나이브 베이지안 방법을 위한 데이터 변환법으로 한국인 급성 심근경색증 환자의 예후를 예측하는 성능의 향상

조선호*, 김정수**, 권혁철*
*부산대학교 컴퓨터공학과
**양산부산대학교병원 심혈관센터
e-mail:{sean, jeongsu.kim, hckwon}@pusan.ac.kr

Development of Performance to Predict the Prognosis of Korean Patients with Acute Myocardial Infarction by Data Transformation for Naïve Bayes Method

Sun Ho Cho*, Jeong-su Kim**, Hyuk-Chul Kwon*
*Dept. of Computer Science & Engineering, Pusan National University
**Division of Cardiology, Cardiovascular Center, Pusan National University Yangsan Hospital

요 약

오늘날 한국에서는 급성 심근경색증으로 인한 사망률이 높은 상태로, 발병 시에 치료까지 신속한 의사결정이 요구되는 위중한 질병이기 때문에, 한국인에게 맞는 급성 심근경색증 연구가 매우 중요하다. 본 연구는 한국인 급성 심근경색증 등록 데이터를 이용해 기계 학습 방법의 한 종류인 나이브 베이지안 방법을 이용해 급성 심근경색증 환자의 예후를 예측하고자, 의료 데이터의 특성에 따른 데이터 변환 방법을 제안한다. 타겟 클래스에서 보다 중요한 의미를 가진 death 값에 대해 각 값을, nominal value, numeric value, 결측치로 구분한 방식에 따라, 확률을 계산해 변환한다. 실험 결과를 통해 결측치를 피처마다 존재하는 값들의 평균을 낸 값으로 대입하였을 때 가장 좋은 성능임을 알 수 있었는데, 기존의 방법에 비해 precision=5.4%, recall=7.0%의 성능이 향상되었다. 따라서 제안한 방법은 나이브 베이지안 방법의 예측 성능 향상에 기여하였다고 판단된다. 이후 적용했던 데이터 변환 방법을 여러 가지 기계 학습 방법에서 판단해보고, 다른 타겟 클래스에도 시험해보고자 한다.

1. 서론

심장질환 중 급성 관상동맥증후군(Acute Coronary Syndrome; ACS)은 서구뿐만 아니라 생활양식이 서구화된 한국에서도 높은 사망률을 보이는 치명적인 질환이기에, 많은 연구가 필요하다. 2013 년 말에 발표한 통계청 자료에 따르면 한국인의 3 대 사인은 암, 심장질환, 뇌혈관 질환인데, 2012 년은 전년 대비하여 심장질환의 사망률이 뇌혈관 질환의 사망률을 앞질렀다. 2012 년의 심장질환 사망자 수는 인구 10 만 명 당 52.5 명이고, ACS 중에서도 사망률이 높은 질병인 급성 심근경색증으로 인한 사망은 10 만 명 당 19.6 명에 이르며, 이 수치는 지난 10 년간 꾸준히 유지되고 있는 추세이다[1]. 급성 심근경색증은 발병부터 치료 행위까지 걸린 시간이 환자의 생명에 직접적인 영향을 끼친다. 이를 위해서는 환자의 상태에 대한 판단과 치료 방법에 대한 의료진의 신속한 의사결정이 중요하다. 의학계에서는 환자 상태에 따른 치료 전략을 세워두고, 새로운 연구 결과에 따른 정보를 반영해 수정한다. 이를 위해서 국가마다 대규모 임상 환자 데이터를 구축해왔고, 이를 기반으로 한 연구가 진행되고 있다. 본 연구는 대규모 한국인 급성 심근경색증 환자 데이터를 이용하여, 기계학습 중에서 나이브 베이지안 방법에 적합한 형태로 변환시켜, 환자의 예후를 예측하는 성능을 향상하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2 장에서는 대규모 심장병 환자 데이터를 이용한 연구 동향을, 3 장에서는 실험에 이용된 한국인 급성 심근경색증 환자 데이터에 대해 설명한다. 4 장에서는 나이브 베이지안 방법에 적합한 데이터 변환 방법에 대해서 설명한다. 5 장에서는 실험 성능의 비교 및 평가를 한다. 마지막으로 6 장에서는 결론 및 향후 연구 방향을 제시한다.

