

사례기반추론을 이용한 대용량 데이터의 실시간 처리 방법론 : 고혈압 고위험군 관리를 위한 자기학습 시스템 프레임워크*

박성혁** · 양근우***

Data Mining Approach for Real-Time Processing of Large Data
Using Case-Based Reasoning : High-Risk Group Detection Data
Warehouse for Patients with High Blood Pressure*

Sung-Hyuk Park** · Kun-Woo Yang***

■ Abstract ■

In this paper, we propose the high-risk group detection model for patients with high blood pressure using case-based reasoning. The proposed model can be applied for public health maintenance organizations to effectively manage knowledge related to high blood pressure and efficiently allocate limited health care resources. Especially, the focus is on the development of the model that can handle constraints such as managing large volume of data, enabling the automatic learning to adapt to external environmental changes and operating the system on a real-time basis. Using real data collected from local public health centers, the optimal high-risk group detection model was derived incorporating optimal parameter sets. The results of the performance test for the model using test data show that the prediction accuracy of the proposed model is two times better than the natural risk of high blood pressure.

Keyword : Clinical Decision Support System, High Blood Pressure, Data Mining, CBR

논문투고일 : 2010년 10월 13일 논문수정완료일 : 2010년 12월 03일 논문게재확정일 : 2010년 12월 10일

* 이 논문은 2010년도 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국연구재단의 지원을 받아 연구되었음(NRF-2010-327-B00188).

** KAIST 경영대학 경영공학전공 박사과정, 주저자

*** 계명대학교 사회과학대학 전자무역학과 조교수, 교신저자

1. 서 론

의사결정지원 시스템(Decision Support Systems, DSS)에서 제공하는 정보를 바탕으로 신상품 개발 전략이나 투자계획과 같은 주요 의사결정을 내리는 경우, 의사결정지원 시스템에서 고품질(high quality)의 정보를 제공할 수 있어야 좋은 의사결정을 할 수 있다. 고품질 정보에 대한 정의는 사용 목적에 따라서 상대적으로 정의하는 것이 일반적이며 [17], 기업에서 시장의 최근 동향을 파악하거나, 선도적인 제품을 개발하려고 하는 경우와 같이 최신 정보를 취급해야 하는 경우, 의사결정지원 시스템에서는 가장 최근의 정보까지 반영한 결과를 제공할 수 있어야 한다. 예를 들어 경쟁사가 신규로 진입한다거나, 국제적 금융위기가 발생하는 상황과 같이 대규모의 외부 환경 변화가 발생할 경우, 이러한 변화가 사람들의 소비행태와 같은 행동방식에 영향을 주어 달라지게 할 수 있다. 따라서 전과 다르게 관찰되는 정보를 의사결정자가 신속하게 파악할 수 있도록 의사결정지원시스템에서 최신의 정보를 빠르게 제공할 수 있어야 한다.

의사결정지원 시스템에서 제공하는 정보가 최신의 것인지를 확인하기 위해서는 의사결정지원시스템이 어떻게 구성되는지를 이해해야 하는데, 일반적으로 데이터를 적재하고 있는 데이터 웨어하우스(Data Warehouse, DW)로 부터 필요한 정보를 추출하는 과정을 거쳐 의사결정에 도움이 되는 수준의 정보로 가공해나가는 일련의 과정으로 설명할 수 있다[16]. 여기서 데이터 웨어하우스는 과거로부터 현재까지 누적된 통합 데이터를 담고 있어 의사결정자가 필요로 하는 정보를 추출하는데 효과적으로 사용된다[9]. 데이터 처리 관점에서 보면, 최근에 발생한 데이터가 데이터 웨어하우스에 적재되고 나서, 사용자가 사전에 정의한 형식에 따라 2차, 3차 정보의 형태로 가공이 되어 제공된다. 이 때, 정보를 가공하는 과정에서 특정 시점에 종속적으로 정의된 상황적(contextual) 모형이 사용되는 경우, 입력 데이터가 최신의 것임에도 불구하고

하고 모형의 결과로 얻어진 정보는 최신의 것으로 보기 어렵다. 예를 들어, 일 년 전에 사용되었던 고객이탈 모형을 사용해오던 중, 경쟁 업체가 새롭게 진입한다거나 산업 내에서 대체 상품이 개발되는 식의 외부 변화가 발생한다면 과거에 개발된 모형이 현재까지 높은 정확성을 유지하는데 무리가 있다. 이것은 특정 시점에서 일회성으로 개발된 모형의 경우, 모형 개발 이후 시점에서 발생한 외부 환경의 변화를 반영하고 있지 않기 때문에, 신규 데이터가 입력되더라도 데이터를 처리하는 기준이 과거의 정보만을 담고 있게 되어 문제가 발생한다. 따라서 특정 시점에서 이벤트 형식으로 개발된 모형에 의하여 생성된 정보를 사용하게 된 의사결정자는 오류를 범할 수 있다.

본 연구에서는 이러한 문제점을 해결하기 위하여 자기학습(self-learning) 방식을 통해 외부 환경의 변화를 실시간으로 반영하여 모형을 수정해나가는 시스템 프레임워크를 제안한다. 구체적으로, 사례기반추론 방법론을 사용하여 예측문제를 해결하는 문제를 정의하고, 의사결정지원 시스템 내부에서 데이터를 처리하는 프로세스를 자동화하는 방식을 제안한다. 연구 방법론 측면에서는, 데이터 웨어하우스 상에서 데이터를 처리하는 프로세스 전체를 자동화하고, 과거에 일회성으로 개발되어 전문가의 개입 없이 모형을 수정하기가 어려운 데이터마이닝 모형의 단점을 해결하기 위하여 사례기반추론 기법을 적용한 시스템을 개발한다.

시스템 개발 및 실험에 사용된 데이터는 한국의 보건소에서 조사한 고혈압 정상인 및 환자들의 관련 질환 설문 응답결과이다. 건강 상태를 측정할 데이터가 실험에 사용된 이유는 다음과 같다. 첫째, 공해나 환경오염 사건과 같은 외부 환경에 의하여 사람들의 식습관이나 질환 발병 패턴이 변하는데, 고혈압에 영향을 미치는 요인이 시간에 따라 어떻게 변화하고 있는지를 실시간 수준으로 관리하는 것이 실제로도 중요한 문제이며, 본 연구에서 제안하고자 하는 의사결정시스템 개발 목적에 부합하기 때문이다. 둘째, 보건소에서

국민의 고혈압과 같은 만성질환을 관리하고자 할 때, 전체 국민들이 발생시키는 데이터의 양이 하루에도 대규모로 발생하기 때문에 이를 실시간으로 처리하는 모형이 개발되어야 하며 이를 통해 본 연구에서 제안한 모형을 검증할 수 있는 환경을 제공하기 때문이다.

