.질환(1~3기) 등과 같이 2진 분류를 하는 SVM을 조합하여 사용하였다.

4.2 실험 결과

표 2는 단일 SVM 분류기와 앙상블 SVM 분류기를 이용한 심혈관질환 질환기의 분류 결과는 나타낸다.

표 2. SVM 앙상블을 이용한 심혈관질환 질환기의 예측 결과

분류기	분류 성능	
	Bagging	Boosting
SVM 분류기 (single classifier)	82.41 %	
SVM 앙상블 (majority voting)	89.54 %	89.31 %
SVM 앙상블 (LSE weighting)	91.53 %	92.32 %
SVM 앙상블(hierarchical combination)	93.12 %	93.74 %

표 2에서와 같이 심혈관질환의 질환단계 예측 문제에 있어서 SVM 분류기를 2-layer로 구성하여 분류 결과를 비선형적으로 취합한 분류기 모델이 다른 앙상블 모델들의 성능보다 상대적으로 좋은 성능을 보였다. 그렇지만 현재의 분류 성능 결과는 총 데이터가 80개뿐인 상태에서 분류를 수행한 결과이기 때문에 이 결과가 충분한 규모의 데이터에 대한 일반화된 분류성능을 보여준다고 할 수 없는 상황이고 이에 대한 적절한 해결방안이 필요하다.

5. 결론 및 향후 과제

논문에서는 SVM 앙상블 방법을 이용하여 심혈관 질환 환자의 압타머칩 데이터를 질환단계 별로 분류하였다. 논문에서 사용한 방법은 어느 정도 유의미한 분류 성능을 보여줬다. 그렇지만, 실험에 사용된 데이터는 심혈관 질환의 실제 임상진단에서 중요 요소로 고려되는 환자의 나이, 흡연여부, 혈중 콜레스테롤 수치 등과 같은 칩데이터 외적 정보를 포함하지 못하였다. 이처럼 임상진단에서 사용되는 중요 지표들을 함께 고려한다면 보다 정확한 질환의 예측이 가능할 것으로 기대되며, 나아가 나이와 성별에 따른 특이 단백질 집합을 발견할 수도 있을 것으로 전망된다. 또한 분류모델의 일반화 성능 향상을 위해서 실험에서 사용한 소규모의 데이터 보다는 좀 더 충분한 수의 샘플 확보도 필요하다.

감사의 글

본 연구는 과학기술부 국가지정연구실(NRL) 사업에 의하여 일부 지원되었으며, 실험에 사용된 심혈관질환 압타머 칩 데이터는 '㈜제노프라' (www.genoprot.com)에서 제공된 데이터로 해당 데이터의 소유권은 '㈜제노프라'에 있음을 밝힙니다.

참고문헌

[1] Hense, H.W., Schulte, H., Lowel, H., Assmann, G., Keil, U., "Framingham risk function overestimates risk of coronary heart disease in men and women from Germany — results from the MONICA Augsburg and the PROCAM cohorts," *Eur. Heart J.*, Vol. 24, pp. 937-945, 2003.

[2] Jackson, R., "Updated New Zealand cardiovascular disease risk-benefit prediction guide," *Brit. Med. J.*, Vol. 320, pp. 709-710, 2000.

[3] Wilson, S., Johnston, A., Robson, J., Poulter, N., Collier, D., Feder, G. et al., "Comparison of methods to identify individuals at increased risk of coronary disease from the general population," *Brit. Med. J.*, Vol. 326, pp. 1436-1438, 2003.

[4] Quaglini, S., Stefanelli, M., Boiocchi, L., Campari, F., Cavallini, A., Micieli, G., "Cardiovascular risk calculators: understanding differences and realising economic implications," *Int. J. Med. Inform.*, Vol. 74, pp. 191-199, 2005.

[5] Vapnik. V., "Statistical Learning Theory," John Wiley & Sons Press, 1998. ■