Journal of The Korea Society of Computer and Information
Vol. 17, No. 10, October 2012
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci/2012.17.10.145

ST 분절 급상승 심근경색 환자들의 단기 재발 사망 예측

임 광 현*, 류 광 선**, 박 수 호**, 손 호 선**, 류 근 호***

Short-term Mortality Prediction of Recurrence Patients with ST-segment Elevation Myocardial Infarction

Kwang-Hyeon Lim*, Kwang-Sun Ryu**, Soo-Ho Park**,
Ho-Sun Shon**, Keun-Ho Ryu***

요 약

현대 사회는 서구화된 식생활 패턴과 흡연, 비만 등의 원인으로 인해 심혈관계 질환들이 급증하고 있다. 특히, 급성심근경색은 심혈관계 질환으로 인한 사망의 대부분을 차지하고 있다. 이러한 추세에 따라 해외 선진국에서는 임상생리학적 오류를 줄이기 위해서 자국민의 데이터를 기반으로 급성심근경색의 발병 및 질병에 영향을 미치는 위험인자를 찾는 연구가 활발히 진행되고 있다. 하지만 한국인에 적합한 급성심근경색 예후 진단 예측 시스템이 미비한 실정이다. 따라서 이 논문에서는 KAMIR(Korea Acute Myocardial Infarction Registry) 데이터베이스에서 제공 받은 급성심근경색 환자의 예후 데이터를 기반으로 ST분절 급상승 심근경색 재발 환자들의 단기 사망률 예측 모델을 찾고자 한다. 실험을 통해 로지스틱 회귀 분석에 의해 추출된 속성 집합을 적용하였을 때 기존의 원시 데이터 보다 높은 정확도를 얻을 수 있었으며, 인공신경망의 경우 다른 분류기법들보다 높은 성능을 보였다. 이를 통해 ST 분절 급상승 심근경색 재발 환자들의 단기 사망률을 예측함으로써 향후 고위험군 환자들의 관리에 도움을 줄 수 있을 것으로 기대한다.

▸ Keywords : 심혈관계질환, 급성심근경색, 데이터마이닝

• 제1저자 : 임광현 • 교신저자 : 류근호
• 투 고 일 : 2012. 06. 01, 심사일 : 2012. 07. 25, 게재확정일 : 2012. 08. 31.
* ㈜에이텍 시스템사업부 (System Business DEPT. ATEC Co.,Ltd)
** 충북대학교 컴퓨터과학과(School of Computer Science, Chungbuk National University, Chungbuk, Korea)
*** 충북대학교 전자계산학과(School of Electronic and Computer Engineering, Chungbuk National University, Chungbuk, Korea)
※ 이 논문은 2012년도 정보(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(No. 한국연구 2012-0000478)

Abstract

Recently, the cardiovascular disease has increased by causes such as westernization dietary life, smoking, and obesity. In particular, the acute myocardial infarction (AMI) occupies 50% death rate in cardiovascular disease. Following this trend, the AMI has been carried out a research for discovery of risk factors based on national data. However, there is a lack of diagnosis minor suitable for Korean. The objective of this paper is to develop a classifier for short-term relapse mortality prediction of cardiovascular disease patient based on prognosis data which is supported by KAMIR(Korea Acute Myocardial Infarction). Through this study, we came to a conclusion that ANN is the most suitable method for predicting the short-term relapse mortality of patients who have ST-segment elevation myocardial infarction. Also, data set obtained by logistic regression analysis performed highly efficient performance than existing data set. So, it is expect to contribute to prognosis estimation through proper classification of high-risk patients

▸ Keywords : Cardiovascular Disease, Acute Myocardial Infarction, Data-mining

I. 서 론

서구화된 식생활 패턴과 흡연, 비만 등의 원인으로 인해 심혈관계 질환들이 증가하고 있는 추세이다. 심혈관계 질환은 심장에 혈액을 공급하는 관상동맥인 우관상동맥, 좌전하행동맥, 좌회선동맥 등에 죽상반이 생기면서 발생한다[1]. 이러한 심혈관계 질환은 조기 사망과 장애를 초래하는 주요 원인으로서 심각한 요양급여 비용을 초래한다. 2009년 통계청 조사에 의하면 심혈관계 질환으로 인한 사망자는 인구 10만 명당 45명으로 10년 전 보다 16% 증가 하였으며, 이는 암을 포함한 모든 전체 질환으로 인한 사망 중 세 번째로 높은 수치이다[2]. 특히, 급성심근경색은 심혈관계 질환으로 인한 사망의 50%를 차지하고 있으며 꾸준한 증가 하고 있다[3]. 또한, 2008년 건강보험심사평가원 조사에 따르면 2002년 이후 급성 심근경색 발병률 증가는 둔화되는 반면, 재발률은 급격히 증가하는 추세를 보이고 있다. 그림 1은 2002년부터 2007년까지의 발병률과 재발률의 증가 추세를 나타낸다. 발병률은 2002년 인구 10만 명당 95.8건에서 2007년 118.4건으로 약 23% 증가율을 보였고 재발률의 경우 15.9건에서 26.6건으로 67%의 높은 증가율을 보였다[4].

