심근허혈 심전도 신호의 자동화된 예측을 위한 출현 패턴 마이닝 기반의 분류 방법[1)]

이헌규, 박명호, 류근호
충북대학교 데이터베이스/바이오인포매틱스 연구실
e-mail:{hglee, bluemhp, khryu}@dblab.chungbuk.ac.kr

An Emerging Pattern Mining based Classification Method for Automated Prediction of Myocardial Ischemia ECG Signals

Heon Gyu Lee, Ming Hao Park, Keun Ho Ryu
Database/Bioinformatics laboratory, Chungbuk NationalUniversity

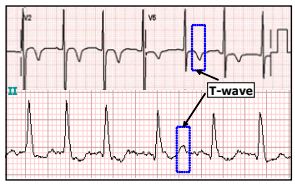
요 약

최근 서구화된 식생활 패턴과 흡연, 비만 등의 원인으로 인해 심근경색, 협심증과 같은 심근허혈(myocardial ischemia) 질환이 급증하고 있다. 이 논문에서는 심전도 신호로부터 허혈성 심장 질환 진단을 위해 출현 패턴 마이닝을 이용하여 심근경색 및 협심증의 진단 신호인 ischemia beat를 분류 하였다. 또한 기존의 출현 패턴 마이닝에 빠른 패턴 탐사와 저장 공간의 효율성을 고려하여 Apriori-T 빈발 패턴 탐사 알고리즘을 출현 패턴 생성이 가능하도록 확장하였다. PhysioNet의 ST-T 데이터베이스로부터 138개의 대조군(정상)과 ischemia beat 데이터에 제안된 분류 알고리즘을 실험한 결과 최소 75% 및 최대 95%의 예측 정확도를 보였다.

1. 서론

최근 심근경색, 협심증과 같은 허혈성 심장 질환에 의한 국내 사망자 수가 급격히 증가함에 따라 심장 질환의 조기 진단 및 진단의 신뢰성은 매우 중요한 문제로 인식되고 있다. 2006년 통계청 조사에 따르면 국내 사망 및 그 원인 통계 결과, 심뇌혈관(심혈관 질환 및 뇌혈관 질환)에 의한 사망자 수는 전체 243,934명의 사망자 중 66,594명으로 27.3%를 차지하며, 이는 암에 의한 사망자에 이어 2위를 차지하며 10년 전에 비해 2배 이상 급증한 것이다[1]. 서구화된 식생활 습관, 흡연, 비만 등이 주요 요인으로 보고되고 있으며 특히 30, 40대의 젊은 연령 층으로 확대되고 있다. 심근 허혈은 심장 근육에 일시적으로 산소가 부족하여 심장이 활동하지 못하는 것이며 허혈의 대표적 증상은 협심증(AP: angina pectoris)이다. 원인으로는 심장의 관상동맥의 일부가 좁아진 혈관 협착이며, 심근경색(myocardial infarction) 및 돌연사 등으로 이어진다. 이러한 심근허혈의 진단 지표로 심전도 신호의 ST-T segments가 사용되는데, segment들의 상승, 하강 또는 역전 여부를 조사함으로서 질환을 진단한다[2]. (그림 1)은 정상인과 전측벽허혈을 가진 환자에서의 역전된 T-wave의 신호의 예이다.

이 논문에서는 심근허혈 질환의 자동화된 진단을 위해서 ST-T segments를 이용한 신뢰성 있는 진단 지표와 정확하고 효율적인 분류 기법 제안을 목표로 한다. 이를 위해서 추출된 ST-T segments 진단 지표를 출현 패턴에 기반한 분류 알고리즘[3], [4]을 확장하여 적용한다. 제안된 전체 수행 단계는 다음과 같다.