고혈압 환자를 관리하는 의료의사결정지원 시스템의 개발이 중요한 이유는 다음과 같다. 정부에서는 주요 국민건강 정책의 일환으로 국민을 대상으로 하는 정기적인 혈압 측정을 통한 고혈압 관련 질환의 조기 발견과 적극적인 치료 및 예방사업을 추진하고 있다. 그러나 현실적인 측면에서 살펴보면 고혈압 치료 및 예방사업의 추진 과정에서 다음과 같은 어려움이 지적된다. 먼저, 고혈압 예방에 어려움을 주는 요인을 살펴보면, 일반 보건소 이용자 중에는 무관심과 같은 개인적인 사유로 혈압 측정을 정기적으로 하지 않는 사람이 대부분이기 때문에 규칙적으로 고혈압 수준이나 관련 증상에 대해서 사전에 관리하기 어렵다는 사실이다. 다음으로 고혈압 치료의 경우, 혈압을 규칙적으로 측정 하는 경우나, 심지어 의사의 처방에 순응하거나 개인이 지속적으로 관리하는 경우에도 정상 혈압 수치로 회복되는 것이 용이하지 않기 때문에 치료가 쉽지 않다. 게다가, 고혈압 환자 개인별로 고혈압에 대한 발병원인이나 치료방법을 확실하게 알지 못한다는 점도 고혈압을 평소에 측정하고 관리해야 하는 이유로 보고된 바 있다[10].

일반적으로 보건소와 같은 공공기관이 고혈압 만성질환 관리 서비스를 국민에게 제공하는 사업의 성과는 한정된 자원을 얼마나 효과적으로 잘 분배하느냐에 따라서 달라질 수 있다. 보건 기관 별로 담당 지역 내에서 문제가 발생할 가능성이 높은 고혈압 고위험군을 사전에 탐지할 수 있다면, 이들을 대상으로 집중적으로 혈압 측정을 유도하는 것은 물론 혈압 조절이 가능하도록 의료 자원을 효과적으로 사용할 수 있을 것이다. 본 연구에서는 이러한 목적을 구체적으로 달성하기 위하여 지역별 보건소를 통해 수집된 자료를 바탕으로 고혈

압 고위험군을 탐지하는 모형을 수립하고자 한다. 지속적인 혈압 측정과 고혈압 관리가 필요한 고혈압 고위험군 환자 중에서도 정기적으로 혈압 측정을 하지 않고 있는 대상자들을 식별하기 위하여 데이터마이닝 기법을 활용한 추정모형을 제안한다. 다음으로, 이렇게 수립된 모형에 대해 실증 데이터를 이용하여 고혈압 고위험군 대상자를 식별해내는 예측 정확도를 평가한다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 고혈압 고위험군 관리 및 의료지원 시스템과 관련된 기존의 연구들을 소개하고, 제 3장에서는 본 연구에서 제안하는 사례추론기반 자기학습 시스템 프레임워크를 설명한다. 다음으로 제 4장에서는 실제 데이터를 사용하여 개발된 시스템을 평가하는 실험을 수행하며, 마지막으로 제 5장에서는 시스템 개발 결과를 평가하고 시사점을 도출한다. 또한, 향후 활용방안에 대해서도 설명한다.

2. 관련 연구

2.1 고혈압 고위험군 탐지 모형

데이터마이닝 기법을 사용하여 고혈압 관리를 위한 의료 의사결정지원 시스템[15] 개발을 목적으로 한 연구[5]에서는 고혈압 환자의 특성과 치료에 따른 예후를 예측할 수 있는 지식을 탐색하고, 이러한 지식의 임상적용이 타당하다는 것을 검증한 바 있다. 또한, 김태수 외 4인의 연구[2]에서는 고혈압 관리에서 위험요인을 분석하고 이를 해결하기 위한 치료방법 탐색의 방안으로 데이터마이닝 기법을 적용하여, 치료결과에 영향을 미치는 주요 요인을 도출하였다. 하지만 과거에 수행되었던 연구들은 모두 특정 기간 동안 수집된 데이터를 대상으로 데이터마이닝 기법을 적용하여 모형을 개발하고, 모형의 검증력을 평가하는 식으로 진행된 것이기 때문에, 실제로 작동할 수 있는 시스템을 개발하려는 노력이 부족하였으며, 시스템 운영에 대한 문제를 해결하려는 시도를 하지 못하였다.

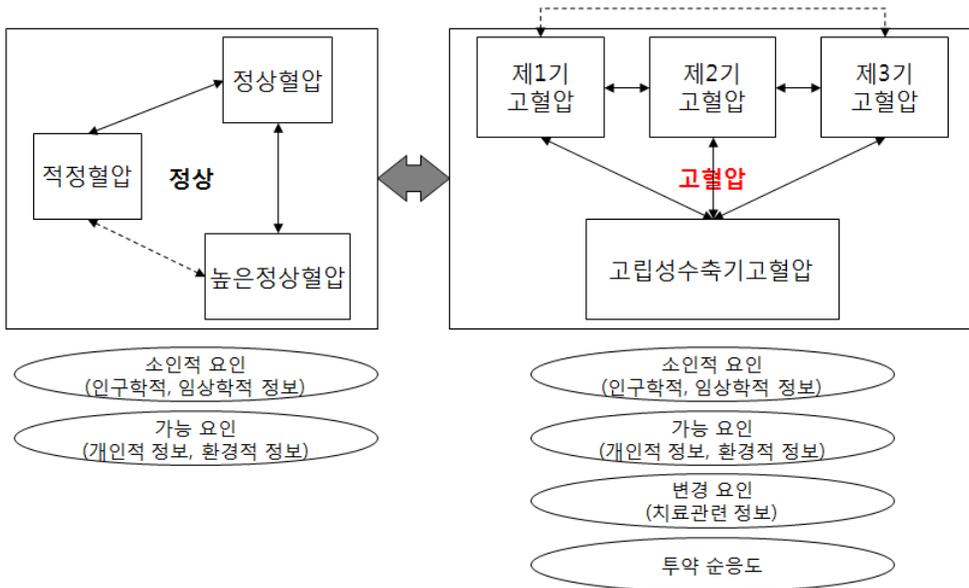
기존 연구와는 달리 본 연구에서 새롭게 수행하고자 하는 것은 지역 보건 기관이 관리해야 하는 수많은 사람들을 대상으로 대용량 데이터 처리에 특화된 의료 의사결정지원 시스템의 개발이다. 의료자원의 효과적인 분배라는 관리 목표를 달성하고, 실시간 운영 가능한 의료정보 시스템을 구축하기 위해서 대용량 데이터를 대상으로 하는 실시간 데이터 처리, 외부환경 변화에 대한 자동학습 기능이 포함된 모형을 개발한다. 따라서 본 연구는, 시스템 개발과 관련하여 데이터 웨어하우스에서 수많은 사람들이 보고하는 고혈압 관련 정보로 구성된 대용량 데이터를 다룰 수 있어야 하며, 고혈압 발병에 영향을 주는 요인을 실시간으로 관리하여 가장 최근에 발생하고 있는 관련 질환과의 인과관계를 파악할 수 있도록 시스템에 반영할 수 있어야 한다. 이 과정에서 사례기반추론 방법론을 사용하여, 가장 최근 시점까지 데이터 웨어하우스에 적재된 데이터를 기반으로 특정 대상자에 대해서 고혈압 고위험군 판단여부에 대한 의사결정을 내릴 수 있도록 해야 한다. 이와 관련된 기존 연구들은 고혈압에 영향을 주는 요인들을 설명하는 부

분과, 방법론 측면에서 사례기반추론 기법을 적용하여 시스템을 개발하는 부분으로 구성된다.