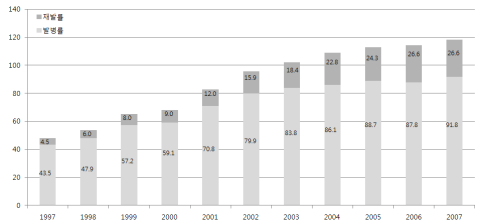

그림 1. 연도별 급성심근경색 발생률 및 재발률 증가 추세
Figure 1. Annual Trend of Attack ratio and Replace Ratio with Acute Myocardial Infarction

이러한 급성심근경색은 사망률과 재발률이 매우 높기 때문에 조기진단의 중요성이 대두 되고 있고 이러한 추세에 따라 연구가 활발히 진행되고 있다[5,6]. 최근 데이터 마이닝 기법은 심혈관계 질환에 영향을 미치는 임상적 요인의 관계를 규명하기 위해 활용되고 있다[7,8,9,10]. 하지만 기존 연구에서 해결하지 못한 내용들을 정리하면 다음과 같다. 첫째, 현재 연구들은 각 나라의 식습관 및 임상학적 요인의 차이로 인해 자국민 데이터를 기반으로 연구가 진행되고 있지만, 현재 데이터마이닝 기법을 기반으로 한국인의 특성을 고려한 급성심근경색 질환의 연구가 부족하다. 둘째, 높은 재발률을 가지고 있는 급성심근경색의 재발에 관한 연구가 필요하다. 셋째, 무작위로 수집된 의료 데이터의 경우 불안정한 데이터를 내포하고 있다. 이에 따라 이들을 효율적으로 처리할 수 있는 연구가 필요하다.

이들 문제를 해결하기 위하여 이 논문에서는 ST분절 급상승 심근경색 재발 환자의 단기 사망률 예측 모델을 제안한다.

이와 같은 목적을 달성하기 위한 이 논문의 구성은 다음과 같다. 2장에서는 관련연구로 급성심근경색 재발에 관한 연구 동향과 기존에 개발되어진 심혈관계 질환의 분류 모델 예측에 대해서 알아본다. 3장에서는 분류모델을 예측하는데 요구 되는 기법들에 대해서 설명한다. 4장에서는 분류 모델의 실험방법과 모델의 비교 및 평가를 한다. 5장에서는 결혼 및 향후 연구 방향을 제시한다.

II. 관련 연구

Tang[5]은 2007년에 발표한 논문에서 전 세계 14개국에서 수집되어진 급성심근경색 환자 1143명의 데이터를 기반으로 통계학적 기법을 이용하여 사망 위험률을 계산하는 통계학적 모형인 GRACE(Global Registry of Acute Coronary Event) 모델을 제안했다.

Kim[6]은 2010년에 발표한 논문에서 한국인에 맞는 급성심근경색 환자들의 예후를 추정한 연구이다. 이 연구에서는 전국의 주요 대학병원에서 수집된 5.458명의 데이터를 기반으로 통계학적 모형을 제시 했으며 ROC curve를 기반으로 GRACE 모델과 비교한 결과 보다 더 낳은 성능을 획득하였다.

Xing[7]이 2007년에 발표한 논문에서 중국 베이징 안진 대학 병원에서 제공받은 532명의 환자 생존 데이터 기반으로 지지도벡터 기계, 인공신경망, 의사결정 트리를 사용하여 관상동맥 질환을 예측 할 수 있는 분류 모델 생성을 제안했다. 사용되어진 데이터 속성은 기본적인 환자 데이터와 단백질 데이터를 사용했으며 분류기의 성능평가는 10차 교차검증을 사용했다. 성능평가 결과 지지도 벡터 기계가 92.1%로 가장 높은 정확도를 보였으며 그 다음으로 인공신경망이 91.0%, C5.0이 85.6%로 가장 낮은 정확도를 보였다.

Palaniappan[8]이 2008년에 발표한 논문에서 의사결정 트리, 베이지안 네트워크, 인공신경망 기반의 지능형 질환 예측 시스템을 제안했다. 이 시스템은 웹 기반으로 심장질환을 예측 할 수 있는 신뢰성 높은 서비스를 제공했다. 이 논문에서는 클리블랜드 심장 데이터베이스에서 제공받은 909개 데이터를 기반으로 사용했으며 이들 데이터들은 455개의 훈련 데이터와 454개의 테스트 데이터로 분리하였다..

M.Anbarasi[9]가 2010년에 발표한 논문에서 유전자 알고리즘을 이용한 특징 추출 기반의 심장질환 진단 분류 모델 생성을 제안했다. 이 논문에서 사용되어진 데이터는 클리블랜드 심장 데이터베이스에서 제공 받은 909개의 데이터를 사용했으며 특징 추출 기법을 이용하여 총 13개 들 중 6개의 속성을 추출하였다. 이들 속성을 기반으로 의사결정트리, 베이지안 네트워크, 분류 비아 클러스터링 모델을 생성하였으며 성능평가 결과 의사 결정트리가 99.2%로 가장 높은 정확도를 보였으며 그 다음으로 베이지안 네트워크가 96.5%, 분류 비아 클러스터링 기법이 88.3%로 가장 낮게 나왔다.

