AptaCDSS – 압타머칩을 이용한 심혈관질환 질환단계 예측 및 진단의사결정지원시스템

엄재홍*0 김병희* 이제근* 허민오* 박영진* 김민혁* 김성천** 장병탁*

*서울대학교 컴퓨터공학부 바이오지능연구실 **제노프라(주)

*{jheom0, bhkim, jklee, moheo, yjpark, mhkim, btzhang}@bi.snu.ac.kr **kimgp@cotech.co.kr

AptaCDSS – A Cardiovascular Disease Level Prediction and Clinical Decision Support System using Aptamer Biochip

Jae-Hong Eom*0, Byoung-Hee Kim*, Je-Keun Rhee*, Min-Oh Heo*, Young-Jin Park*, Min-Hyeok Kim*, Sung-Chun Kim**, Byoung-Tak Zhang*

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

**Genoprot Co., Ltd., 2FL Saeseoul B/D 94-1, Guro 6-dong, Guro-gu, Seoul 152-841, Korea

요 약

최근 연구결과에 의하면 심장질환을 포함한 심혈관질환은 성별에 관계없이 미국 및 전 세계적으로 질병사망의 주요 요인으로 조사되었다. 본 연구에서는 심혈관질환은 보다 효율적으로 진단하기 위해 개발된 진단의사결정 보조시스템에 대해서 다룬다. 개발된 시스템은 혈청 내의 특정 단백질의 상대적 양을 측정할 수 있는 바이오칩인 압타머칩을 이용해 생성한 환자들의 칩 데이터를 Support Vector Machine, Neural Network, Decision Tree, Bayesian Network 등의 총 4가지 기계학습 알고리즘으로 분석하여 질환단계를 예측하고 진단을 위한 보조정보를 제공한다. 논문에서는 총 135개 샘플로 구성된 3K 압타머칩 데이터에 대해 측정된 초기 시스템의 질환단계 분류성능을 제시하고 보다 유용한 진단의사결정 보조 시스템을 구성하기 위한 요소들에 대해서 논의한다.

1. 서론

1.1. 연구의 배경

최근 연구결과에 의하면 심장질환을 포함한 심혈관질환(Cardiovascular Disease; CVD)은 성별에 관계없이 미국 및 전 세계적으로 질병사망의 주요 요인으로 보고되고 있으며 미국의 경우 최근 연간 질병 사망자의 약 40%가 CVD와 연관되어있다고 보고되었다[1]. 일반적으로 CVD가 약 65세 이상의 고령자들에게서 보편적이었던 과거와는 달리 최근에는 생활습관 및 식습관 변화에 따른 영양상태 변호로 약 15세에서 34세이 이르는 젊은 세대에서도 CVD와 관련된 질환들이 보고되고 있다[2].

심혈관질환은 심부전, 고혈압성 심장질환, 부정맥, 판막질환, 선천성 심장질환, 심근증, 심낭질환과 같은 심장질환과 뇌졸중, 말초혈관질환, 동맥류 등의 혈관질환을 포함하는 질병으로 심장질환들 중에서 중요한 부분을 차지하는 관상동맥 질환은 대개 동맥경화에 의해 심장에 혈액을 공급하는 관상동맥이 막히거나, 좁아져 발생하는 것으로 심근경색증이나 협심증이 이에 해당한다. 관상동맥질환은 우리나라에서도 지난 약 30여 년 동안 급격히 증가하여왔는데 급속한 경제 발전에 따른 식이의 변화와 주로 관련이 있을 것으로 생각되고 있다. 현재 심혈관질환은 암 다음으로 높은 사망원인으로 기록되고 있어 질환에 대한 초기진단은 질환치료에 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.

최근에는 Aptamer를 이용한 압타머칩 기술이 생체시료의 발현 측정에 활용되기 시작했다. 압타머칩은 혈액 내의 혈청(serum) 에 포함된 특정 단백질의 상대적 양을 직접 측정할 수 있는 바이오칩으로 질병진단 등의 의학적 응용에 활용될 수도

있으며 기존의 마이크로어레이 분석기법을 그대로 적용할 수 있다는 장점을 가진다. 압타머를 이용해 혈액 내의 특정 단백질 양을 측정하는 압타머 바이오칩은 유전자를 이용한 기존의 마이크로어레이 분석방법과는 달리 생리적 작용에 보다 가까운 위치에서 영향을 미치는 '단백질'을 측정함으로써 보다 정확한 인과관계 추정이 가능할 것으로 여겨지고 있다.