(그림 1) 전측벽허혈환자(위)와 정상적인 심전도 신호(아래)의 T-wave 비교

① ST-T segments의 진단 지표 추출 : PhysioNet의 ESC (european society of cardiology) ST-T 데이터베이스의 원시 심전도 신호로부터 5가지의 진단 지표를 추출한다.
② 데이터의 이산화 : 출현 패턴 마이닝 적용을 위해 엔트로피 기반의 이산화를 수행한다.
③ 출현 패턴 마이닝 : 목표 클래스에 대해 높은 발생 빈도를 갖는 출현 패턴들을 효율적으로 발견하기 위해 트리 구조의 출현 패턴 탐사 알고리즘을 제안하며, 질환 진단을 위한 분류 모델에 적용한다.
④ 심근허혈의 분류 : 138명(대조군 49명, 환자 89명)의 심전도 데이터를 이용하여 제안한 질환 분류 모델을 평가한다.

논문의 구성은 다음과 같다. 2장에서는 ST-T segments 진단 지표의 정의 및 추출 과정을 기술하고 3장에서 데이터의 이산화 및 확장된 트리 기반 출현 패턴 마이닝 기법에 대해 설명한다. 제안한 심근허혈 질환 분류 모델의 실

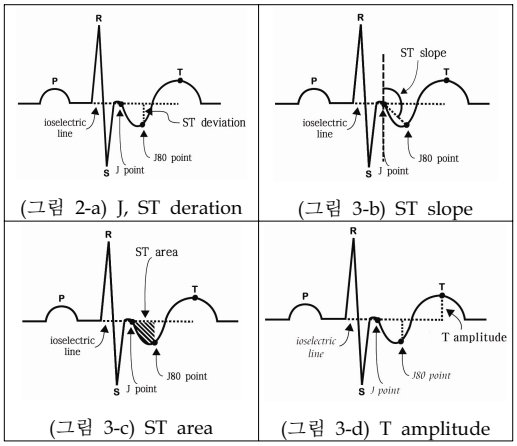
1) 이 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구 (R01-2007-000-10926-0)이며, 한국과학재단에서 지원하는 우수연구센터사업 종양의료 개인특화를 위한 기기·시스템 연구센터(ERC)의 2008년도 연구과제 지원에 의한 결과입니다.

험 및 결과 분석은 4장에 기술하며, 5장에서는 이 논문에 대한 결론을 맺는다.

2. ST-T segments 진단 지표 추출

심전도 신호의 QRS complex는 웨이블릿 변환을 응용하여 검출한 후 ST-segments 진단 지표를 추출한다. QRS complex는 웨이블릿 특성을 이용하여 5~30Hz을 추출하여, 정확히 QRS를 검출한다. QRS complex 검출 후 R-peak를 검출하여 ST segment 시작점인 J point와 J point로부터 80ms 떨어진 지점인 J80 point 추출한다[5]. 심근 허혈의 진단 지표로 추출된 5가지의 ST-T segments는 다음과 같다.

- J point : ST segment의 시작점 J (그림 2-a)
- ST deviation : Isoelectric line (높이 0인 지점)으로부터 J80인 저점의 높이 (그림 2-a)
- ST slope : J 지점과 J80 지점을 연결한 선의 기울기 (그림 2-b)
- ST area : Isoelectric line에서 J 지점과 J80 지점의 면적 (그림 2-c)
- T-wave amplitude : Isoelectirc line에서 T 파의 peak까지의 높이 (그림 2-d)



(그림 2-a) J, ST deration (그림 3-b) ST slope

(그림 3-c) ST area (그림 3-d) T amplitude

(그림 2) ST-T segments 진단 지표

3. 데이터 이산화 및 출현 패턴 마이닝

3.1. ST-T segments의 데이터 이산화

추출된 ST-T segments의 진단 지표들은 실수형 속성 값이다. 따라서 출현 패턴 마이닝 수행을 위해서는 범주형 속성 값을 갖도록 이산화되어야 한다. 이 절에서는 클래스를 고려하고 구간의 순도를 최대화하는 방식으로 분리점을 배치하는 엔트로피 기반의 이산화[6]를 적용한다. k는 클래스의 개수이고 m_i는 분할의 i번째 구간에 속하는 값들의 수, m_{ij}를 구간 i의 클래스 j의 값의 수라고 하면, i번째 구간의 엔트로피 E_i는 다음과 같다.