K.Srinva[10]가 2010년에 발표한 논문에서 심장 질환들의 효율적인 분류를 위해 베이지안 네트워크의 확장 알고리즘인 ODANB와 NCC2 사용해서 분류기의 정확도를 평가했다. 사용되어진 데이터집합은 UCI 연구실에서 공개 데이터 집합인 Heart-c, Heart-h, Heart - statlog을 사용했고 NCC2 분류기가 모든 데이터 셋에서 우수한 성능을 보였다.

이들 연구들은 각 나라 별 식습관 및 유전적 요인으로 인한 임상생리학적 오류를 줄이기 위해 각 나라의 데이터를 기반으로 연구가 진행되었다. 하지만 현재 데이터마이닝 기반에 한국인의 특성을 고려한 진단 모델이 미비한 실정이다.

III. ST 분절 급상승 심근경색 재발 환자들의 단기 사망률 예측

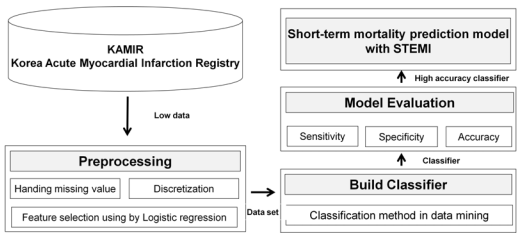

그림 2. ST 분절 급상승 단기 재발 사망률 예측을 위한 프레임 워크
Fig 2. Framework for Short-Term Mortality Prediction of Replace Patients with STEMI

그림 2는 이 논문에서 제안하는 ST분절 급상승 심근경색 환자들의 단기 사망률 예측 과정을 보여주고 있다. ST 분절 급상승 심근경색 재발 환자들의 단기 사망률 예측을 위한 단계는 4단계로 이루어진다. 첫째, 전처리 단계에서 데이터의 무결성을 보장하기 위해 이상치 제거 및 결측치를 처리하고 연속형 속성을 이산화 한다. 둘째, R language 기반으로 로지스틱 회귀분석을 적용하여 유의한 속성들을 추출한다. 셋째, 유의한 속성들을 기반으로 C4.5, 인공신경망, 지지도벡터 기계를 이용하여 분류기를 생성한다. 넷째, 10차 교차 검

증 기반으로 분류기들의 성능 평가를 통해 ST분절 급상승 심근경색 재발 환자들의 사망률 예측에 가장 적합한 분류기를 획득한다.

1. 데이터 전처리

KAMIR 데이터베이스에 포함되어 있는 데이터의 결측치는 분류 알고리즘을 통해 생성되는 모델의 성능을 저하시키기 때문에 평균 대체 방법을 적용하여 결측치를 대체함으로서 분석에 필요한 표본자료를 확보했다. 또한, 제공받은 데이터 속성 26개중 (Age, Height, Weight, BMI, Abdominal Circumference, Hip Circumference, SBP, DBP, Heart Rate, Lv ejection fraction, Regional wall mortion score, Glucose, Creatinine, Maximum CK, Maximum CK-MB, Maximum TroponinI, Total Cholesterol, Triglyceride, HDL, LDL, hsCRP, NT-proBNP) 23개가 연속형 데이터 이므로 분류 기법에 적용하기 위해서 데이터 이산화 작업이 필요하다[11].

이산화 기법은 인접한 구간을 분할하거나 병합으로 나눌 수 있다. 분리 기준점 선정은 목표 클래스 정보 활용 여부에 따라 유감독 이산화와 무감독 이산화로 나눌 수 있다. 이 논문에서는 분할 유감독 이산화 기법인 MDLP(Minimum Description Length Principle) 방법을 적용했다[12]. 이 기법은 정보획득에 바탕을 둔 하향식 이산화 기법으로 Fayyad와 Irani가 제안한 방법이다. 이 이산화 방법은 3단계로 구성된다.

1 단계 : 각 클래스의 속하는 데이터 비율을 기반으로 분할 지점 계산 S를 데이터 집합, $|S_i|$는 S의 부분집합인 S_i의 원소의 개수이며, $Ent(S_i)$는 (S_i)의 클래스 정보이득이다. 식(1)과 같다.

$$Ent(S_i) = -\sum_{j=1}^{k} P(C_j, S_i) Log(P(C_j, S_i)) \quad (1)$$

2단계 : 최소 엔트로피 구간을 최적의 분리 기준점으로 선정하고 S 데이터집합, X를 연속형 변수, T를 분할 점이라 할 때, T에 의한 평균 클래스 엔트로피 $E(X,T,S)$를 이용하여 평균 클래스 엔트로피를 최소화하는 T_x를 분할 점으로 한다. $E(X,T,S)$는 식(2)과 같다.