더욱이, 최근 들어 질병 진단 방법에 있어서 기존의 임상 진단 방법에만 의존하던 것과는 달리 CDSS (Clinical Decision Support System)와 같은 전문 진단환경 시스템을 이용하는 등의 다양한 방향으로 변화가 이루어지고 있다. 이에 본 논문에서는 압타머 바이오칩 데이터 및 이러한 진단환경 시스템에서 활용 가능한 4가지 기계학습 기법들을 이용하여 심혈관 질환의 질환단계를 예측하고 진단보조를 위한 시스템을 구축하였다.

1.2. 관련 연구

전체 질환에서 심혈관 질환이 차지하는 비중이 상대적으로 높기 때문에, 초기 진단의 중요성이 대두되어 심혈관질환 진단을 위한 다양한 연구들이 진행되어왔다.

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.

CDSS에 관련된 연구는 본 논문에서 언급하기 힘들 정도로 많은 연구가 진행되고 있다. 최근의 연구 결과 중 본 논문의 내용과 연관된 일부 연구로 Decision Tree를 이용한 잠재적 바이오마커탐색 (potential biomarker finding) [5][6], Bayesian Network을 이용한 MDSS (Medical Decision Support System) 개발 [7], 다양한 종류의 심장병(heart diseases) 진단을 위한 Neural Network을 이용한 MDSS 개발 [8], Support Vector Machine을 이용한 바이오마커탐색[9] 등을 들 수 있다.

또한, AptaCDSS와 관련해서 CVD 질환단계를 예측하기 위한 사전 시도로 본 저자들에 의한 SVM 앙상블 방법[10]과 기계학습 기반 방법[11] 등의 사전연구가 진행되었다.

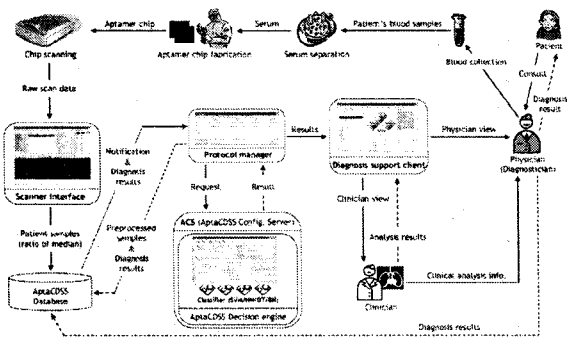

그림 1. AptaCDSS 시스템 흐름도

2. AptaCDSS

2.1. 시스템 구성도

그림 1은 Aptamer 바이오칩을 이용한 실혈관질환의 질환단계 예측 및 진단보조 시스템의 활용을 보여준다. AptaCDSS 시스템은 Windows 기반으로 구성되었으며 스캐너인터페이스, AptaCDSS 데이터베이스, AptaCDSS 통신서버, ACS, 진단클라이언트의 약 5개의 주요 요소로 구성된다.

2.2. 시스템 구성요소

스캐너인터페이스 — 스캐너인터페이스는 그림 2와 같이 구성된다. 스캐너인터페이스는 압타머 바이오칩을 이용하여 환자의 혈액내의 혈청 내의 단백질을 측정한 칩 데이터를 읽어 생성된 원본 데이터 파일을 입력으로 받아 원본파일의 특정 필드를 이용하여 환자별 데이터집합을 구성한다. 이렇게 구성된 환자별 데이터는 AptaCDSS DB에 저장되고 이후 분석에 활용된다.

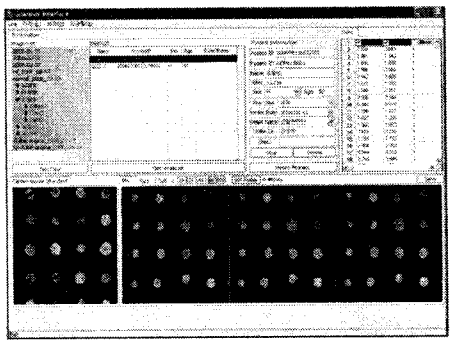

그림 2. 스캐너인터페이스

AptaCDSS 데이터베이스 — AptaCDSS가사용하는 모든 정보를 담고 있으며 환자의 압타머칩 데이터 및 전 처리된 결과 데이터, 기계학습 모델 데이터 및 진단 결과, 각종 로그를 저장하며 Oracle 10g가 사용되었다.