$$E_i = \sum_{i=1}^{k} p_{ij}\log_2 p_{ij},\quad E = \sum_{i=1}^{n} w_i e_i \qquad \text{식 1, 2}$$

$p_{ij} = m_{ij}/m_i$는 I 번째 구간에서 클래스 j의 확률이며, 분할의 전체 엔트로피 E는 각 구간의 엔트로피의 가중치 평균이 된다. 여기서, m은 값의 수이고, $w_i = m_i/m$은 i번째 구간의 값들의 비율이며, n은 구간의 개수이다. 연속형 속성에 대한 분할은 두 구간이 최소 엔트로피를 가지도록 초기 값들을 이분할 하는 것이다. 또한 분리 과정은 다른 구간에 대해 반복되며, 중단 기준이 만족될 때까지 구간을 선택한다.

3.2. T-tree (total support tree) 구조의 출현 패턴 마이닝 알고리즘

출현 패턴(emerging pattern)이란 성장률(growth-rate)의 적절한 분류 기준을 적용하여 서로 다른 클래스에 해당되는 데이터 집합의 분명한 변화와 차이를 보이는 속성 값들의 조합으로 지지도를 증가시키는 항목집합을 출현 패턴이라고 한다[3], [4]. 일반적으로 연관(association) 분석에서 자주 발생하는 패턴과는 달리 출현 패턴은 높은 구별력(discriminating power)으로 분류 문제에 적용되어 더욱 유용하다고 증명되어 있다. 출현패턴에 대한 문제 정의는 다음과 같다.

- 성장률 (growth rate) : 두 개의 서로 다른 클래스에 해당되는 두 집합 D_1, D_2에 대해, 패턴 X의 D_1에 대한 D_2의 성장률은 다음과 같이 정의된다.

$$GrowthRate(X) = GR(X) = \begin{cases} 0 & If\ sup_1(X)=0\ and\ sup_2(X)=0 \\ \infty & If\ sup_1(X)=0\ and\ sup_2(X)>0 \\ \sup_2/\sup_1 & otherwise \end{cases}$$

여기서, D_1을 배경(background) 데이터 집합, D_2를 목표(target) 데이터 집합이라고 하며, 출현패턴은 배경 데이터로부터 목표 데이터 집합에 대해 높은 성장률을 가지는 패턴을 의미한다. 또한 성장률 임계값 $\rho > 1$에 대해서 패턴 X가 $GrowthRate(X) \gg \rho$의 성장률을 가질 때, 패턴 X를 ρ-Emerging Pattern($\rho - EP$)라 한다.

- JEP (jumping emerging pattern) : 점핑 출현 패턴이란 배경 집합 D_1으로부터 목표 집합 D_2에 대해, 성장률(GR)이 무한대(∞)를 갖는 출현 패턴이다.

출현 패턴 생성을 위해 기존의 트리 구조의 빈발 패턴 탐사 알고리즘인 Apriori-T [7]를 확장하여 출현 패턴을 발견한다. 알고리즘의 확장은 T tree가 지지도만을 계산하는 것에서 출현 패턴 생성을 위해 클래스별 지지도 및 성장률을 계산함으로서 확장된다. 제안한 EP-T (emerging pattern-total support tree) 트리의 구성 과정을 다음과 같다.