$$E(X,T,S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \quad (2)$$

3 단계 : 2단계에서 계산되어진 정보이득 값을 다음 식이 만족할 때 까지 재귀적으로 계산되어진다. 만약 이 식을 만족하지 않은 속성들은 이산화에서 기각되어진다. 즉, 유감독 이산화를 통한 이산화의 효과가 낮음을 의미한다. 이럴 경우 정략적 이산화 방법으로 삼분위수, 사분위수로 이산화 한다. $Gain(X,T;S) = Ent(S) - E(X,T;S)$이며, log는 밑을 2로 하고 K, K_1, K_2는 집합 S, S_1, S_2에 속해있는 클래스의 개수, $N = |S|$이며 식(3)과 같다.

$$Gain(X,T;S) \le \frac{Log(N-1)}{N} + \frac{Log(3^k - 2) - k_a Ent(S_a)}{N} \quad (3)$$

2. 특징 추출

특징 선택 기법은 많은 독립 변수 중에서 목표 변수에 큰 영향을 미치는 변수들을 추출함으로서, 많은 속성의 수로 인해 발생하는 차원의 저주 문제를 효율적으로 해결할 수 있는 기법이다.

이 논문에서는 최근 특징 추출에서 활용되고 있는 로지스틱 회귀 분석 기법을 이용하여 속성을 추출 하였다[13]. 역학적 연구나 의학 데이터 등에서는 종속 변수가 질병의 발생 유무 혹은 생존과 사망 등으로 표시되는 이항적인 사건을 다루기 때문에 일반 회귀분석법을 그대로 적용할 수 는 없다. 왜냐하면 회귀분석은 연속형 데이터 분석에 대한 예측만이 가능하므로 범주형 데이터의 이항 사건에 대한 문제는 다룰 수 없다. 이러한 문제를 해결하기 위한 방법으로 로지스틱 회귀 분석이 사용될 수 있다. 일반적으로 회귀분석에서의 모형은 주어진 독립변수$(X_1, X_2, ..X_k)$에 종속변수의 평균이 독립변수에 대한 선형 식으로 표현된다.

즉 $E[Y|X] = \beta_0 + \beta_1 x_1 .. + \beta_k x_k$로 나타낼 수 있다. 로지스틱 분석에서는 반응 변수가 2개의 가능한 값을 갖는 이항 반응인 경우가 대부분이다. 따라서 반응 변수가 이항변수 $E[Y|X] = px$ 가 되고 px는 식(4)과 같이 모형화 된다.

$$px = \frac{\exp(\beta_{0+}\beta_1 x_1 + ... + \beta_x x_k)}{1 + \exp(\beta_{0+}\beta_1 x_1 + ... + \beta_x x_k)} \quad (4)$$

이때 로지스틱 모형은 식(5)과 같다.

$$(\frac{p_x}{1-p_x}) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k \quad (5)$$

로지스틱 모형을 통해 궁극적으로 알고자 하는 것은 유의한 독립변수가 반응변수를 변동시키는 정도가 과연 얼마 정도인가 이다. 여기서 독립변수가 반응 변수를 변동 시키는 정도를 대응 위험도(odds ratio)라 한다. OR의 경우 로짓 값의 차이로 나타낼 수 있는데 가령 X_1의 효과를 보고자 할 때 다른 공변량을 고정시킨 상태에 $(\frac{p_1}{1-p_1}) - (\frac{p_0}{1-p_0})$을 해주면 $X_1 = 0$일 때에 비해 $X_1 = 1$일 때 $\log(OR)$값을 구할 수 있다. 따라서 $(OR) = b$가 되고 OR은 $\exp(b1)$으로 나타낼 수 있다[14].

3. 분류 기법

분류 기법은 의미 있는 패턴이나 규칙을 추출 할 수 있는 데이터 마이닝의 대표적인 기법이다[20,21]. 이 논문에서는 최근 들어 생물학, 유전자, 의학 분야에서 널리 사용되고 있는 C4.5, 인공신경망(artificial neural network), 지지도 벡터 기계(support vector machine)와 같은 분류 기법을 이용하여 연구를 진행하였다[15,16,17].

3.1 C4.5

의사결정 트리는 분류 모델의 생성 결과를 나무 구조로 나타나기 때문에 이해가 쉽고 해석이 용이한 장점을 가지고 있다. 이 논문에서 사용하는 C4.5는 ID3 알고리즘을 확장한 것이다. ID3의 경우 명목형 속성 밖에 처리 할 수 없다는 한계점과 속성의 분할 기준점이 많을수록 불순 척도 값이 낮아지는 데이터 편향 문제를 가지고 있다. C4.5는 이러한 문제점들을 해결하기 위해서 엔트로피 기반 이득 비율을 사용하여 데이터 편향을 제거 했다. 이득 비율은 식(6)과 같다.

$$GainRatio(X,T) = \frac{Gain(X,T)}{SplitInfo(X,T)} \quad (6)$$

$SplitInfo$는 어떤 속성 X에 의해 의사결정트리가 n개의 부분 노드로 분할되면서 발생하는 정보의 양을 의미한다. 분리 정보를 계산하여 정보 이익을 보정함으로써 이득 비율을 얻을 수 있다. C4.5는 가장 큰 이득 비율 갖는 속성을 분리 속성으로 선택한다[15].