AptaCDSS 통신서버 — AptaCDSS의 각 구성요소별 통신을 처리하는 시스템으로 HL7에 준하는 규격을 이용하여 통신을 처리한다. AptaCDSS DB에 새로운 데이터가 저장되었거나 신규 진단요청 데이터가 올라온 경우 ACS에 진단을 요청한다.

ACS (AptaCDSS Configuration Server) — 4가지 기계학습 알고리즘 모델을 생성/학습/평가하기 위한 인터페이스를 제공하고 데이터의 전처리 및 속성 선택을 지원하여 효율적 모델 학습 및 학습된 모델을 이용한 신규 환자 데이터에 대한 진단을 수행한다.

진단클라이언트 — 임상의 및 진단의사에게 환자에 대한 시스템의 진단 결과를 제공하고 진단에 참고가 되는 보조 정보를 제공한다. 임상의사의 경우 칩에 사용된 전체 3,000개의 단백질 중에서 선정된 중요 단백질 정보 및 진단 결과를 이용하여 진단에 중요 정보를 제공하는 바이오마커 탐색에 활용할 수 있다. 진단의사의 경우 진단의사의 진단지식 이외에 칩 데이터

분석을 통한 시스템 예측 결과를 활용하여 진단이 쉽지 않은 환자의 진단에 대한 보조정보를 제공받을 수 있다.

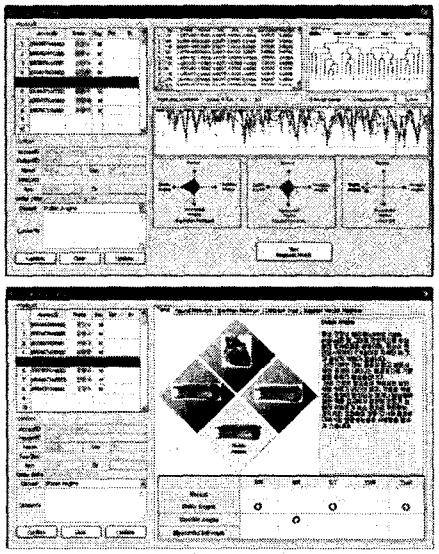

그림 3. 진단클라이언트. 임상의용 인터페이스 (상) 및 진단의용 인터페이스 (하)

3. AptaCDSS의 CVD 질환단계 분류기

AptaCDSS는 아래와 같은 4가지 기계학습 분류기를 이용하여 CVD 환자의 질환단계를 예측한다.

3.1. Support Vector Machine

Support Vector Machine (SVM)은 고차원공간으로의 사상(mapping)을 통해 데이터 샘플을 분류하는 최적 hyperplane을 탐색하는 알고리즘으로 다양한 응용에서 좋은 성능을 보이어 시스템 구현에 사용하였다. 총 4단계의 CVD 질환단계 예측을 위해 2진 분류기 여러 개를 사용하여 다중분류 문제를 처리하도록 구성하였으며 사용자로부터 사용할 커널 타입 및 기타 파라미터들을 입력받도록 구성하였다.

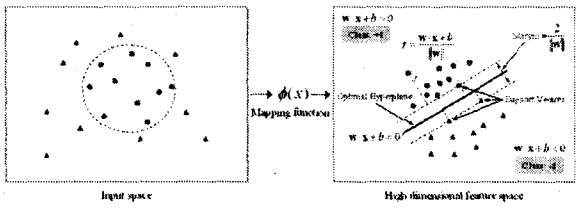

그림 4. 고차원 공간으로의 사상을 통한 SVM의 데이터 분류

3.2. Neural Network

Neural Network (NN)은 CDSS 및 MDSS 구성에 가장 널리 사용되는 기계학습 방법 중의 하나로 AptaCDSS에서는 NN 분류기로 3-layered MLP를 사용하였다. 시스템의 NN 모듈은 전처리 과정을 통해 축소된 각 샘플의 차원(단백질 수)에 해당하는 입력 노드와 사용자가 지정한 개수의 은닉노드, 학습률, momentum, epoch를 입력 받는다. 출력 노드는 CVD 구분단계 각각을 표현하도록 구성된다.