- 전체 데이터 집합에 대한 전처리를 통하여 단일 항목 집합들에 대한 지지도를 카운트한 다음, 지지도에 따라서 내림차순으로 정렬한다. 각 데이터 집합의 레코드도 단일 항목들의 지지도에 따라서 내림차순으로 정열하고 지지도를 만족하지 못하는 항목은 레코드에서 삭제한다.
- 전처리를 한 데이터 집합을 이용하여 T-tree를 구성한다. T-tree에는 클래스별 지지도 및 성장률을 저장 한다: <첫번째 클래스 카운트, 두 번째 클래스 카운트, 성장률>

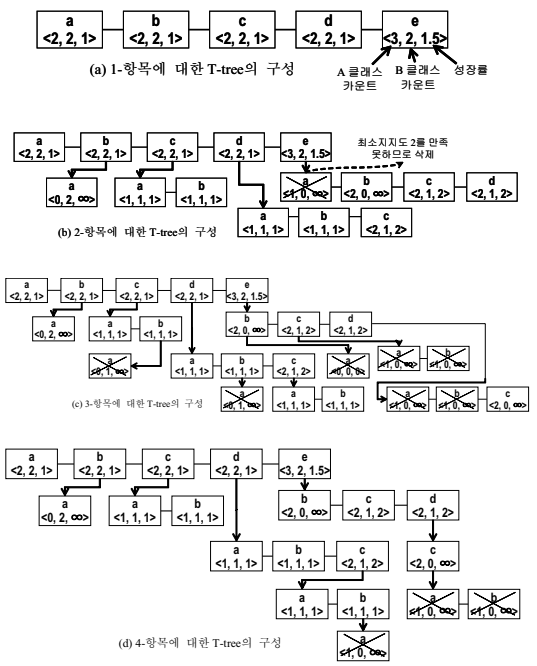
예를 들어, 입력 데이터가 <표 1>과 같을 경우 이를 이용하여 열거 트리 형태인 T-tree를 구성한다. 기존의 T-tree

알고리즘과는 다르게 T-tree에 저장되는 정보는 <A클래스 지지도, B클래스 지지도, 성장률>이다 (최소지지도 카운트는 2로 성장률은 2% 이상으로 가정한다.).

<표 1> 트랜잭션 데이터 집합

TID	Class	Itemset
100	D_1	a c d e
200	D_1	a
300	D_1	b e
400	D_1	b c d e
500	D_2	a b
600	D_2	c e
700	D_2	a b c d
800	D_2	d e

출현 패턴 발견을 위한 확장된 T-tree의 구성은 [7]에서 소개한 T-tree와 동일한 방법으로 진행된다. 그러나 출현 패턴 탐사 과정 중 현재 부분 항목집합이 최소지지도 2를 만족하지 못할 경우 더 이상 그 상위 노드로의 확장은 불필요함으로 성장 하지 못하고 현재 지지도를 만족 못하는 노드를 삭제하게 된다. 처음 1-항목을 가지는 출현 패턴으로부터 최대 4개의 항목을 가지는 출현 패턴 발견을 위한 T-tree의 구성은 (그림 3)에 나타내었다[2].



(그림 3) EP-T 트리의 구성

모든 필수 출현 패턴 생성 후, 새로운 데이터에 대한 분류는 [4]에서 소개된 score를 계산하여 가장 높은 score 값을 가지는 클래스로 분류하게 된다. 분류를 위한 score 계산식은 다음과 같다.

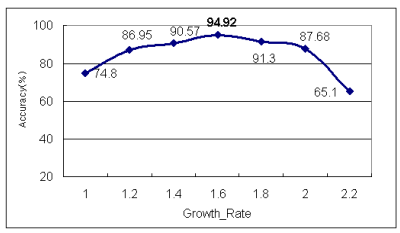
$$score(s, C) = \sum_{e \subseteq s, e \in E(C)} support_c(e) \cdot \frac{growth\ rate(e)}{growth\ rate(e)+1} \quad \text{식 4}$$

여서기 s는 분류될 데이터 인스턴스이고, E(C)는 클래스 C에서 발견된 필수 출현 패턴이다.

5. 실험 및 평가

심근허혈의 예측을 위한 분류 모델 생성과 평가를 위해서 PhysioNet[8]의 European society of cardiology의 ST-T 데이터베이스로부터 대조군 (49명) 및 환자군(89)의 총 138개의 심전도 신호를 사용하였다.