2.2 고혈압과 고혈압 영향요인

보건복지가족부의 고혈압관리지침에 의하면 고혈압 환자는 수축기혈압과 확장기혈압 측정값을 바탕으로 판정되며, 수축기 140mmHg 이상 또는 확장기 90mmHg 이상인 경우에 고혈압으로 구분한다. 본 연구에서도 동일한 기준으로 고혈압 환자를 정의하였으며, 정상 혈압 범위를 벗어나고 있는 환자를 고혈압 고위험군으로 분류해 혈압측정과 관리에 대한 우선대상자로 선정하였다.

혈압에 영향을 주는 요인은 매우 다양하고 요인들 간의 상호작용도 큰 것으로 알려져 있다. 지금까지의 다양한 연구를 바탕으로 고혈압에 영향을 주는 요인을 분류하면 [그림 1]에서 설명하는 것처럼 인구학적, 임상학적 정보를 의미하는 소인적 요인과 개인적, 환경적 정보에 해당 하는 가능 요인으로 나눌 수 있다[2, 5, 10]. 먼저, 소인적 요인의 경우 연령과 성별 등을 포함하는 인구통계학적



[그림 1] 고혈압 영향 요인

요인을 포함하는데 성별, 연령대별로 고혈압 관련 질병 발생 확률을 달리한다. 또한 고지혈증, 관절염, 흉부질환, 골다공증, 심혈관 가족력, 염좌, 뇌졸중, 빈혈증, 비만 등 다양한 질환 보유 여부와 특정 질환의 분류 단계 등이 포함된다. 또한 통계학적으로 고혈압과 유의한 연관성을 보이는 것으로 알려진 가능 요인으로 활용도가 높은 정보들이 있는데, 학력, 직업, 결혼상태, 생활수준, 주거유형, 총가구원수, 흡연여부(양), 음주여부(횟수), 운동여부(양) 등이 여기에 해당한다.

특정 환자가 고혈압으로 진단을 받고 치료를 시작할 경우 처방에 따라 복용하는 약품에 의해 해당 환자의 혈압 상태는 변하게 된다. 이러한 치료 관련 정보인 변경 요인에는 평균처방일수, 평균방문주기, 총 방문횟수, 단위기간 평균방문 횟수, 일일 처방횟수평균, 일일 복용약평균 등이 포함되며 복용하는 약품에 대한 특정 환자의 투약 순응도도 혈압에 영향을 미치는 요소로 작용하게 된다.

이러한 소인적 요인, 가능 요인, 변경 요인, 투약 순응도 등이 고혈압에 영향을 주며 개별 환자에 대한 이러한 요인 측정값을 활용하여 고혈압 고위험군을 추정하는 모형에 활용하는 것이 가능하다. 본 연구에서는 소인적 요인과 가능 요인들을 활용하여 정기적인 고혈압 측정이 되지 않고 있는 고위험군 환자들에게 예방적 활동을 가능케 하고자 고혈압 고위험군 추정 모형을 개발한다.

2.3. 사례기반추론

사례기반추론은 특정 문제를 해결하기 위해서 과거에 이와 유사한 문제를 해결했던 사례를 바탕으로 솔루션을 찾아 적용하거나, 보다 개선된 해결책을 마련하여 문제를 풀고자 하는 대표적인 기법이다[14]. 특히 의학이나 법률 분야와 같이 과거 사례에 대한 경험과 전문성이 중요시되는 분야에서 사례기반추론 방법론이 사용되는 것이 적합한 것으로 알려져 있으며, 실제로 해외 연구에서는 의학 및 영양학 분야에서 이 기법이 적용된 바 있

으며[11-13], 국내에서도 사례기반추론을 바탕으로 운동처방지원 시스템을 개발하거나[1] 데이터 웨어하우스를 설계한 연구가[3] 있다. 최근에는 고객관계관리(CRM) 분야에서 다이렉트 마케팅의 고객 반응예측 모형을 통합하는데 사례기반추론이 적용되는 등[6] 널리 이용되고 있다.

사례기반추론은 크게 네 단계로 진행되는데, 각 단계에 대한 설명은 다음과 같다[7, 14].

- ① 사례 조회(Retrieve) : 사전에 적재해놓은 사례 데이터 저장소에서 현재 해결하고자 하는 문제와 가장 유사한 사례를 검색하는 단계로, 해당 문제를 푸는데 가장 도움이 되는 과거의 대표 사례를 찾는 과정이다.
- ② 사례 적용(Reuse) : 앞 단계에서 조회한 대표 사례를 바탕으로, 현재 풀고자 하는 문제에 대한 해결책을 찾는 단계로, 사전에 정해진 절차에 따라 기존에 문제를 해결했던 답안으로부터 현재 문제를 해결하는데 유용한 정보를 찾아 적용한다.
- ③ 사례 교정(Revise) : 과거에 해결되었던 문제 상황과 현재 풀고자 하는 문제 상황이 100% 정확하게 일치하지 않는 것이 보통이므로, 이 간극을 메우기 위하여 기존의 해결책을 수정하여 실제 적용 가능한 최종 수정안을 마련하는 단계이다. 과거 사례를 적용하였을 때 성공적인 결과가 얻어지지 않은 경우 사례에 대한 해결책을 수정해야 한다.
- ④ 사례 학습(Retain or Learning) : 과거 사례가 현재 문제를 해결하는데 도움이 되었던, 그렇지 않던 간에 성공과 실패에서 배울 수 있는 교훈을 지식화하여 사례 저장소에 저장하는 단계이다. 만약 사례 교정단계에서 새로운 사례가 개발된 경우 신규 사례로 기존 사례 저장소에 등록될 수 있다.

