

# 심근허혈 질환 진단을 위한 스트림 데이터 처리<sup>1)</sup>

박진형\*, Khalid E.K.Saeed\*, 이종범\*, 이현규\*\*, 류근호\*

\*충북대학교 데이터베이스/바이오인포매틱스 연구실

\*\*한국전자통신연구원 우정물류기술연구부

e-mail:neozean@dblab.chungbuk.ac.kr

## Stream Data Processing Prototype Development for Automated Prediction of Myocardial Ischemia

Jin Hyoung Park\*, Khalid E.K.Saeed\*, Jong Bum Lee\*, Heon Gyu Lee\*\*,  
Keun Ho Ryu\*

\*Database/Bioinformatics Laboratory, Chungbuk National University

\*\*Electronics and Telecommunications Research Institute, Postal & Logistics  
Research Department

### 요 약

실시간으로 심장 질환의 진단을 위하여 심전도 신호의 스트림 처리 및 데이터 마이닝 프로토타입을 구현하였다. 신체부착형 센서로부터 전송되는 심전도를 전처리하여 심장질환의 진단 지표를 추출하였고 실시간 진단을 위한 출현 패턴 마이닝 알고리즘을 구현 및 적용하였다.

이를 기반으로 심혈관계 질환에 대하여 실시간 자동 진단 및 예측이 가능한 생체 신호 스트림 데이터 처리·분석 프로토타입을 구현하였다.

### 1. 서론

최근 서구화된 식생활 패턴과 흡연, 비만 등의 원인으로 인해 심근경색, 협심증과 같은 심장 질환이 급격하게 증가하고 있다. 지난 10년간 심장질환의 사망률이 인구 10만 명당 1997년 35.6명에서 2007년 43.7명으로 8.1명 증가(2007년도 통계청 조사)하였으며, 2007년 통계청 조사에 따르면 심뇌혈관(심혈관 질환 및 뇌혈관 질환)에 의한 사망자수는 인구 10만 명당 103.3명으로 크게 증가하였으며 전체 244,874명의 사망자중 50,773명(20.7%)으로 암에 의한 사망자에 이어 2위를 차지하고 있다[1]. 지금까지 이러한 심장 질환의 조기 진단을 위한 많은 연구 노력이 진행되었고, 심전도는 과거부터 현재까지 건강의 지표로서 가장 중요한 요소 중의 하나로 여겨지고 있다. 의료기관에서만 실시하던 심전도 측정 장비는 초소형화, 저전력화 기술의 발전으로 인하여 휴대가 가능한 측정 센서 형태로 개발되고 있으며, 센서 네트워크 기술과의 융합을 통하여 실시간으로 환자의 건강관리 및 응급 상황에 대해 대처할 수 있게 되었다. 대표적인 예로 USN(ubiquitous sensor network)의 한 응용분야로 U-헬스케어는 정보통신 기술과 보건의료를 연결하여 언제 어디서나 예방, 진단, 치료, 사후 관리의 보건의료 서비스의 제공에 목적을 두고 서비스를 제공하고 있다. 이러한 목적으로 무선 통신망을 이용한 MobiHeath[2] 건강 서비스가 개발되었다. 이 서비스는 혈

압 또는 생체 신호를 센서에서 계측하여 원격 시스템에 전송하고 이를 모니터링 하여 환자의 위급상황에 대처할 수 있는 서비스를 개발하였으나 모든 환자 데이터에 대한 부하가 집중되는 현상이 발생되고 있다. 이러한 문제점을 극복하기 위하여 자동화된 생체 정보 분석이 가능한 데이터 마이닝 기술이 접목되고 있다. 환자의 심전도 데이터를 수집하여 오프라인 상태에서 분석하고 그 결과를 통해 진단을 내리는 방식의 연구가 이뤄지고 있다[3], [4], [5]. 이 기법들은 정확한 진단율을 갖으나 실제 위급상황일 경우에도 환자의 상태를 기록만 할 수 있는 치명적인 단점을 가진다. 또한, 모바일 장치에서 심전도를 PDA를 통해 모니터링하고 기계학습 알고리즘을 적용하여 환자의 건강 상태를 진단하는 시도[6]를 하고 있으나 하드웨어 성능상의 문제로 인하여 실제 의료에 적용하기 어렵다.