출현 패턴 마이닝은 최소지지도 및 성장률 임계값 파라미터[3]를 가지므로 최적의 임계값 설정을 위해 성장률 변화에 따른 분류 모델의 정확도를 분석하였다. (그림 4)은 최소지지도 1%에서의 성장률 변화에 따른 정확도이다.



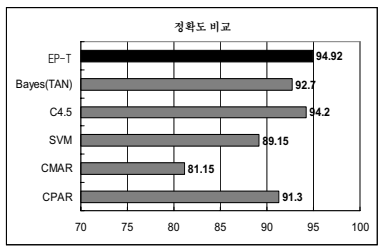
(그림 4) 최소 성장률에 따른 정확도 비교

최적 임계값 설정(최소지지도 1%, 성장률 1.6) 후에, 138건의 데이터 분류 결과에 대한 혼잡 매트릭스는 <표 2>와 같다.

<표 2> 분류 모델의 혼잡 매트릭스

		Predicted class	
		Control	Ischemia
Actual class	Control	49	0
	Ischemia	7	82

제안된 EP-T 알고리즘을 기존의 분류 기법들과의 성능 비교를 수행하였다. [7]에서 제공하는 연관적 분류 기법인 CMAR과 CPAR를 프로그램을 적용하였으며, Java Weka 프로그램[9] 중 베이지안 분류기, C4.5 의사결정트리 그리고 SVM을 각각 비교하였으며 그 결과는 (그림 5)과 같다.



(그림 5) 분류 모델의 정확도 비교

2) 상세한 T-tree 구성 과정은 [7]를 참조

3) 최소지지도는 분류 모델의 정확도에 큰 영향을 미치지 않으므로 낮은 지지도하에서 성장률 변화만을 비교하였음.

6. 결론

이 논문에서는 최근 급증하는 심근허혈 질환의 자동화 되고 정확한 예측을 위해 ST-T segments의 진단 지표를 추출하여 분류 모델을 생성하였다. 또한 기존의 분류 기법보다 더 신뢰성 있는 분류 결과를 위해서 T-tree 기반의 EP-T 출현 패턴 마이닝 알고리즘을 제안하였다. 실험 결과 최소성장률 1.6에서 대조군과 환자군의 분류 정확도는 최대 약 95%이였으며, 기존의 분류 기법들과의 성능 비교에서도 더 정확한 결과를 보여 주었다.

참고문헌

[1] 통계청 인구동향과, "2006년 사망 및 사망원인통계결과," pp. 17-18, 2007.

[2] R. Detrano, D. Mulvihill, K. Lehmann, P. Dubach, A. Colombo, D. McArthur, V. Froelicher, "Exercise-induced ST depression in the diagnosis of coronary artery disease. A meta-analysis," Circulation, American Heart Association, Vol. 80, pp. 87-98, 1989.

[3] G. Dong, J. Li, X. Zhang, "Discovering jumping emerging patterns and experiments on real datasets," 9th Int'l Database Conf. on Heterogeneous and Internet DB, pp. 155-168, 1999.

[4] G. Dong, X. Zhang, L. Wong, J. Li, "Classification by aggregating emerging patterns," Proc. 2nd Int'l Conf. on Discovery Science, pp. 30-42, 1999.

[5] 노기용, 김원식, 이헌규, 이상태, 류근호, "심전도 패턴 판별을 위한 빈발 패턴 베이지안 분류," 정보처리학회논문지 D 제11-D권 제4호, pp. 1-11, 2004.

[6] U. Fayyad, K. Irani, "Multi-Interval discretization of continuous-valued attributes for classification learning," Proc. Int'l Joint Conf.. on AI, pp. 1022 - 1027, 1993.

[7] F. Coenen, LUCS-KDD group, Department of Computer Science, The University of Liverpool, UK, Available : "http://www.cSc.liv.ac.uk/~frans/KDD/," 2004.

[8] European Society of Cardiology, Pisa, Italy, European ST-T database, 1991.

[9] IH. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," San Mateo, CA: Morgan Kaufmann, 1999.