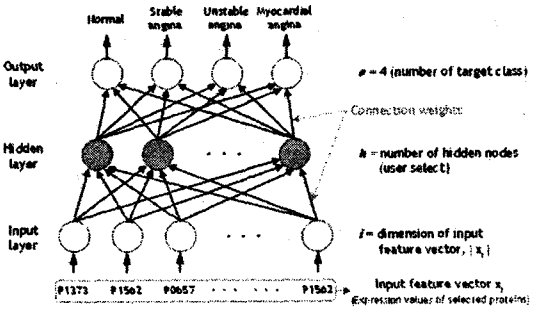

그림 5. AptaCDSS의 3 layered MLP 구조 예시

3.3. Decision Tree

Decision Tree (DT)는 학습결과를 사람이 읽기 쉬운 규칙으로 구성할 수 있어 진단 보조정보 생성을 위해 AptaCDSS에 포함되었다.

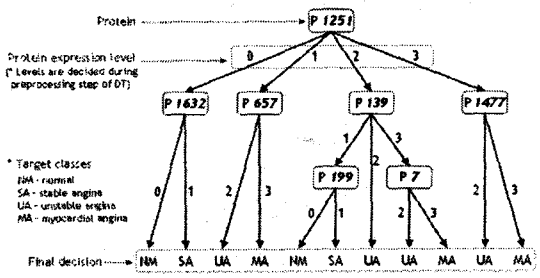

그림 6. CVD 분류를 위한 Decision Tree의 예 (말단 노드는 CVD 분류의 4가지 클래스중 하나를 의미)

3.4. Bayesian Network

Bayesian Network (BN)은 DT와 마찬가지로 압타머 바이오칩을 이용한 CVD 질환단계 예측에 있어 중요하게 고려되는 단백질들(바이오마커) 간의 연관성을 보다 직관적으로 표현할 수 있어 AptaCDSS 시스템에 포함하였으며 모델학습을 통해 얻어진 네트워크 정보를 진단가시화에 활용할 수 있도록 하였다.

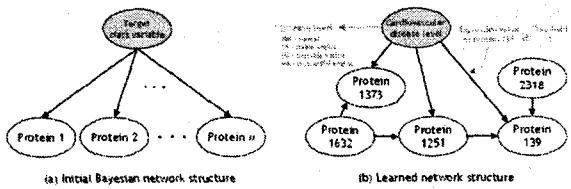

그림 7. CVD 분류를 위한 Naive Bayes 형태의 BN 기본 모델(a)과 데이터 학습 후 주요 단백질(압타머)의 상호작용을 표현하는 Bayesian Network의 예

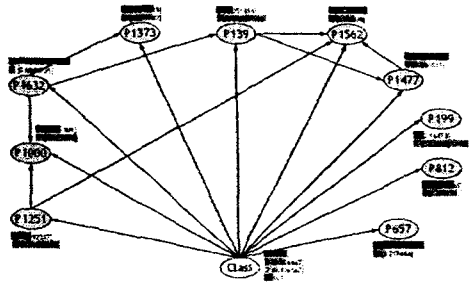

그림 8. 선정된 10개 단백질을 이용한 주요 단백질 (압타머)의 상호작용 네트워크의 예

4. 시스템 성능분석

4.1. 데이터 및 전처리

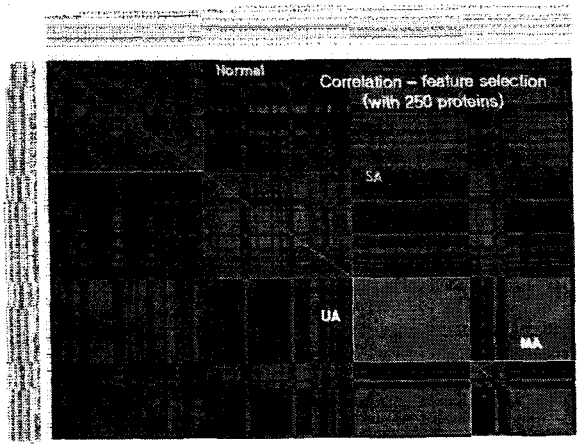

그림 9. 135 샘플 데이터의 클래스간 상관관계 (3K, 135 샘플, 3,000 → 250 속성선택 데이터 사용)

논문에서는 캐제노프라의 135샘플 3K 압타머 칩 데이터를 사용하였다. 아래의 표 1은 논문에서 심혈관 질환을 분류에 사용된 전체 데이터에 대한 정보를 나타낸다.