2. 관련연구

서구에서는 ACS 연구를 위해 발병 환자의 데이터를 수집하고 이를 이용해 다양한 예측 모델을 개발해왔다. Cleveland Clinic Foundation Multi-Center[2], New York State, American College of Cardiology-National

Cardiovascular Data Registry[3] 등이 있으나, 가장 대표적인 것은 전 세계 30 개국에서 ACS 환자의 자료를 대량으로 수집해서 만든 GRACE(The Global Registry of Acute Coronary Events) 데이터이다[4]. GRACE 데이터를 이용해 Tang 은 2007 년에 1,143 명의 ACS 환자 정보를 통계학적 분석을 통해, 환자의 사망을 예측할 수 있는 주요 인자와 값을 제시한 GRACE risk model 을 제안했다[5]. 이 모델은 ACS 발병 후 장기간의 생존 여부를 예측할 수 있는 7 가지 주요 인자로 환자의 나이, 허혈성 질환 이력, 심장마비 이력, 입원 시의 심박 수, 초기의 세럼 크레아티닌 레벨, 근육 괴사 증거, 관상동맥 중재술 여부를 제시했다. 이와 같이 각 국가의 의료기구나 단체에서는, GRACE 데이터를 이용해 자국민에게 적합한 risk model 로 수정하여 검증하거나, GRACE 데이터에서 고려한 인자들을 중심으로 환자 데이터를 수집하여 연구하였다.

한국에서는 급성 심근경색증 환자의 자료를 수집한 한국인 급성 심근경색증 환자 등록(Korean Acute Myocardial Infarction Registry; KAMIR) 데이터를 구축하였다[6]. 이를 이용해 Kim 은 2011 년에 한국인의 특성에 맞는 급성 심근경색증 환자들의 예후를 추정하기 위한, GRACE risk model 을 개량한 KAMIR risk model 을 제안했다[7]. 이 연구는 한국의 주요 대학병원에서 두 기간에 걸쳐 수집된 총 5,458 명의 급성심근경색증 환자 데이터를 이용, 모델을 생성하여 GRACE 모델과 성능을 비교했다. KAMIR 모델은 6 개의 주요 인자들로 환자의 나이, Killip class, 세럼 크레아티닌, PCI 시술 여부, 좌심실 구혈률, 입원 시의 포도당 수치를 제시했다.

대용량의 환자 데이터의 분석과 예측을 위해 기존 의학계는 student's t-tests, $\chi 2$-tests, multivariate logistic regression analyses 를 이용한다. 이러한 방법들은 해당 도메인의 전문가에 의해서 미리 선택된 피처만 고려되고, 피처마다 중요도에 따른 가중치 조절이 직접적으로 어렵다는 한계가 있다.

Mair J. 등은 1995 년에 급성 심근경색증의 빠른 진단을 위해 의사결정 트리 방법을 고려하였다[8].

2000 년에 Cynthia J. Sims 등은 임산부의 상태에 따라 제왕절개 분만(cesarean delivery)을 결정하는데 logistic regression models 와 비교를 통해 기계학습 방법의 하나인 규칙 기반 의사결정 트리 방법을 사용할 수 있음을 확인하였다. 규칙 기반 의사결정 트리 방법의 결과는 의사가 더 쉽게 이해할 수 있고, logistic regression models 에서 감지되지 않는 변수 간의 인과관계의 종속성을 표시한다는 장점을 부각하였다[9]. 이후 의학 데이터 분석에 기계학습 방법을 이용한 많은 연구 시도가 있었다.

현재 기계학습 방법을 이용해 가장 활발히 연구되고 있는 심장병 분야는, Support Vector Machines 을 이용해 환자의 심장박동파형이나 심혈관 이미지에서 심장병 종류를 분류하는 연구들이다. R. Gouripeddi 등은 Support Vector Machines 를 이용해 Drug Eluting Stent 시술에 따른 합병증의 위험을 예측하는 연구를 했다[10]. 이와 같이 ACS 나 심근경색증 환자의 임상 데이터를 기계 학습 방법을 이용한 연구들 대부분은, 임상환자 수가 수십에서 수백 명 이내인 소량의 데이터를 이용하는데 그쳤거나, 국소적인 의학적 목표로 설계된 데이터를 사용한 연구라는 한계점을 지니고 있다.