3.2 인공신경망

인공신경망은 인간의 뇌 작동 원리를 이용하여 개발된 분류 방법이다. 인공 신경망을 통한 분류기 생성은 매우 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾는데 유용하다. 인공신경망은 입력층, 출력층, 은닉층으로 구성된다. 인공신경망은 반복적으로 입력되는 정보의 패턴 신호에 대하여 가중치를 목적에 맞도록 변환시킴으로써 분류 모델을 학습 시킨다. 각 신경망의 노드마다 다른 가중치를 가지며 신경망에 의해 계산된 실제 출력 값과 목표 출력 값과의 오차를 최소화하기 위해 신경회로망의 가중치를 자동적으로 변환 시킨다. 인공신경망은 실제 세계에서 나타나는 상당히 복잡한 현상이나 문제를 단순화된 수리 모델을 사용하여 분석하는데 유용하다. 즉, 질병 진단과 같은 여러 속성들의 긴밀한 상호 연관성에 일어나는 사건과 같은 예측에 장점을 가지고 있다.

3.3 지지도 벡터 기계

지지도 벡터 기계는 Vanpnik에 의해 개발된 분류 알고리즘으로 객체 인식, 텍스트 분류 등 많은 문제들을 성공적으로 해결한 기법이다. 지지도 벡터 기계는 SRM(Structural Risk Minimization)이라 부르는 통계적 학습 원리를 적용하여 클래스들을 분류하기 위한 최대 마진을 가지는 초평면을 찾기 위하여 학습한다. 지지도 벡터 기계에서는 특정 공간에 대한 선형 분리가 불가능한 경우, 원래의 입력 공간을 새로운 고차원의 특징 공간으로 사상하여 복잡한 비선형 형태의 데이터들도 각 클래스로 분류 할 수 있다. 또한, 지지도 벡터 기계는 최적 초평면을 선택함으로서 훈련 데이터의 과잉적합 문제를 방지 할 수 있다. 따라서 의료 데이터와 같이 표본의 수가 변수의 수보다 작은 대용량 데이터를 분석하는데 적합하다.

Ⅳ. 실험 및 평가

이 장에서는 ST분절 급상승 심근경색 재발 환자들의 사망을 예측하기 위해서 KAMIR 데이터베이스로부터 목표 모집단을 설정하고 MDLP 이산화 기법을 적용하여 모집단을 생성한다. 생성된 모집단을 기반으로 로지스틱 회귀분석을 이용하여 속성을 추출 하고 분류기법을 이용하여 생성된 예측 모형을 민감도, 특이도, 정확도를 비교하여 성능평가 한다.

1. 실험 환경

이 논문에서 제안 하는 ST분절 급상승 심근경색 환자들의 단기 사망 예측 정확도 평가를 위해 Intel Core 2 Duo E4600 @2.40GHz, Ram 2GB를 사용하는 Microsoft

Window 7 시스템에서 실시하였으며 Weka 3.64와 R 2.13.2를 사용하였다[18,19].

2. KAMIR 데이터베이스

KAMIR 데이터베이스는 급속하게 증가하고 있는 급성심근경색 환자들의 사망률을 둔화시키기 위해서 전국 24개 병원에서 2005년 11월 1월부터 2008년 1월 31일까지 급성심근경색에 대한 광범위한 임상적 데이터를 모은 것이다.

3. 연구 모집단 생성

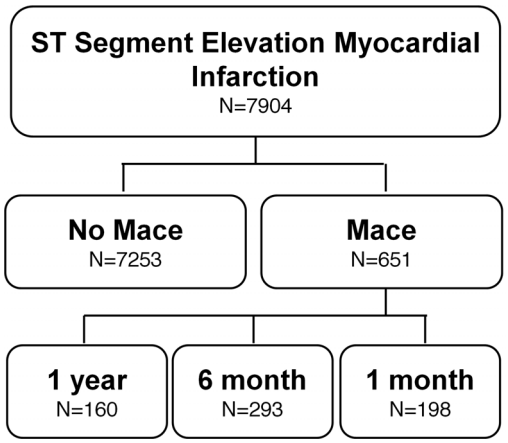

그림 3 연구 모집단
Fig 3 Population

ST 분절 급상승 심근경색 환자들의 단기 재발 사망률 예측을 위해서 연구 모집단 생성은 그림 3과 같다. 이 논문에서는 잠재적 혼돈 변수를 제거하기 위해서 전체 ST분절 급상승 심근경색 환자 7904 명중에 MACE(Major Cardiac Event)가 발생한 651명 중 1달 안에 MACE가 발생한 198명 환자들로 구성했다.

4. 데이터 전처리

이 연구에서 사용하는 데이터베이스에는 다수의 연속형 속성이 포함되어 있다. 따라서 이들 연속형 데이터 값을 범주형 값으로 변환하는 MDLP 이산화 방법을 적용했다. 표1은 이산화 결과이다.