사례기반추론의 장점은 데이터베이스 관리 분야에서 다루어지는 데이터 처리 기법을 적용하여 대용량 데이터를 처리할 수 있을 뿐만 아니라, 새로

운 지식을 기존 사례 저장소에 추가함으로써 유지 보수 차원에서 시스템을 손쉽게 운영해나갈 수 있다는 것이다[18]. 이러한 장점 덕분에 본 논문에서 제안하려고 하는 대용량 데이터의 실시간 처리 문제와, 신규 지식을 추가하거나 신규 지식에 의하여 기존 지식들을 수정하는 문제를 해결하는데 사례기반추론 기법을 기반으로 데이터 웨어하우스를 개발하도록 한다.

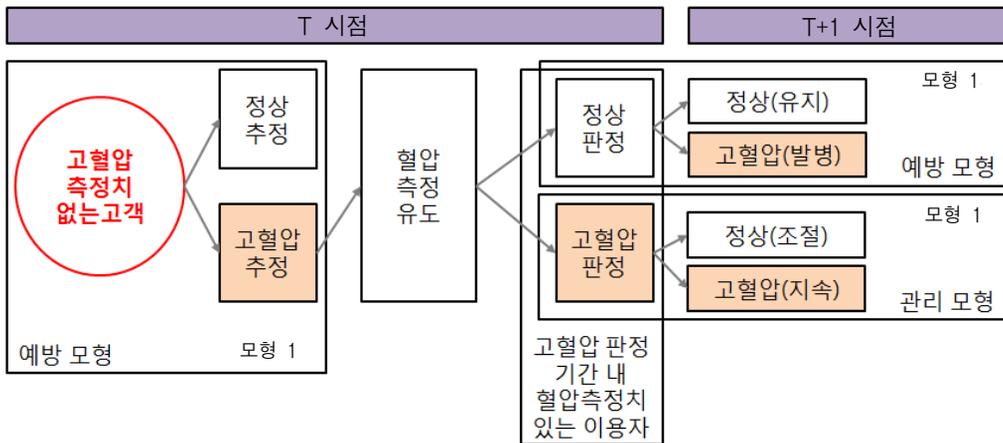
3. 고혈압 고위험군 탐색을 위한 자동학습 데이터웨어하우스

본 연구에서 제안하는 고혈압 환자 추정 및 관리를 위한 통합 모형은 [그림 2]와 같다. 특정 시점(T)에서 고혈압 측정 수치가 존재하지 않는 환자 중 고혈압으로 의심되는 고위험군 환자를 추정하는 모형을 개발하여 이러한 환자들에 대해 적극적으로 혈압 측정을 유도한다. 모형에 의해 선별된 고혈압 고위험군에 해당하는 그룹이 혈압을 측정한 후 그 결과가 정상 판정일 경우에는 다음 혈압 측정 시점(T+1)까지도 정상 혈압을 유지할 수 있는 예방 모형을 작동하게 된다. T시점에서 측정된 혈압이 고혈압으로 판정되었을 경우 적절한 조치를 통해 정상 혈압 수준으로 혈압을 낮출 수 있는 관리

모형을 작동시킬 수 있다.

고혈압 고위험군 탐지 모형을 개발하기 위한 방법론으로는, 고혈압 발생 확률 모형과 데이터 마이닝 기법이 주로 사용된다[2, 4, 5]. 본 연구에서는 외부환경 변화에 강인하면서 대용량 데이터 처리를 원활하게 하기 위하여 사례기반추론 기법을 바탕으로 하는 확률모형이 사용되었다. 사전에 학습한 사례기간의 확률정보를 바탕으로 특정인이 고혈압 환자일 가능성이 높은지를 판단하고, 사례에서 학습시킨 각 규칙에 의거하여 고혈압 고위험군을 분류해내는 방식으로 진행된다. 이를 위해 데이터웨어하우스 상에 이미 존재하는 다른 고혈압 여부 판정자들의 소인적, 가능 요인에 대한 확률 값을 계산한 다음, 특정 이용자가 처한 상황과 비교하는 확률적인 추정 모형을 적용하게 된다. 예를 들어 전체 대상자 중에서 일반적으로 고혈압이 발병하는 자연 확률 $P(\text{고혈압}) = 27\%$ 를 알고 있고, 특정 보건소 이용자가 처한 소인적, 가능 요인 정보에 대해서 $P(\text{고혈압}|60대, 남자, 당뇨병보유) = 71\%$ 라고 확인이 되었다면, 그 이용자에 대해서는 확률적인 관점에서 고혈압 발병 가능성이 높으니 고혈압 수치를 측정해보는 것을 권장할 수 있으며, 이러한 사실은 의료정보로서도 가치가 있다[8].

개발된 모형에 의해 고혈압 환자로 추정되었다



[그림 2] 고혈압 환자 추정 및 관리를 위한 통합 모형

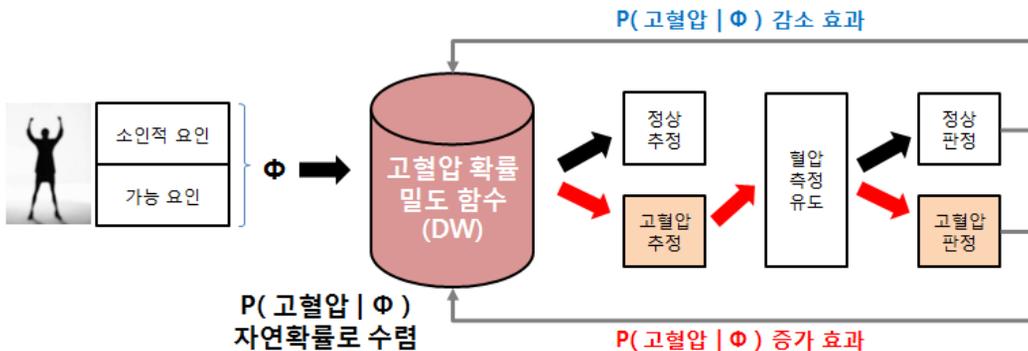
면, 통합모형에 의하여 혈압 측정을 유도한다. 이 과정에서 실제 측정치를 얻은 다음, 고혈압 판정 기준에 따라 고혈압 유무를 판정해 그 결과를 ‘고혈압 확률 데이터웨어하우스’의 구성 레코드로 업데이트 해줌으로써, 고혈압 추정 모형에서의 확률 값이 자동으로 조정되도록 설계하였다. 이를 최초 모형설정 단계 이후에 전문가의 개입 없이도 지속적으로 모형을 정교화하고 외부 환경 변화를 자동으로 반영하여 시스템의 추정 성능을 향상시키게 된다. 예를 들어, P(고혈압|40대, 남자, 미혼, 대도시 거주) 값이 높게 얻어져서 고혈압 환자로 추정된 사람이 실제 혈압 측정 결과 정상으로 얻어지면, 그 결과가 고혈압 확률 데이터 웨어하우스에 다시 반영되므로 다음 추정 시에는 특정 사람에 대해서 P(고혈압|40대, 남자, 미혼, 대도시)를 계산해준 결과가 기준 보다는 낮은 수준으로 조절된다. 이러한 조정은 실제 판단 결과에 따라 수행되며, 외부 환경 변화에 의해 더 이상 고혈압 고위험군 탐지에 유의하지 않다고 판단되는 규칙은 대수의 법칙에 의해 자연확률 값으로 수렴하게 되어 손쉽게 관리할 수 있게 된다([그림 3] 참조).