3. 데이터

연구에 사용한 KAMIR 데이터는 한국 심근경색증 연구회(Korea Working Group on Myocardial Infarction)에서 주관하여 한국인의 실정에 맞는 급성 심근경색증의 치료에 관한 연구를 진행하기 위해 국내 41 개 일차적 관상동맥 중재술(PCI)이 가능한 병원에서 인터넷을 이용한 급성 심근경색증에 대한 데이터베이스의 구축과 등록 연구의 결과물이다[6].

KAMIR 데이터는 2005 년 11 월부터 2008 년 1 월까지, 일관된 목적 하에 환자마다 동일한 문항(이후 피처 단위가 됨)을 수집하였다. 원시 데이터의 구성은 의료 데이터 분석에 적합한 형태였지만, 일부 피처들은 수집 당시에 규칙 없이 의료진마다 자유롭게 입력한 문자열 그대로를 포함하고 있었다. Data sparseness 문제를 막고자, 이를 의학 전문가의 도움으로 의학적인 의미를 가진 nominal value 를 가진 피처로 분류·변환하여 정제하였다. 또 시각과 같은 이산적인 피처들로부터 의학적인 관점의 numeric 피처를 도출하였다.

<표 1> KAMIR 데이터에서 분류한 카테고리들

Category	No.	Features
환자 신체	7	Age, Gender, BMI, Central Obesity, Ratio of Waist & Hip
환자 이송	7	First contact medical center, Vehicles to first medical center, Transferred from other hospital?, Resuscitation prior to arrival, etc.
환자 진단	28	Symptoms on admission, Previous angina before MI symptom, Chest Pain, Dyspnea, Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, etc.
평소 복용약	21	Aspirin, BetaBlocker, ACEi, ARB, Cilostazol, Clopidogrel, Digoxin, Diuretics, Eztrole, CCB, etc.
혈액 검사	13	Creatinine_on_admission, Total_cholesterol, HDL, LDL, etc.
시술 치료	61	Initial therapeutic strategy in STEMI, Initial therapeutic strategy in NSTEMI, PCI, etc.
원내 처방약	15	Aspirin, BetaBlocker, ACEi, etc.
합계	152	

<표 2> 실험 데이터의 타겟 클래스에 따른 구분

구분	survival (생존)	death (사망)	전체
발병 후 6 개월 이내	7,588 명 (88.2%)	1,020 명 (11.8%)	8,608 명 (100%)

당초 14,871 명의 환자 데이터 중에서 생사 정보가 누락되었거나 정확하게 판단하기 어려운 사례를 제외한 후에, 생사 판단이 가능한 환자의 합계는 8,608 명이었다. 이때 전 기간에 걸쳐서 12.7%의 환자만이 사망하였다. 기계 학습을 통한 분류를 위해, 환자의 사망 시기에 맞춰 의학계에서 분류하는 방법대로 원내 사망 여부, 퇴원 후 1 개월 내 사망 여부, 퇴원 후 6 개월 내 사망 여부, 퇴원 후 12 개월 내 사망 여부를 표시하고 구분하였다.

본 연구에서는 GRACE risk score model[5], KAMIR risk score model[7]과의 관점 비교를 위해, 급성 심근경색증의 발병으로부터 6 개월 이내의 생사 여부를 타겟 클래스로 정하여 예측 실험을 하였다.

4. 연구 방법

연구에 사용한 KAMIR 데이터는 34.6%의 결측치를 포함하고 있고, 타겟 클래스에서 'survival'(환자의 생존을 의미)과 'death'(환자의 사망을 의미) 값이 약 9:1 로 상당히 편중된 데이터임을 알 수 있다. 이러한 데이터를 나이브 베이지안 방법으로 우수한 예측 성능을 얻기 위해, 다음과 같은 점을 고려한 데이터 변환 방법을 고안하였다. (1) 약 35%나 되는 결측치 대신에 대표 값을 부여해서 학습 능력을 강화한다. (2) 특정한 분포 영역에서 크게 벗어난 값의 인근에 결측치가 존재한다면, 이것에 대표 값을 적용해 강화 또는 약화함으로서 스무딩 할 수 있다. (3) 의학적으로 분류되는 경계 구간에 인접한 값인 경우에 그 특성이 무시되는 점을 보완하여, 경계에 위치한 값의 특성을 강화한다.