표 1. 이산화된 속성들의 분리 기준점
Table 1. Division Standard Point of Discretized Attributes

Attributes	Division point
BMI	28.5
hsCRP	0.545
Creatinine	1.65
NT-proBNP	24772.5

5. 속성 추출

속성 추출 방법으로는 로지스틱 회귀 분석을 실행한 후 p-value의 결과를 통해 의미 있는 속성들을 선택하였다. 유의속성 추출 결과 키 (height) P〉0.0303, 몸무게(Weight) P〉0.03535, 엉덩이 둘레(Abdominal ircumference) P〉0.01726, 체지방율(BMI) P〉0.03528, 킬립클레스 4(Killip class) P〉0.00761, 크레아티닌(Creatinine) P〉0.00137이 유의속성으로 추출되었다. 하지만 기존의 급성심근경색의 강력한 예측 인자라고 밝혀진 NT-proBNP , 고감도C반응성 단백질(hsCRP), 전체 콜레스테 롤(Total Cholesterol), 고밀도지단백 콜레스테롤(High Density Lipoprotein Chol esterol), 저밀도 지단백 콜레스테롤(Low Density Lipoprotein Cholesterol)등은 급성심근경색 재발에 관한 환자들의 단기 사망률 예측에 유의 하지 않았다. 따라서 이 논문에서는 유의속성으로 확인된 키, 몸무게, 엉덩이둘레, 체지방율, 킬립클레스4, 클레아티닌 속성을 이용하여 분류기법에 적용하였다.

6. 성능 평가 방법

이 논문에서 분류기법들의 평가를 위해 의학 분야에서 널리 활용되고 있는 혼동행렬 기반의 민감도(Sensitivity), 특이도(Specificity), 정확도(Accuracy) 이용하여 비교 분석하였다.

민감도는 목표 클래스 라벨을 옳다고 예측 했을 때 그 결과가 옳은 것들을 의미한다. 식은 다음과 같다.

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

특이도는 목표 클래스 라벨이 틀리다고 예측 했을 때 그 결과가 틀린 것들을 의미한다. 식(8)과 같다.

$$Specificity = \frac{TN}{FP + TN} \quad (8)$$

정확도는 민감도와 특이도의 합에 연산이다. 식(9)과 같다.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

7. 모델의 비교 및 평가

이 장에서는 로지스틱 회귀 분석을 통한 속성 추출 전과 후에 따른 분류기의 민감도, 특이도, 정확도를 비교한다.

그림 4은 로지스틱 회귀분석을 이용하여 추출된 속성들을 분류기에 적용한 것과 기존 데이터를 적용한 분류기법들 간에 민감도를 비교한 것이다. 그 결과, 인공신경망의 경우 민감도는 3.5%향상된 70.7%를 보였다. C4.5는 2.6% 상승된 66.7%의 민감도를 얻었다. 지지도 벡터 기계는 민감도가 3%증가한 69.7%를 보였다. 민감도 성능평가 결과, 로지스틱 회귀분석을 이용한 속성 추출 적용한 경우 인공신경망, C4.5, 지지도벡터 기계 기법들의 성능이 개선을 볼 수 있었다. 특히, 인공신경망의 경우 다른 분류 기법보다 실제 급성 심근경색 재발 환자들의 사망을 가려낼 수 있는 능력이 뛰어나다는 사실을 설명해 주고 있다.

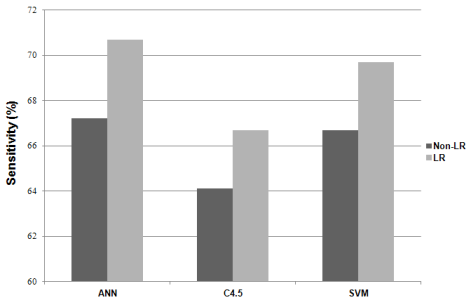

그림 4. 분류기들의 민감도
Fig 4. Sensitivity of Classifiers

그림 5은 로지스틱 회귀분석을 이용하여 속성을 추출한 데이터를 분류기에 적용한 것과 그렇지 않은 분류기들 간에 특이도를 비교한 것이다. 인공신경망의 경우 특이도가 5.3% 향상된 71.7%를 보였다. C4.5는 5.3% 향상된 68.8%의 특이도를 얻었다. 지지도 벡터 기계에서는 4%증가한 70.3%를 보였다. 특이도 성능평가 결과 로지스틱 회귀분석을 이용한 속성 추출을 적용한 경우 인공신경망, C4.5, 지지도벡터 기계 기법들의 성능 향상을 볼 수 있었으며 특히, 인공신경망의 경우 다른 분류기들 보다 높은 특이도를 보였다. 즉, 다시 말하면 인공신경망의 경우 다른 분류 기법보다 실제 급성심근경색 재발 환자들의 단기 사망 하지 않는다는 사실을 가려낼 수 있는 능력이 더 좋다는 것을 의미한다.