샘플별로는 칩 스캔 데이터 중에서 median 값을 칩 데이터로 사용하였고 분산분석(ANOVA)을 이용하여 각 단백질 별로 p-value 측정 한 후 샘플별로 전체 3,000개의 단백질 들 중에서 p-value가 낮은 순으로 상위 250개의 단백질들만을 선택하는 방식으로 전처리를 수행하였다. 그림 9는 이렇게 전처리 후 데이터의 각 클래스별 상관관계를 나타낸다.

표 1. 실험에 사용한 데이터의 각 질환 단계별 샘플 정보

데이터베이스	데이터 개수	전체 단백질 수 (사용된 수)
정상	40	3,000 (250)
질환 1기	38	3,000 (250)
질환 2기	29	3,000 (250)
질환 3기	28	3,000 (250)
계	135	3,000/샘플

4.2. 성능평가 결과

표 2는 AptaCDSS 시스템에 포함되어있는 각 기계학습 분류기들의 CVD에 대한 분류성능을 나타내며 각 성능은 10-fold cross validation으로 측정되었다.

표 2. AptaCDSS 시스템 분류기들의 심혈관질환 질환단계 예측 결과

분류기	분류 성능
Support Vector Machine	79.10 %
Neural Network	85.07 %
Decision Tree	70.00 %
Bayesian Network	75.12 %

표 2에서와 같이 심혈관질환의 질환단계 예측 문제에 있어서 AptaCDSS의 각 알고리즘은 전반적으로 약 70% ~ 85%의 성능을 보이고 있다. 그렇지만 현재의 분류 성능 결과는 총 데이터가 135개뿐인 상태에서 분류를 수행한 결과이기 때문에 이 결과가 충분한 규모의 데이터에 대한 일반화된 분류성능을 보여준다고 할 수 없는 상황이다. 때문에 통제된 환경에서 생산된 보다 충분한 수의 데이터 확보를 위한 노력이 필요할 것으로 사료된다.

5. 결론 및 향후 과제

5.1. 결론

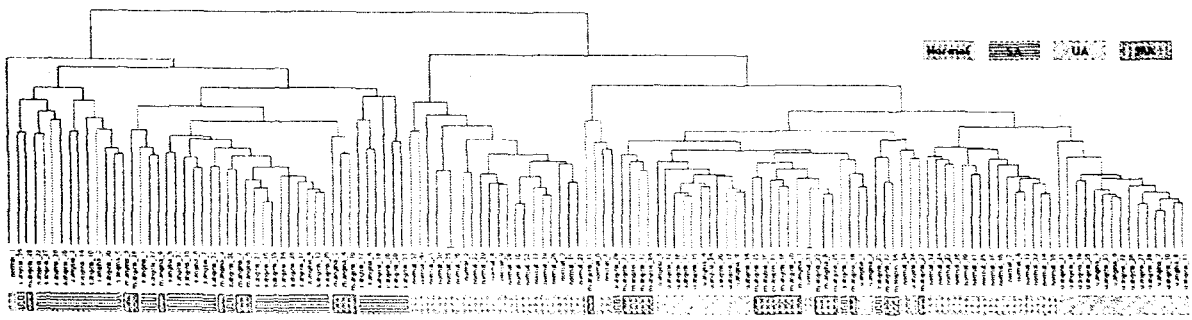

그림 10. 135 샘플 데이터의 계층적군집화 결과 (Hierarchical clustering with 'average linkage', 3K, 135 샘플, 250단백질 사용)

논문에서는 4가지 기계학습 알고리즘을 활용한 심혈관질환 질환단계 예측 및 진단보조 시스템인 AptaCDSS에 대해서 다루었다. 분류기 중에서는 NN이 가장 좋은 성능 결과를 보였다. 현재 개발되고 있는 다양한 CDSS나 MDSS의 성능이 70~80% 정도의 성능을 갖는다는 점을 고려하면 일단 비교될 정도의 성능을 얻고 있다고 할 수 있다. 그렇지만 모델 학습에 사용한 데이터가 적고 보다 신뢰성 이는 결과를 위해서는 데이터를 확충하는데 보다 노력을 기울여야 할 것이다.

또한, 논문에서 사용한 방법은 어느 정도 유의미한 분류 성능을 보여주고는 있지만, 실험에 사용된 데이터는 심혈관 질환의 실제 임상진단에서 중요 요소로 고려되는 환자의 나이, 흡연여부, 혈중 콜레스테롤 수치 등과 같은 칩데이터 이외의 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.