이러한 서비스가 확대되기 위해서는 센서로부터 발생된 심전도 신호 데이터를 실시간으로 처리하는 기술과 수집된 데이터를 짧은 시간에 진단 위한 데이터마이닝 기법이 필요하며, 환자 상태에 따른 맞춤형 자동 진단 및 질환 예측 프로토타입이 요구된다. 이 논문에서 제안한 프로토타입은 실시간으로 환자들의 심전도 데이터를 수집함으로써 응급상황 발생시 최적의 의료 서비스를 제공할 수 있게 된다. 또한 제안한 프로토타입 모델은 실제 응용분야인 U-헬스케어 시스템에 활용 가능하며, 실시간 생체신호 분석 기술을 통해 환자 개인맞춤형 의료진단이 가능해 지고 위급상황에도 빠른 대처를 가능케 해줌으로써, 의료산업 분야 및 의학 분야에 새로운 패러다임을 제시할 수 있다.

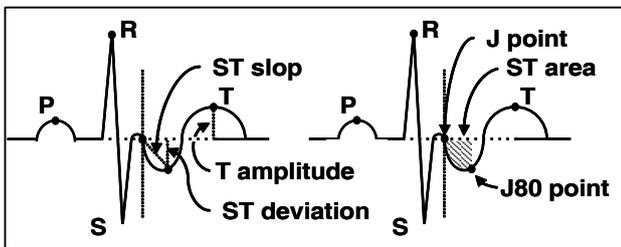
1) 이 논문은 2009년도 정부(과학기술부)의 재원으로 한국과학재단(R01-2007-000-10926-0)과 2009년 교육과학기술부(지역거점연구단육성사업/충북BIT연구중심대학육성사업단)의 지원을 받아 수행된 연구임.

## 2. 데이터 전처리 및 출현 패턴 마이닝

이 장에서는 심전도 파형으로부터 진단 지표를 추출하고, 데이터 전처리 과정을 통해 특정 환자군에서 발생하는 출현 패턴을 찾는 과정에 대해 기술한다.

### 2.1. 심장 질환 진단 지표 추출

심장 질환의 진단 지표로 심전도 신호의 ST-segment가 사용되는데, segment의 상승 또는 하강 여부를 조사함으로써 질환을 진단한다. ST-segments 진단 지표를 추출하기 위하여 심전도 신호의 QRS complex에 웨이블릿 변환을 응용하여 검출한다. QRS complex는 웨이블릿 특성을 이용하여 5~30Hz를 추출하여, QRS를 검출한다. QRS complex 검출 후 R-peak를 검출하여 ST segment 시작점인 J point와 J point로부터 80ms 떨어진 지점인 J80 point를 추출한다. 진단 지표로 추출된 5가지의 ST-T segments는 (그림 1)과 다음과 같다.



(그림 1) ST-segment 진단 지표 (ST slope, ST deviation, T amplitude, J point, J80 point, ST Area)

추가적으로, 심장 질환의 진단 지표로써, 수집된 ECG로부터 주파수 및 시간 영역의 심박동변이도 진단 지표를 추출한다(그림 2).

### 2.2. ST-T segments의 데이터 이산화

ST-segments 및 HRV의 모든 진단 지표들은 연속형 속성 값이다. 따라서 목표 클래스에 대한 출현 패턴 마이닝 수행을 위해서는 범주형 속성 값을 갖도록 이산화 되어야 한다. 이 절에서는 클래스를 고려하고 구간의 순도를 최대화하는 방식으로 분리점을 배치하는 엔트로피 기반의 이산화[7]를 적용한다.

$k$ 는 클래스의 개수이고  $m_i$ 는 분할의  $i$ 번째 구간에 속하는 값들의 수,  $m_{ij}$ 를 구간  $i$ 의 클래스  $j$ 의 값의 수라고 하면,  $i$ 번째 구간의 엔트로피  $E_i$ 는 다음과 같다.

$$e_i = \sum_{j=1}^k p_{ij} \log_2 p_{ij} \quad (\text{식 1})$$

SDRR [ms] : RR 간격의 표준편차	
SDSD [ms] : 인접한 RR 간격사이의 차이값의 표준편차	
RRm [ms] : 모든 RR 간격의 평균	
TP (total power) [ms <sup>2</sup> ] : 0.4 Hz이하의 RR간격들의 주파수 영역에서의 총 파워	
VLF (very low frequency) [ms <sup>2</sup> ] : 0.04 Hz 이하에서의 파워	
LF(low frequency) [ms <sup>2</sup> ] : 0.04~0.15 Hz 에서의 파워	
nLF [nu] : LF의 정규격화 값	
$(nLF = \frac{LF - VLF}{TP} \times 100)$	
HF [ms <sup>2</sup> ] : 0.15 - 0.4 Hz 에서의 파워	
nHF [nu] : HF의 정규격화 값	
$(nHF = \frac{HF - VLF}{TP} \times 100)$	
LF/HF : LF/HF 비율	

(그림 2) 선형적 비선형적 특징의 진단 지표

$p_{ij} = m_{ij}/m_i$ 는  $i$ 번째 구간에서 클래스  $j$ 의 확률이며, 분할의 전체 엔트로피  $E$ 는 각 구간의 엔트로피의 가중치 평균이 된다.

$$E = \sum_{i=1}^n w_i e_i \quad (\text{식 2})$$

여기서,  $m$ 은 값의 수이고,  $w_i = m_i/m$ 은  $i$ 번째 구간의 값들의 비율이며,  $n$ 은 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.