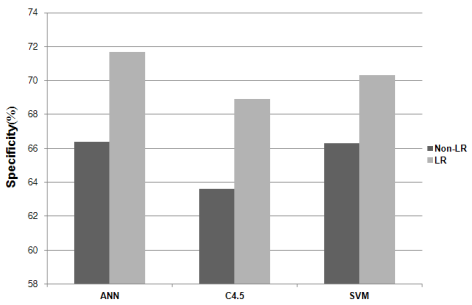

그림 5. 분류기들의 특이도
Fig 5. Specificity of Classifiers

그림 6은 로지스틱 회귀분석을 이용하여 추출된 속성을 기반으로 한 데이터 집합을 분류기에 적용한 것과 속성을 추출하지 않고 분류기에 적용시킨 분류기들 간에 정확도를 비교한 것이다. 인공신경망에서 기존의 데이터 집합에서 67.2%의 정확도를 보였으며 로지스틱 회귀분석을 이용한 속성 추출 후 정확도는 3.5%향상된 70.7%를 보였다. C4.5에서 기존 모집단 기반으로 분류 모델 정확도는 64.1%을 가졌으며 로지스틱 회귀분석을 적용하였을 때 2.6% 상승된 67.7%의 정확도를 얻었다. 지지도 벡터기계의 경우 기존의 데이터 집합에서 66.7%를 얻었으며 로지스틱 회귀분석을 이용한 속성 추출한 후 정확도는 3%증가한 69.7%를 보였다. 정확도 성능평가 결과 로지스틱 회귀분석을 이용한 속성 추출을 적용한 경우 인공신경망, C4.5, 지지도벡터 기계 기법들의 성능의 개선을 볼 수 있었다. 특히, 인공신경망의 경우 다른 분류 기법보다 실제 ST분절 상승 급성심근경색 단기 재발 환자들의 사망률 예측에 적합하다는 것을 알 수 있다.

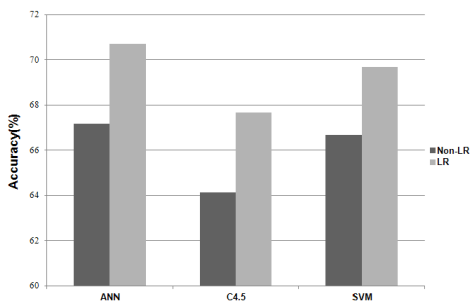

그림 6. 특징추출 전과 후 분류기들의 정확도
Fig 6. Accuracy of Classifiers

V. 결론 및 향후연구

급성심근경색은 심혈관계 질환으로 인한 사망의 대부분을 차지하고 있으며 지속적으로 증가하고 있다. 특히, 국내의 경우 급성심근경색 재발의 비율이 급격하게 증가하고 있다. 하지만 현재 한국인의 특성을 고려한 급성심근경색 재발 진단 마이너가 미비한 실정이다. 따라서 이 논문에서는 ST 분절 급상승 심근경색 환자들의 단기 재발 사망률 예측모델을 제안한다. 이와 같은 목적을 달성하기 위해 원시 데이터들에 포함되어진 잠재적 혼돈 변수들을 전 처리하였으며, 이들 데이터들을 기반으로 기존에 개발되어진 분류기법에 적용하여 성능을 평가 하였다. 성능평가 결과, 로지스틱 회귀 분석을 통해 추출된 속성을 사용한 분류기들의 성능은 민감도, 특이도, 정확도에서 속성 추출 적용하지 않은 분류기들보다 성능이 향상되었다. 특히, 인공신경망은 다른 분류 기법들에 비해 더 좋은 예측 정확도를 확인 할 수 있었다. 따라서 ST 분절 급상승 심근경색 재발 환자들의 단기 사망률을 예측하는 데 있어서 로지스틱 회귀분석을 이용해 추출되어진 속성을 기반으로 데이터 마이닝의 분류기법인 인공신경망을 이용한 분류 기법이 적합하다는 결론을 얻었다. 이 연구를 통해 고위험군 환자에 대한 ST 분절 급상승 심근경색 재발 환자들의 단기 사망률 예측에 도움을 줄 수 있을 것으로 기대한다.

향후연구로는 좀 더 다양한 분류 기법과 속성 추출 기법들 간에 비교를 통한 가장 적합한 데이터 마이닝 모델을 개발하기 위한 연구가 진행되어야 할 것이다.

참고문헌

〔1〕 "Cardiovascular Update", vol.8, no.6, pp 10-39, 2006.

〔2〕 Population Trends and Statistics, Statistics Korea,"In 2009, deaths and causes of death statistics", pp.9, 2009.

〔3〕 I. S. Kim, "The present condition and trend of five major causes of death in Korean. Korean J Med Assoc 1995, vol.38, 132-45", 1995.

〔4〕 "Occurring trend research of medical expenses with acute myocardial infarction", Health Insurance Review and Assessment Service , pp 45-49, 2008.

〔5〕 E. W. Tang, C. K. Wong and P. Herbison, "Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome" America Heart journal, vol.154, no.5, pp.29-35, 2007.

〔6〕 H. K. Kim, M. H. Jeong, Y. Ahn, J. H. Kim, S. C. Chae, Y. J. Kim, S. H. Hur, I. W. Seong, T. J. Hong, D. H. Choi, M. C. Cho, C. J. Kim, K. B. Seung, W. S. Chung, Y. S. Jang, S. W. Rha, J. H. Bae, J. G. Cho, S. J. Park and Other Korea Acute Myocardial Infarction Registry Investigators, "Hospital Discharge Risk Score System for the Assessment of Clinical Outcomes in Patients With Acute Myocardial Infarction (Korea Acute Myocardial Infarction Registry〔KAMIR〕Score)", The American Journal of Cardiology, vol.107, no.7, pp.965-971, 2010.