5.2. 보다 유용한 CDSS/MDSS 구성을 위한 제언

AptaCDSS 개발을 통해 확인한 바에 의하면 실제로 CDSS나 MDSS를 사용하는 관련 전문가를 위해서는 다양한 요소들이 고려되어야 하겠지만 최소한 다음과 같은 사항들이 고려된다면 국내외 CDSS/MDSS와 비교하여 어느 정도 경쟁력을 가지는 시스템을 구성할 수 있을 것으로 전망된다.

시스템 목적의 구체화 — 환자데이터 분석을 통한 바이오마커의 탐색 및 발굴 또는 병원정보시스템(HIS)의 일부로 환자 진단 보조용 시스템 등의 목적을 구체화 하여 시스템 전체 설계를 다르게 하여야 한다.

목적 분야의 요구사항 조사 — 실제 시스템을 사용하는 임상(진단)의의 경험적 정보 및 필드에서 원하는 상황에 대한 조사 및 반영은 다른 어떤 요소보다 먼저 조사·분석되어야 하는 사항으로 이에 따라 시스템 전체의 방향을 미리정해야 한다.

진단결과 제시방법의 다양화 — CDSS 구성을 위해 사용하게 되는 모델들의 출력결과를 다양한 방식으로 해석하여 가시화하는 것은 유용한 CDSS/MDSS 구현에 있어 중요한 요소 중의 하나이다. 이를 위해서는 "목적 분야의 요구사항 조사"를 보다 적극적으로 수행하여 필드에서 일반적으로 사용되는 방식 및 필드의 요구사항 수렴에 더해 새로운 가시화 방법을 디자인 하는 것으로, 다양한 분야의 시스템 사용자 또는 특정 분야의 시스템 사용자들 중에서도 서로 다르게 요구되거나 선호되는 방식을 시스템이 포괄할 수 있는 수용력 및 사용자 집단의 만족도를 높여준다.

데이터의 증분축적 및 마이닝 기능 강화 — 의학진단 분야의 데이터는 대량 수집이 쉽지 않기 때문에 진단을 위해 제공되는 환자 데이터 및 이에 대한 진단의사의 최종 판단결과를 기초로 진단시스템의 학습 데이터를 축적하고 이를 기반으로 한 성능 향상 전략의 수립이 필요하다. 이는 국내외 CDSS/MDSS가 아직 구현하고 있지 못한 기능이지만 데이터부족문제 및 시스템 성능의 점진적 향상에 도움이 된다는 다양한 연구들이 진행되어왔다.

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. 더불어 이 연구를 위해 연구 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에 감사드립니다.

참고문헌

[1] CDC's Report 1, http://www.cdc.gov/nccdphp/overview-.htm (accessed May 29, 2006).

[2] CDC's Report 2, http://www.cdc.gov/nccdphp/publications/aag/cvh.htm (accessed May 29, 2006).

[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] Qu, Y., Adam, B.-L., Yasui, Y., Ward, M. D., Cazares, L. H., Schellhammer, P. F., Feng, Z., Semmes, O. J., & Wright, G. L. Jr. (2002). Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48(10), 1835-43.

[6] Won, Y., Song, H., Kang, T. W., Kim, J., Han, B., & Lee, S. (2003). Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons. *Proteomics*, 3(12), 2310-6.

[7] Sadeghi, S., Barzi, A., Sadeghi, N., & King, B. (2006). A Bayesian model for triage decision support. International *Journal of Medical Informatics*, 75(5), 403-11.

[8] Yan, H.-M., Jiang, Y.-T., Zheng, J., Peng, C.-L., & Li, Q.-H. (2006). A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Systems with Applications*, 30(2), 272-81.

[9] Prados, J., Kalousis, A., Sanchez, J. C., Allard, L., Carrette, O., & Hilario, M. (2004). Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents. *Proteomics*, 4(8), 2320-32.

[10] 엄재홍, 장병탁, "SVM 앙상블을 이용한 심혈관질환 질환단계 예측", *한국컴퓨터종합학술대회 2006 논문집*, 제33권 1(A), pp. 76-78, 2006.

[11] 김병희, 장병탁, "기계학습에 의한 압타머칩 데이터 기반 심혈관 질환 단계의 예측", *한국컴퓨터종합학술대회 2006 논문집*, 제33권 1(A), pp. 85-87, 2006. ■