### 2.3. 자동화된 예측을 위한 출현 패턴 마이닝

출현 패턴(emerging pattern)이란 성장률(growth-rate)의 분류 기준을 적용하여 특정 데이터 집합에서 높은 지지도를 갖고 다른 데이터 집합과 분명한 변화와 차이를 보이는 속성 값들의 조합으로 된 항목집합을 출현 패턴이라고 한다[8][9]. 즉, 출현 패턴은 두 개 이상의 분할된 데이터 집합을 명확하게 구분해 주는 패턴을 말하는 것으로, 이러한 패턴들은 하나의 데이터 집합에서 다른 클래스를 갖는 데이터 집합 사이에 명확한 차별점을 가진다. 일반적으로 연관(association) 분석에서 자주 발생하는 패턴과는 달리 출현 패턴은 높은 구별력(discriminating power)으로 분류 문제에 적용되어 더욱 유용하다고 증명되어 있다. 출현 패턴에 대한 문제 정의는 다음과 같다.

**[정의 1]** 성장률 (growth rate) : 두 개의 서로 다른 클래스에 해당되는 두 집합  $D_1, D_2$ 에 대해, 패턴  $X$ 의  $D_1$ 에 대한  $D_2$ 의 성장률은 다음과 같이 정의된다.

$$GrowthRate(X) = \begin{cases} 0 & \text{If } sup_1 = 0 \ \& \ sup_2 = 0 \\ \infty & \text{If } sup_1 = 0 \ \& \ sup_2 \neq 0 \\ sup_2 / sup_1 & \text{otherwise} \end{cases} \quad (\text{식 3})$$

여기서,  $D_1$ 을 배경(background) 데이터 집합,  $D_2$ 를 목표(target) 데이터 집합이라고 하며, 출현 패턴은 배경 데이터로부터 목표 데이터 집합에 대해 높은 성장률을 가지는 패턴을 의미한다. 또한 성장률 임계값  $\rho > 1$ 에 대해서 패턴  $X$ 가  $GrowthRate(X) \gg \rho$ 의 성장률을 가질 때, 패턴  $X$ 를  $\rho$ -Emerging Pattern( $\rho$ -EP)라 한다.

[정의 2] 출현 패턴  $X$ 의 강도(strength)는 (식 4)와 같다.

$$strength(X) = \frac{GR(X)}{GR(X+1)} \cdot sup(X) \quad (식 4)$$

[정의 3] JEP (jumping emerging pattern) : 점핑 출현 패턴이란 배경 집합  $D_1$ 로 부터 목표 집합  $D_2$ 에 대해, 성장률(GR)이 무한대( $\infty$ )를 갖는 출현 패턴이다. 따라서 점핑 출현 패턴은  $strength(X) = sup(X)$ 인 특별한 형태의 출현 패턴이다.

예를 들어 임계값인 최소성장률(minimum growth rate)  $\rho = 2$ 라 하고, <표 4>의 복합 진단 지표에 대한 환자 및 정상인 두 클래스에 해당되는 출현 패턴은 다음과 같다.

- $X_1$ :  $\{J=(-0.52865\sim 0.47055), J80=(0.0662\sim 0.28495), SDRR)=(42.525379\sim \infty)\}$ 의 지지도가 환자군에 대해 2/9, 정상인군에 대해 3/5를 가질 경우, 성장률  $GR(X_1)=2.7$  이므로  $X_1$ 은 정상인군에 대한 출현 패턴이다.
- $X_2$ :  $\{J=(0.47055\sim \infty), Area=(23.9315\sim \infty)\}$ 은 환자군에서 단 한번 발생되고(1/9), 정상인군에서 0의 지지도를 갖는다면, 성장률  $GR(X_2)=\infty$ 이므로 점핑 출현 패턴이 된다.

그러나  $X_2$ 의 경우, 패턴의 길이가 길고 상세한 패턴이므로 목표 데이터 집합인 환자군에서도 한번 발생되므로 분류 패턴으로는 적합하지 않는 noise 패턴이다. 따라서 이러한 noise 패턴을 제거하기 위해서 연관규칙에서의 최소지지도(minimum support) 임계값을 적용한다.