이를 위해 KAMIR 데이터에서 피처마다 여러 가지 형태와 범위를 가지는 개별 값들을, 타겟 클래스의 특정 값과 대응될 확률을 구해 0~1 사이의 값으로 모두 변환하는 방법을 제시한다. 기존 데이터는 각 피처마다 각기 다른 범위와 값의 형태를 가지고 있기 때문에 피처 개수만큼의 차원을 갖게 되지만, 데이터 변환 후에는 각 피처의 값들이 타겟 클래스의 특정한 값을 가지는 확률로 바뀌므로 차원은 0 에서 1 사이의 확률로 줄어든다. 본 연구에서는 기준을 타겟 클래스의 두 값인 'survival'과 'death' 중에서 의학적 관점에서 더 중요한 'death'로 지정했다.

나이브 베이지안의 변환식은 다음과 같다.

$$C_{naive\ Bayes} = argmax_{C_i} P(c_i) \prod_j P(v_j|c_i)$$
$$(C_i \in \{death, survival\})$$

Nominal value 는 타겟 클래스의 값이 'death'인 인스턴스 수에 따라 확률을 구할 수 있다. 예를 들어 전체 환자가 100 명인 데이터에서, '성별' 피처가 '남자'와 '여자'라는 두 값으로 구성되고, 'death'라는 타겟 클래스 값을 가진 '남자'는 10 명, '여자'는 30 명일 때, 기존의 '성별' 피처의 값이 '남자'라면 '0.1', '여자'라면 '0.3'인 'death'의 확률로 대치된다.

Numeric value 의 변환은 nominal value 의 변환 방법과 유사하나, 특성상 주변의 값도 반영해야 한다.

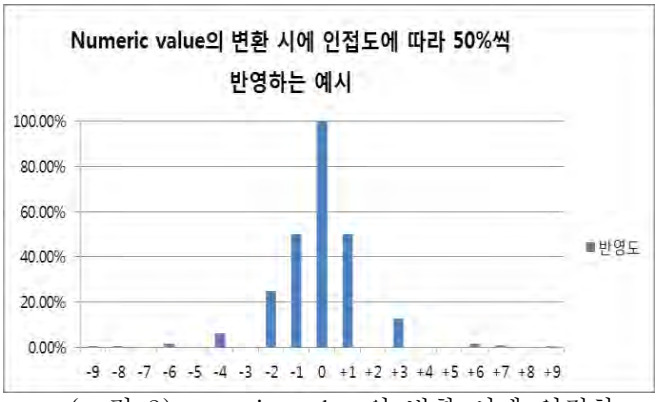
	남자	여자
survival	80	50
death	20	10

변환 전인 피처의 값

	남자	여자
value	0.125	0.0625

변환 후 해당 피처의 값

(그림 1) nominal value 의 변환 예시



(그림 2) numeric value 의 변환 시에 인접한 정도에 따라 50%씩 줄여나가며 반영하는 예시

의학적 분석에서 수치를 구간으로 나누어 판단하는 피처라면, 구간의 경계에 있으나 중심부에 있으나 같은 의미를 지니게 된다. 이대로라면 구간의 경계에 인접한 값은 그 구간을 대표하는 값과 동일하게 판단하지만, 실제로는 이산적으로 나눠서 분포되어 있지 않기 때문에 이 값의 특성을 제대로 반영하지 않을 수 있다. 따라서 특정 numeric value n 의 타겟 클래스가 'death'일 확률을 구한 뒤, 이 값 n 에서 일정 거리 이내의 인접한 값들도 모두 확률을 구해 반영한다. 이 때 (그림 2)와 같이 인접한 값이 어느 정도의 거리 차이가 나느냐에 따라 반영되는 비율이 달라지도록 한다.