사례 저장소에 기록되는 사례를 Φ_i 라고 표현하자. 여기서, Φ_i 는 고혈압 유무를 나타내는 변수를 제외한 나머지 변수들에 대해서 If Then 조건의 나열로 구성된다. 그러면 특정 사례, Φ_i 에 해당하는 사람이 고혈압에 걸릴 확률은 다음과 같이 표현된다.

$P(\text{고혈압}|\Phi_i) = \Phi_i$ 에 해당하면서 고혈압 보유자의 수/ Φ_i 에 해당하는 사람 수

Φ_i 에 해당하는 새로운 사람에 대한 정보가 추가될 때 마다, 위 식에서 분모에 해당하는 Φ_i 에 해당하는 사례 수는 증가하며, 이 사례에 해당하는 사람이 고혈압 고위험군에 해당하는지 아닌지에 따라서 분자 항은 증가하거나 동일한 수준으로 유지된다. 결과적으로는 새로운 사례가 추가될 때 마다 P(고혈압| Φ_i) 값이 변화하게 된다. 마찬가지로 고혈압 발병 기준 확률값, $P(\text{고혈압}) = \text{고혈압 보유자의 수/전체 사람 수}$, 자체도 변화하게 된다. 물론 기준에 수많은 사례를 바탕으로 계산된 확률 값들이 한 명의 사람에 대한 정보가 추가된다고 해서 크게 바뀐다고 보기는 어렵지만, 결국 고혈압 고위험군을 탐지하는 과정에서 특정 사례가 이 업무를 수행하는데 중요한 사례로 승격되거나 탈락되는 계기가 되기도 하므로, 변화를 실시간으로 감지할 수 있으면 세세한 수준으로 사례를 관리할 수 있어야 한다.

고혈압 고위험군 여부를 탐지하고자 하는 특정 대상자를 평가하는 기준은, 이 대상자가 해당하는 사례 Φ 에 대해서, 현재 사례 저장소에서 계산해준 $P(\text{고혈압}|\Phi)$ 값이 의사결정 기준을 초과하는지 아닌지에 의존한다. 만약에 고혈압이라고 판정이 된 경우, 이 사람의 데이터가 일차적으로 데이터 저장소에 업데이트 되면, 사전에 정의된 데이터 처



[그림 3] 고혈압 추정 모형의 자동학습 기능

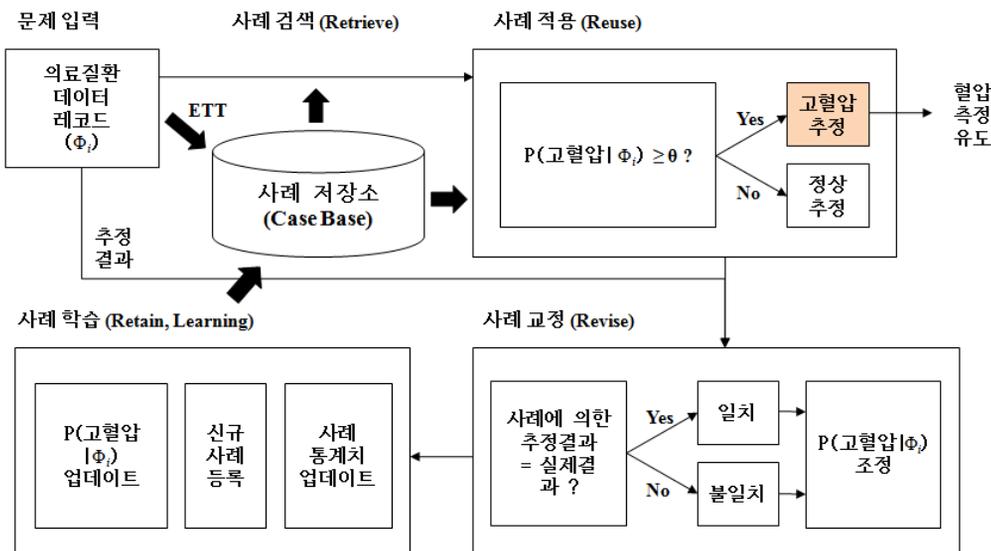
리 과정에 의해서 사례 저장소에서 해당 사례, Φ , 에 대한 $P(\text{고혈압}|\Phi)$ 값이 증가하게 된다. 만약에 사례 Φ 에 해당하는 고혈압 환자들의 이력이 계속 발생한다면, $P(\text{고혈압}|\Phi)$ 값 역시 증가하게 되며, 해당 사례는 고혈압 고위험군 대상자를 탐지하는데 중요한 사례가 될 것이다. 반대로, 사례 Φ 에 해당하는 사람들이 고혈압 환자가 아닌 것으로 판정된다면, $P(\text{고혈압}|\Phi)$ 값이 점차 감소할 것이며, 의사결정 기준보다 낮은 확률 값을 가지게 된다면 고혈압 고위험군 대상자를 탐지하는데 중요하지 않은 사례로 교정될 것이다. 이러한 절차가 사례 기반추론을 구성하는 각 단계에 해당된다.

본 연구에서 제안하는 고혈압 고위험군 추정 모형은 데이터 웨어하우스에 저장된 데이터를 대상으로 분석을 하기 위한 SQL 프로시저 형태로 작성하였으며, 고혈압 사례 저장소에서 각 사례에 대한 고혈압 고위험군 판단 확률을 관리하기 위한 $P(\text{고혈압}|\Phi)$ 값을 앞서 제시된 절차를 통해 계산되어 데이터 웨어하우스 상에서 사례 저장소에 적재된다. 다음으로, 특정 대상자에 대해서 위 단계를 거쳐 고혈압 고위험군 여부를 판단해준 후 실제로 고혈압 환자인지 아닌지에 대한 정보가 추가

되었을 때, 고혈압 환자로 판정되었으면 해당 사례에 대한 고혈압 추정 확률이 증가하도록 교정되어야 할 것이며, 고혈압 환자가 아닌 것으로 판정되었으면 고혈압 추정 확률이 감소하도록 교정시켜주어야 한다.

이상에서 설명한 고혈압 고위험군 탐지를 위한 시스템 프레임워크는 사례기반추론의 순환과정에 따라[7] 다음과 같이 네 단계로 설명할 수 있다 ([그림 4] 참고).