[정의 4] score : 모든 필수 출현 패턴 생성한 후에 새로운 데이터에 대한 분류는 [9]에서 소개된 score를 계산하여 가장 높은 score 값을 가지는 클래스로 분류하게 된다. 분류를 위한 score 계산식은 (식 5)와 같다.

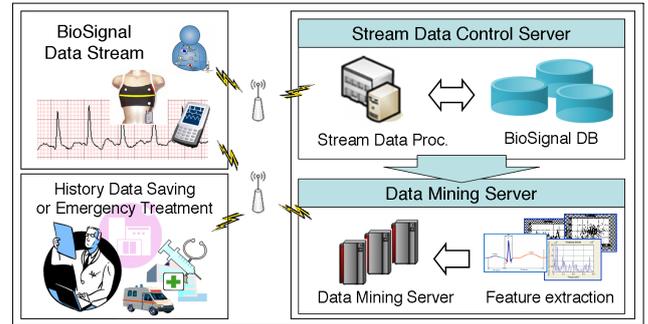
$$score(s, C) = \sum_{e \in s, e \in E(C)} support_c(e) \cdot \frac{growth\ rate(e)}{growth\ rate(e)+1} \quad (식 5)$$

여기서  $s$ 는 분류될 데이터 인스턴스이고,  $E(C)$ 는 클래스  $C$ 에서 발견된 필수 출현 패턴이다.

### 3. 스트림 데이터 처리 프로토타입

이 장에서는 연속적인 생체신호의 실시간 스트림 데이터 처리와 출현 패턴 마이닝기반의 생체 신호 분석을 통

한 자동 진단 프로토타입에 대해 설명한다. 구현된 프로토타입의 개념도는 (그림 3)과 같다.



(그림 3) 자동 진단 프로토타입 개념도

센서에서 측정된 데이터가 중계기를 거쳐 TCP/IP 형태로 서버로 전송된다고 할 때, 구현한 프로토타입은 (1) 센서 네트워크 환경에서 생체신호 스트림 데이터 수집하고 처리할 수 있는 모듈과 (2)원시 생체신호(ECG)로부터 심혈관계 질환의 진단에 필요한 복합 진단 지표 추출 모듈, (3)출현 패턴을 이용하여 생체 정보로부터 심혈관계 질환을 진단 및 예측할 수 있는 생체정보 데이터마이닝 모듈로 구성된다.

스트림 데이터 처리 모듈은 센서로부터 동시 다발적으로 수집되는 데이터를 실시간으로 데이터베이스에 삽입시킴으로서 시스템의 부하를 줄이는 역할을 수행한다.

복합 진단 지표 추출 모듈은 수집된 원시 신호로부터 진단 예측에 필요한 특성들을 추출하고 훈련데이터로부터 생성된 엔트로피 이산화 구간 스키마를 이용하여 추출된 진단 지표들을 이산화 한다.

마지막으로 생체정보 데이터마이닝 모듈은 트리 구조의 빈발패턴 탐사 알고리즘인 Apriori-T를 확장한 EP-T (Emerging Pattern-Total support tree) 알고리즘을 적용하였다[10]. 분류를 위해 의료기관에서 검증된 생체 신호 데이터(Training Set)로부터 추출된 출현 패턴을 기반으로 분류 모델을 생성한 후에 센서로부터 수집된 데이터(Test Set)를 분류한다. 분류된 결과를 환자나 의사, 응급센터 기관에 알림으로써 심혈관 질환에 의한 사망 사고를 예방하도록 한다.

### 4. 시스템 구현

(그림 5)는 제안된 프로토타입의 구현화면이다. 구현 환경은 CPU DualCore Intel Core 2 Duo E6600 2400MHz, RAM 3Gbyte, 운영체제 Microsoft Windows XP Professional를 사용하였다. 사용한 프로그래밍 언어는 JAVA 1.6.을 사용하였고, 사용한 개발 도구는 Eclipse 2.3 이다. 구현에 사용한 데이터는 피조넷(Physo Net)[11] 84 건의 데이터를 사용하였다.

(그림 4-1)은 데이터 베이스 서버 접속과 통신 서버를

구동시키는 화면이다. (그림 4-2)는 서버가 정상 작동시 다수의 센서로부터 데이터를 수집하고 있는 화면이다. (그림 4-2)에서 보는 바와 같이 스트림 데이터 관리 서버의 구현 화면은 총 3가지 부분으로 구성된다. 왼쪽 부분에는 현재 접속한 센서ID(환자ID) 리스트를 보여주고 있으며, 오른쪽 위 부분은 왼쪽 센서리스트에서 특정 항목을 클릭하였을 경우 접속된 센서를 사용하는 사용자의 간단한 정보와 수집되는 데이터의 수치를 볼 수 있다. 오른쪽 아래 부분은 접속과 접속 해지에 대한 로그 정보를 볼 수 있다.