결측치에 대해서는 대상 전체의 사망자 비율로 통일하여 변환하는 방법, 피처마다 존재하는 값의 평균으로 변환하는 방법, 피처마다 최빈값으로 변환하는 방법을 적용해서 서로 비교하도록 한다.

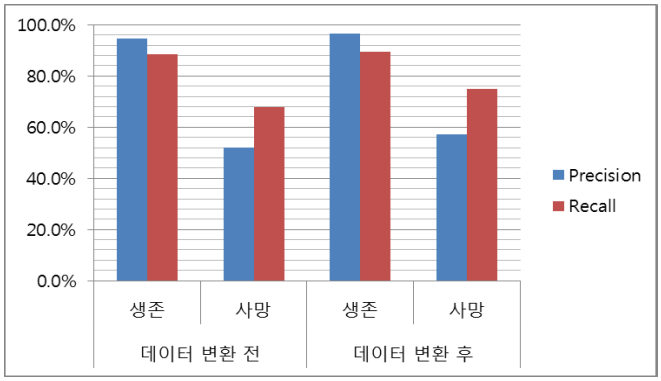
전체 인스턴스 8,608 개를 10-fold cross validation 으로 training data 와 test data 를 9:1 로 나눈 뒤, training data 를 데이터 변환하여 학습 모델을 만든다. Training data 의 데이터 변환 시에 피처마다 변환 전후의 값을 변환 테이블에 저장해두고, test data 의 값을 저장해둔 변환 테이블에서 찾아 대치한다. 만약 변환 테이블에 존재하지 않는 값이라면, training data 의 해당 피처에서 변환한 최빈값으로 바꾼다. 대치한 test data 는 학습 모델을 이용해 'survival', 'death'로 분류, 예측한다.

5. 실험 결과 및 분석

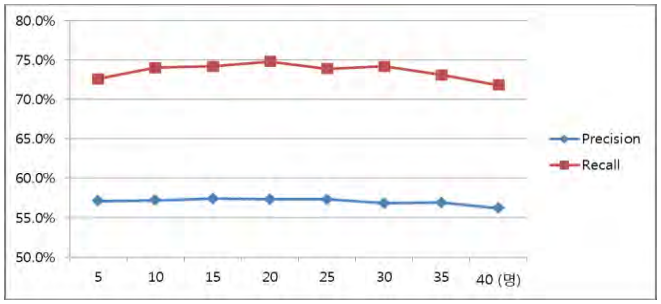
나이브 베이지안 방법에서 실험 데이터 변환 전후의 예측 성능은 <표 3>과 같다. 'Survival'(생존)을 예측하는 경우 precision=96.5%, recall=89.3%이고, 의학적 관점에서 보다 더 중요한 'death'(사망)를 예측하는 성능은 precision=57.3%, recall=74.8%이다.

<표 3> 나이브 베이지안 방법에서 실험 데이터의 변환 전후에 따른 예측의 최고 성능

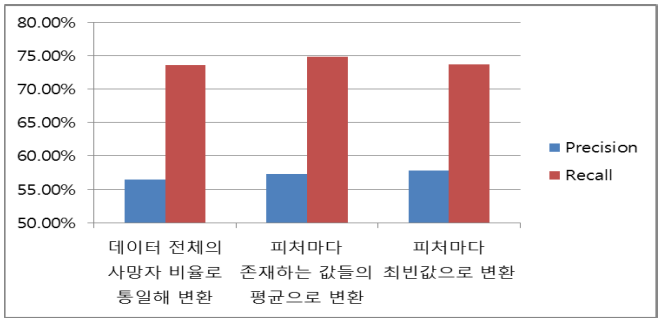
	데이터 변환 전		데이터 변환 후	
	Precision	Recall	Precision	Recall
survival	94.5%	88.4%	**96.5%**	**89.3%**
death	51.9%	67.8%	**57.3%**	**74.8%**