- <사례조회> 추정 대상자의 의료 데이터 레코드(Φ)를 받아 결측치가 아닌 속성들의 데이터 값을 기준으로, 사전에 구축된 사례 저장소 내의 고혈압 고위험군 유효 사례 중에서(비교 대상을 줄여 시스템 성능을 개선하기 위한 조치임) 추정 대상자와 가장 유사한 조건을 보유한 사례를 조회함. 이 때, 추정 대상자의 데이터 레코드에서 결측치가 아닌 부분에 대해서 정확하게 일치하는 사례들이 검색되는 경우에는 그 결과를 선택하며, 정확하게 일치하지 않는 경우에는 기존 데이터 저장소의 사례들과 matching coefficient(비교 대상 변수 중에서 두 값이 동



[그림 4] 고위험군 탐색 시스템 프레임워크

일하면 '1', 상이하면 '0'으로 설정한 다음, '1'의 비율을 계산하는 방식)를 계산하여 유사도를 측정하는 다음, 가장 높게 얻어진 사례들을 선택함.

- <사례적용> 전 단계에서 검색한 사례에 대한 고혈압 추정확률값이, $P(\text{고혈압}|\Phi)$, 의사결정기준 이상일 경우, 고혈압 고위험군 추정 사례로 등록하고, 그렇지 않을 경우 고혈압 고위험군 추정 사례에서 제외함. 이 때, 실험을 통하여 사전에 정의된 의사결정기준 값을 사용함
- <사례교정> 고혈압 여부가 확인된 데이터 레코드가(추정 결과 또는 신규 데이터 레코드) 한 건 이상 신규 추가되면, 해당 의료 레코드(Φ) 값에 대한 $P(\text{고혈압}|\Phi)$ 값을 조정해줌. 구체적으로 고혈압 추정자가 고혈압으로 판정되면, 해당 사례의 고혈압 확률 값은 증가하고, 그렇지 않은 경우는 감소함
- <사례학습> 사례에 대한 고혈압 확률 값 사례 저장소에 업데이트 하며, 신규 사례의 경우 새로 등록하도록 함. 또한, 고혈압 고위험군 탐지에 적합한 사례들에 대한 통계치를 사례 저장소에 반영함([그림 5] 참조).

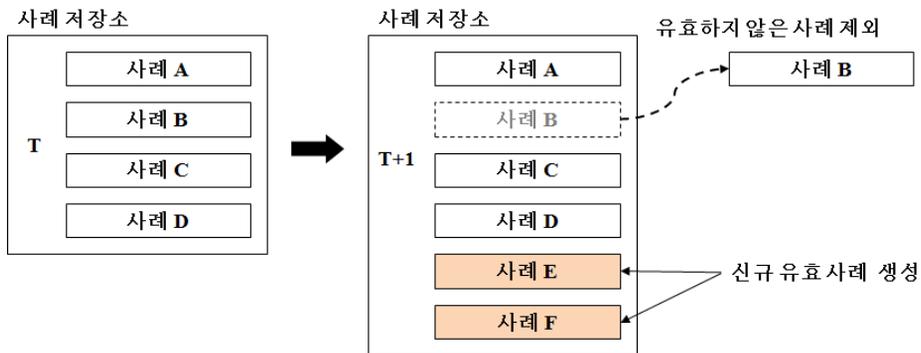
4. 실험

고혈압 고위험군 관리를 위한 자기학습 시스템 프레임워크의 유용성과 실제 적용가능성을 검토하기 위하여 실제 보건소 이용자의 고혈압 여부 및

관련 질환 정보를 포함하는 데이터를 수집하였으며, 이를 바탕으로 실험을 수행하였다. 먼저, 고혈압 고위험군 추정을 위하여 사례 저장소에 사례들을 생성해야 하는데, 사례를 구성하는 세부 속성들을 선정하기 위한 작업을 수행하였다. 총 19개의 변수가 선정되었으며, 각 단계 별로 아래와 같이 진행되었다.

- ① 첫째로, 국내외 고혈압 관리 지침이나 관련 연구 문헌에서 고혈압과 관련 요인으로 밝혀진 것들 중에서, 실제 추출 가능한 변수를 선정함
- ② 둘째로, 고혈압 여부에 영향을 주는 변수를 탐색하는 통계적 유의성 검증(feature selection)에서 통과한 변수를 선택 함. 이 때, 명목형 변수들간의 상관관계를 검증하기 위하여 cross-tab 방식에 의한 유의성 검증을 함
- ③ 셋째로, 변수 유의성 검증에서는 선택되지 않았지만, 보건기관 전문가 의견에서 설명 변수로 포함하도록 권장된 변수를 추가하였다.

위와 같은 세 단계 절차를 거쳐, 최종적으로 선정된 소인적 요인은 성별, 연령대, 기타질환유무, 흡연여부, 음주여부, 운동여부, 고지혈증유무, 비만유무, 타질환유무, 당뇨병조절유무이었으며(연도구분 : T년도, T-1년도), 가능 요인은 거주자의 행정 지역구분, 건강관리사업등록유무, 결혼상태, 생활수준, 주거유형이었다(<표 1> 참조).



[그림 5] 사례 학습에 의한 사례저장소 업데이트 예시

〈표 1〉 후보 변수 유의성 검증 결과

변수 명	검증 결과	최종 선정
성별	유의함*	O
연령대	유의함*	O
기타질환유무_T	유의함*	O
기타질환유무_T-1	유의함*	O
흡연여부_T	유의함*	O
흡연여부_T-1	유의함*	O
음주여부_T	유의함*	O
음주여부_T-1	유의함*	O
운동여부_T	유의함*	O
운동여부_T-1	유의함*	O
고지혈증여부_T	유의함*	O
비만여부_T	유의안함	O(구제)
그 외 다른 질병_T	유의안함	O(구제)
당뇨병조절유무	유의함*	O
거주지행정지역구분	유의함*	O
건강관리사업등록유무	유의함*	O
결혼상태	유의함*	O
생활수준	유의함*	O
주거유형	유의함*	O

주) 위 검증 결과에서 *는 95% 신뢰수준을 통과함을 의미.

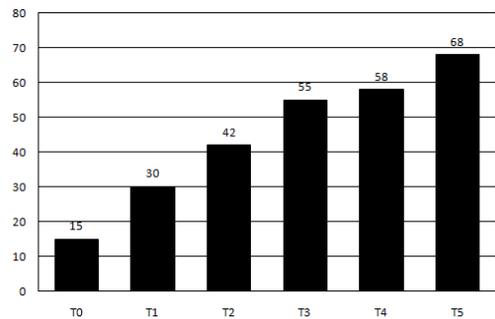
이렇게 정의된 소인적 및 가능 요인으로 구성된 데이터 레코드를 바탕으로, 동일한 조건을 갖는 데이터 레코드들을 모아 사례로 정리하였으며, 이때 각 사례에 해당하는 사람 수와 실제 고혈압자 비율에 해당하는 확률 값을 계산하여 저장하였다.