(그림 4-1) 초기 설정 화면



(그림 4-2) 동작 화면

(그림 5)는 수집된 생체 신호 데이터에서 필요한 진단 지표들을 추출하고, 추출된 지표를 바탕으로 분류를 하는 동작 화면이다. 왼쪽부분은 현재 접속된 센서의 목록을 보여주며, 가운데부분은 추출된 지표 정보를 보여준다. 오른쪽 위부분은 진단 예측에 관련된 데이터 마이닝 로그 정보를 볼 수 있으며, 오른쪽 아래부분은 응급 상황 발생시 응급 상황에 대한 정보를 보여준다.

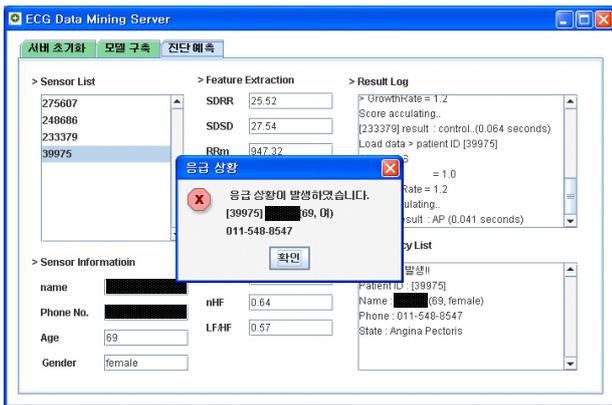


그림 5 생체 신호 데이터 마이닝 서버 동작 화면 (응급상황 발생시)

### 5. 결론

이 프로토타입의 구현을 통하여 실시간으로 발생하는 스트림 데이터 처리 및 질의 처리, 실시간 데이터마이닝 기술을 정립하였다. 이러한 프로토타입 모델은 실제 응용 분야인 U-헬스케어 시스템에 활용 가능하다. 즉, 심근경색, 뇌졸중과 같은 심혈관계 질환자를 위한 주기적·비주

기적 점검 및 의료 정보 전송 시스템에 적용이 가능하다.

또한, 실시간 생체신호 분석 기술을 통해 환자 개인맞춤형 의료진단이 가능해 지고, 위급상황에도 빠른 대처를 가능케 해줌으로써, 의료산업 분야 및 의학 분야에 새로운 패러다임을 제시할 수 있다.

### 참고문헌

- [1] 통계청 인구동향과, "2007년 사망 및 사망원인통계 결과," pp. 8-18, 2008.
- [2] MobiHealth Project, <http://www.mobihealth.org>
- [3] 노기용, 김원식, 이현규, 이상태, 류근호, "심전도 패턴 판별을 위한 빈발 패턴 베이지안 분류," 정보처리학회논문지 D 제11-D권 제4호, pp. 1-11, 2004.
- [4] Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, Hammond WE. "Medical data mining: knowledge discovery in a clinical data warehouse," Proceeding of AMIA annual Fall Symp., Vol.101, No.5, 1997.
- [5] Jiuyong Li, Ada Wai-chee Fu, Hongxing He, Jie Chen, Huidong Jin, Damien McAullay, Graham Williams, Ross Sparks, Chris Kelman, "Mining risk patterns in medical data," Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 770-775. 2005.
- [6] Jimena Rodriguez, Alfredo Goñi, Arantza Illarramendi, "Real-time classification of ECGs on a PDA," IEEE Transactions on Information Technology in Biomedicine, Vol.9, No.1, pp. 23-34, 2005.
- [7] U. Fayyad, K. Irani, "Multi-Interval discretization of continuous-valued attributes for classification learning," Proc. Int'l Joint Conference. on AI, pp. 1022 - 1027, 1993.
- [8] G. Dong, J. Li, X. Zhang, "Discovering jumping emerging patterns and experiments on real datasets," Proc. 9th Int'l Database Conf. on Heterogeneous and Internet Databases, pp. 155-168, 1999.
- [9] 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.
- [10] 이현규, 박명호, 류근호, "심근허혈 심전도 신호의 자동화된 예측을 위한 출현 패턴 마이닝 기반의 분류 방법," 한국정보처리학회 추계학술발표대회 논문집, Vol.15, No. 2, 2008, pp. 19-22.
- [11] <http://www.physionet.org/>