〔7〕 Y. Xing, J. Wang , Z. Zhao, Y. H. Gao, "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease", International Conference on Convergence Information Technology", pp.868-872, 2007.

〔8〕 S. Palaniappan, R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IEEE.ACS International conference Computer system and application, vol.8, no.8, pp.340-384, 2008.

〔9〕 M. Anbarasi, E. Anupriya, N. CH. S. N. IYENGAR, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", ISSN: International Journal of Engineering Science and Technology, vol.2, no.10, pp.5370-5376, 2010.

〔10〕 K. Srinivas, B. K. Rani, A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", OJCSE: International Journal on Computer Science and Engneering, vol.2, no.2, pp. 250-255, 2010.

〔11〕 L. Peng, W. Qing, G. Yujia, "Study on comparison of Discretization Method,"International Conference on Artificial Intelligence and Computational Intelligence, vol.4, pp.380-384, 2009.

〔12〕 U. M. Fayyad, K. B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning", Artificial Intelligence, vol.13, pp.1022-1027, 1993.
〔13〕 V. Bourders, J. Ferrieres, J. Amar, E. Amelineau, S. Bonnevay, M. Berlion, N. Danchin, "Prediction of persistence of combined evidence-based cardiovascular medication in patients with acute coronary syndrome after hospital discharge using neural network", Medical and Biological Engneering and Computing, vol.49, no.8, 2011.
〔14〕 C. T. J. Peng, K. L. lee, G. M. Ingersoll, "An Intoroduction to Logistic Regression Analysis and Reporting", The Journal of Educational Research, pp.3-14, 2002.
〔15〕 J. R. Quinlan, C4.5: Programs for machine learning",Morgan Kaufman, San Fransisco, 1993.
〔16〕 Abraham, A., "Artificial Neural Networks", In Handbook of Measuring System Design, pp.901-908, 2005.
〔17〕 D. C. Li and C. W. Liu, "A class possibility based kernel to increase classification accuracy for small data set using support vector machines," Expert Systems with Application, vol.37, no.4, pp.3104-3110, 2010.
〔18〕 The University of Waikato, http://weka.wikispaces .com/, 2012
〔19〕 The R Project for Statistical Computing http://www.r-project.org/, 2012
〔20〕 M. Y. Hwang, C. H. Jin, U. Yun, K. D. Kim, K. H. Ryu, "Building of Prediction Model of Wind Power Generationusing Power Ramp Rate", The Korea Society of Computer and Information, Vol.17, No.1, pp.211-218. 2011.
〔21〕 D. H. Suh, K. I. Kim, K. D. Kim, K. H. Ryu, "Predicting Power Generation Patterns Using the Wind Power Data", The Korea Society of Computer and Information, Vol.16, No,11, pp.245-254. 2011

저 자 소 개

임 광 현

2003 : 한양대학교 전자계산학과 석사
2011 : 충북대학교 컴퓨터과학과 박사 수료
현 재 : (주)에이텍 시스템사업부 근무
관심분야 : 멀티미디어 데이터마이닝
Email : khlim888@naver.com

류 광 선

2012 : 충북대학교 퓨터과학과 석사
현 재 : 충북대학교 컴퓨터과학과 박사과정
관심분야 : 메디컬인포매틱스, 데이터마이닝
Email : ksryu@dblab.chungbuk.ac.kr

박 수 호

2011 : 충북대학교 컴퓨터공학과 공학사
현 재 : 충북대학교 컴퓨터과학과 석사과정
관심분야 : 메디컬인포매틱스, 데이터마이닝
Email : soohopark@dblab.chungbuk.ac.kr

손 호 선

2010 : 충북대학교 컴퓨터과학과 이학 박사
현 재 : 충북대학교 이터베이스/바이인포메틱스 Post-Doc
관심분야 : 메디칼인포메틱스, 데이터마이닝
Email : shon0621@dblab.chungbuk.ac.kr

류 근 호

1976 : 숭실대학교 전산학과 이학사
1980 : 연세대학교 공학대학원
전산전공 공학석사
1988 : 연세대학교 대학원
전산전공 공학박사
1976 ~ 1986 :
육군군수 지원사 전산실(ROTC 장교),
한국전자통신연구원(연구원),
한국방송통신대 전산학과(조교수) 근무
1989 ~ 1991 :
Univ. of Arizona Research Staff
(TemplS 연구원, Temporal DB)
1986년 ~ 현 재 :
충북대학교 전기전자 컴퓨터공학부 교수
관심분야 : 시간 데이터베이스,
시공간 데이터베이스,
Temporal GIS,
지식기반 정보검색 시스템,
유비쿼터스 컴퓨팅 및 스트
림데이터처리, 데이터 마이닝,
데이터베이스, 보안,
바이오 인포메틱스
Email :
khryu@dblab.chungbuk.ac.krr