전체 사례들 중에서 고혈압 고위험군 대상자를 추정하는데 실제로 사용될 유효 사례를 정의하기 위하여 두 가지 기준을 정의하였는데, 각 사례 별로 (1) 사례에 해당하는 사람 수가 최소 10건 이상으로 얻어져 일반적으로 사용될 수 있는 수준이면서, (2) 사례에 해당하는 사람들 중에서 실제 고혈압 보유자의 비율이 27%(고혈압 자연 발생률 : 전체 데이터 웨어하우스상에서 조사된 값) 이상으로 얻어진 경우를 유효 사례로 분류하였다.

시간이 지남에 따라 외부 환경의 변화가 발생하

고, 그 변화의 결과로 사람들의 고혈압 관련 질환 역시 달라진다. 본 모형에서 제안하는 고혈압 고위험군 탐지를 위한 시스템 프레임워크가 이러한 변화를 감지하고 반영할 수 있다는 것을 보여주기 위하여 다음과 같은 실험을 수행하였다.

- <실험 1> (사례 발생) 최초 시점에서(T0) 사례 저장소에 생성되지 않았지만, 시간이 지남에 따라 새롭게 발생할 수 있는 유효 사례들이 얼마나 발생하는지를 관찰함. 만약 시간이 지남에 따라 신규로 생성되는 유효 사례가 관찰된다면, 고혈압 고위험군 탐지를 위한 시스템에서는 이것을 반영한 결과를 제공하는지는 확인해야 함
- <실험 2> (사례 변화) 최초 시점에서(T0) 사례 저장소에 생성된 유효 사례들이 시간이 지남에 따라 계속 유효 사례로 남게 되는지, 아니면 덜 중요한 사례가 되어 유효 사례에서 탈락하게 되는지를 추적하고 변화를 관찰함. 만약 시간이 지남에 따라 유효 하지 않은 사례로 분류되는 사례가 관찰된다면, 고혈압 고위험군 탐지를 위한 시스템에서는 이것을 반영한 결과를 제공하는지는 확인해야 함



[그림 6] 시간이 지남에 따라 사례 저장소에 기록된 유효 사례가 증가함

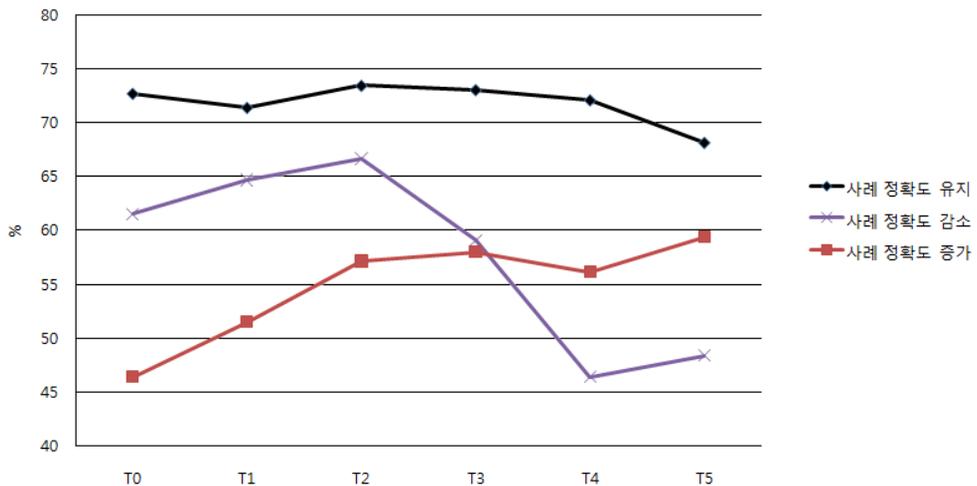
먼저 실험 1을 수행하기 위하여, 전체 데이터 발생 기간 18개월을 3개월씩 총 6개로 구분하여, 최초 관측 기간부터 T0, T1, ..., T5로 설정하였다. 매 시점 별로 유효 사례로 판정된 사례의 수는 [그림

6]과 같이 얻어진다. 초기 T0시점에서는 최초 3개월 동안 발생한 고혈압 및 관련 질환 이력을 바탕으로 유효 사례를 탐색하였으며, 그 결과 신규 15개 사례가 사례저장소에 기록되었다. 다음으로 T1시점까지 추가적으로 발생한 데이터를 사용하여 유효 사례를 탐색해준 결과, 총 30개의 유효 사례가 사례저장소에 기록된 것을 확인하였다. 유사하게, T5시점까지 수집된 데이터를 사용하여 유효 사례를 탐색해준 결과, 최종적으로 68개의 유효 사례가 사례 저장소에 기록된 것이 관찰되었다.

[그림 6]에서 살펴 볼 수 있듯이, 외부 환경 변화로 인하여 추가적으로 만들어지는 유효 사례는 시간이 지남에 따라 크게 증가한다. 그렇기 때문에, 만약 T0시점까지만 수집된 정보를 바탕으로 사례기반추론을 수행하고자 한다면, 그 이후 시점에서 발생한 신규 사례들을 반영하지 못하여 고혈압 고위험군을 효과적으로 탐색할 수 없게 된다. 즉, T0시점의 15개의 사례가 T5시점에서도 여전히 유효한 사례라고 하더라도, T5시점에서 사례저장소에 학습된 68개에서 T0시점에서 보고된 15개의 유효 사례를 제외한 나머지 53개에 해당하는 사례를 신규 사례 학습을 통해서 사례저장소에 기록하지 못하였다면, 이들 53개에 의해서 탐지되는

고혈압 고위험군 대상자들은 시스템에 의하여 추천을 받지 못하게 된다. 이러한 결과는 시간이 지남에 따라 사례 학습을 통하여 사례저장소에 최신 정보들을 반영해주는 것이 중요하다는 점을 시사한다.