(그림 3) 나이브 베이지안 방법에서 실험 데이터의 변환 전후에 따른 예측 성능 비교



(그림 4) Numeric value 의 인접한 값을 참조한 인원 수 변화에 따른 사망 예측의 성능 비교



(그림 5) 결측치 처리 방법에 따른 사망 예측의 비교

(그림 4)에서 numeric 값의 참조한 인원수는 20 명일 때 사망을 예측하는 성능이 가장 높으나, 큰 영향을 미치는 요소는 아니라고 판단된다. 또한 참조한 값의 인접한 정도에 따른 반영 비율 실험 결과에서는, 50%를 반영한 경우에 대비해 0.1% 미만의 정확도 차이를 보임으로서 거의 영향을 주지 않았다. (그림 5)에서 결측치의 처리 방법에 따른 성능은 거의 비슷하나, 피처마다 존재하는 값만으로 평균을 낸 값으로 대입하였을 때의 death 예측이 상대적으로 precision, recall 이 1% 정도 높았다.

실험 결과를 통해 numeric 값은 20 명을 참조하였을 때, 결측치 값은 각 피처마다 존재하는 값들의 평균을 낸 값으로 대입하였을 때 가장 좋은 성능임을 알 수 있다. 제시한 데이터 변환 방법은 변환 전에 비해 최대 precision=5.4%, recall=7.0%의 성능 향상을 보이고 있어 나이브 베이지안 방법에 적합하다고 판단된다.

6. 결론

한국인 급성 심근경색증 환자의 6 개월 내 사망 여부를 기계학습의 하나인 나이브 베이지안 방법으로 예측하기 위해, 데이터의 복잡성을 단순화시키고 학습 대상에 지향적인 데이터 변환 방법을 제안하였다. KAMIR 데이터를 이용한 실험을 통해, 제안된 방법은 나이브 베이지안 방법의 예측 성능 향상에 기여함을 확인하였다. 이후에는 다른 기계 학습 방법들에도 적용할 수 있도록 연구할 것이다. 또한 다른 의학적 관점에 맞는 타겟 클래스를 대상으로 실험할 것이다.

참고문헌

[1] 만성질환 사망률, 사망원인통계, 국가승인통계 제 10154 호, 통계청, 2013.

[2] S.G. Ellis, W. Weintraub, D. Holmes, R. Shaw, P.C. Block, and etc., “Relation of Operator Volume and Experience to Procedural Outcome of Percutaneous Coronary Revascularization at Hospitals with High Interventional Volumes,” Circulation, 95, pp.2479-2484, 1997.

[3] R.E. Shaw, and etc., “Development of a Risk Adjustment Mortality Model using the American College of Cardiology-National Cardiovascular Data Registry Experience: 1998-2000,” J. Am. Coll. Cardiology, 39, pp.1104-1112, 2002.

[4] Welcome to GRACE, http://www.outcomes-umassmed.org/grace, Center for Outcomes Research.

[5] Tang, Eng Wei, Cheuk-Kit Wong, and Peter Herbison, “Global Registry of Acute Coronary Events (GRACE) hospital discharge risk score accurately predicts long-term mortality post acute coronary syndrome.”, American heart journal, 153(1), pp.29-35, 2007.

[6] 한국인 급성 심근경색증의 현황에 대한 등록 연구, http://kamir3.kamir.or.kr, 한국 심근경색증 연구회

[7] Hyun Kuk Kim, etc., “Hospital Discharge Risk Score System for the Assessment of Clinical Outcomes in Patients With Acute Myocardial Infarction (Korea Acute Myocardial Infarction Registry Score),” American Journal of Cardiology, 107(7), pp.965–971, 2011.

[8] Mair J., and etc., “A Decision Tree for the early Diagnosis of Acute Myocardial Infarction in Nontraumatic Chest Pain Patients at Hospital Admission,” Chest, 108(6), pp.1502–1509, 1995.

[9] Cynthia J. Sims, “Predicting cesarean delivery with decision tree models,” American Journal of Obstetrics & Gynecology, 183(5), pp.1198–1206, 2000.

[10] Gouripeddi, R. K., et al. “Predicting risk of complications following a drug eluting stent procedure: A SVM approach for imbalanced data,” Computer-Based Medical Systems, CBMS 2009, 22nd IEEE International Symposium on IEEE, 2009, pp.1-7, 2009.