일반적으로 T시점에서 T+1시점으로 변화할 때, T시점에서 이미 존재하던 사례의 정확도는 (1) 유지되거나, (2) 낮아지거나, (3) 높아지게 된다(단, T+1시점에서 신규로 생성된 사례는 정확도가 높아지게 되는 것으로 생각할 수 있음). [그림 7]에서는 시간의 변화에 따른 사례 정확도의 변화를 보여주기 위한 세 가지 사례에 대한 분석 결과를 설명하고 있다. 마름모 모양 (◆)으로 연결된 사례의 정확도는 T0시점에서 시작하여 시간이 지날수록 낮아지거나 높아지는 소폭의 변화를 보이면서, 대체로 정확도에 큰 변화가 없는 것으로 관찰된다. 반면에, X표 모양 (X)으로 표기한 두 번째 사례는, T0부터 T2까지는 사례 정확도가 소폭 증가하다가, T3부터 T4시점 동안 크게 하락하여, 최종 T5시점에서는 정확도가 50% 이하로 관찰되었다. 마지막으로 네모 모양 (■)으로 표기한 사례는 T0부터 T5시점까지 정확도가 증가하는 추세를 보이며, 최종 T5시점에서는 거의 60% 수준의 정확도를 보이



[그림 7] 시간에 따른 사례 정확도의 변화

는 것이 관찰되었다. 이러한 분석 결과는 각 시점마다 동일한 사례이지만 외부 환경 변화에 의해서 정확도가 달라질 수 있다는 것을 보여준다. 실제로 T0시점에서 임의로 집계된 40개의 사례에 대해서 T5시점에서의 정확도 변화를 집계한 결과, 유지의 기준을 $\pm 10\%$ 로 정의하였을 때, 감소, 유지, 증가의 비율이 각각 15%, 55%, 30% 수준으로 얻어졌다. 이는 사례의 정확도가 감소하거나 증가하는 것으로 변화하는 비율이 유지되는 비율과 거의 유사한 수준이라는 것을 보여주며, 시점에 따라 사례 정확도의 변화를 관리하는 것이 바람직하다는 것을 지지한다.

실험 결과를 고혈압 고위험군 관리라는 연구 목적에 따라 해석해보면 다음과 같은 시사점이 얻어진다. 시간이 지남에 따라 새롭게 학습되는 사례들은, 과거에는 고혈압에 영향을 미치지 않는 요소라고 생각되었지만, 최근 들어 새롭게 나타나는 고혈압 관련 질환에 대한 규칙이기 때문에, 시스템에서 자동적으로 변화를 감지하여 새로운 정보가 반영될 수 있도록 하는 것이 중요하다. 다음으로, 시간이 지남에 따라 고혈압 고위험군 예측 정확도가 변화할 수 있기 때문에, 시스템에서 이러한 변화를 감지하여 효과적으로 사례를 관리해야 한다. 즉, 과거 시점에서 고혈압 고위험군을 탐지하는데 높은 정확도를 보였던 사례의 경우라도, 시간이 지남에 따라 사람들의 식습관 변화 및 의료 혜택과 관련된 정책의 변화, 의료 기술의 발달로 인하여 더 이상 그러한 사례에 의해서는 고혈압에 걸리지 않게 될 수도 있으므로, 고혈압 고위험군 탐지라는 목적에 적합한 정보가 시스템 내부에서 유지되도록 해야 하는 것이 중요하다.

외부 환경 변화를 반영한 결과를 제공한다는 것이 확인되었다는 것과 별도로, 고혈압 고위험군 추정이라는 예측 문제를 잘 수행하는지 여부도 평가되어야 한다. 성능 평가를 위한 기준으로 정분류율과 민감도를 고려할 수 있는데, 실제 고혈압인 사람들을 정확히 분류하는 것이 보건기관의 목표와 부합하므로, 민감도를 높이는 것에 비중을

더 주었다. 즉, 모형의 성능을 높이기 위해서는 민감도 값을 만족할만한 수준을 유지하면서, 고혈압 예측 결과 중 실제 고혈압 환자 비율이 높은 모형을 만드는 것이 중요하다.

제안된 고혈압 고위험군 탐색 방법론을 바탕으로 지역 보건소에서 수집된 의료 데이터를 이용하여 고혈압 고위험군 대상자 관리를 효율적으로 하기 위한 유효사례 기준을 강화하고자 하였으며, 이 때, 데이터 웨어하우스 상에서 보고된 고혈압 자연 발병률 대비 25% 향상된 값으로 유효 사례 선정 기준을 상향 조정하였다. 다음으로, 각 시점마다 고혈압 환자 정보로 구성된 평가용 데이터를 사용하여 고혈압 환자를 올바르게 예측하는 비율을 측정해준 결과, 실제 고혈압인 사람들 중에서 고혈압 고위험군 대상자로 분류되는 비율이 평균 68.2% 수준으로 나타났으며, 모형에 의한 고혈압 추정자 중 실제 고혈압 보유자 비율은 약 45.4%로 얻어졌다.

본 연구에서 사용된 데이터는 국내 보건기관에서 개인을 대상으로 수집한 건강 상태 조사 결과이며, 서울과 경상도 지역에 소재한 보건기관에서 집계되었다. 관찰기간은 2007년 6월부터 2008년 12월까지 18개월 동안 약 80만 건의 건강 이력 데이터(소인적 요인 및 가능 요인)가 수집되었다. 데이터 수집 과정에서 개인 프로파일 정보는 제공받지 않았으며, 개인을 식별할 수 있는 정보는 암호화 처리가 되었기 때문에 개인 정보 보호와 관련된 문제가 발생하지 않도록 하였다.

5. 결 론

본 연구에서는 보건기관이 효율적으로 고혈압 관리 대상자를 탐색하고, 고혈압 관련 요인에 대한 지식을 효과적으로 관리할 수 있도록 하는 고혈압 고위험군 추정 모형 및 우선 사업 대상자 탐색 모형을 제안하였다. 특히, 대용량 데이터 처리 및 실시간 시스템 운영, 외부 환경 변화를 고려한 자동 학습과 같은 현실적인 제약 조건을 해결하

는 모형을 개발하는 것을 주목표로 하였다. 실험 결과에서 알 수 있듯이, 시간 변화에 따라 사례추론모형에서 사용되는 유효 사례들의 변화 수준이 크다는 것이 확인되었고, 이러한 변화들이 매 시점마다 반영되어 최신 정보를 바탕으로 고혈압 고위험군 탐색을 할 수 있도록 시스템을 설계하였다. 최종적으로 개발된 사례추론기반 데이터웨어하우스를 대상으로 고혈압 환자 정보로 구성된 평가용 데이터를 사용하여 고혈압 환자를 올바르게 예측하는 비율을 측정한 결과, 시스템을 사용하지 않았을 때보다 약 2배 정도로 향상된 고혈압 환자 예측력이 보고되었다. 이는, 모형을 사용하지 않을 경우보다 약 2배 효율적인 추정 결과를 얻으면서도, 전체 관리 대상자의 68.2% 수준으로 우선 대상자의 범위를 축소할 수 있다는 것을 의미한다. 또한, 시스템 유지보수 측면에서도 대용량 데이터 처리 문제 및 외부 환경 변화에 강인한 자동학습을 가능하게 하는 방법론을 제안하였다. 향후 유사 질환에 대해서도 본 방법론의 확대 적용이 가능하다.

연구 한계점에 대해서는 데이터의 한계와 방법론의 한계로 구분하여 정리할 수 있다. 먼저, 실험에 사용된 데이터가 특정 보건 기관에 한정되어 수집된 것이기 때문에 고혈압 고위험군 관련 사례를 생성함에 있어서 정보량이 충분하지 않을 수 있으므로, 더 많은 수의 보건기관에서 데이터를 수집하여 추가 실험이 수행되어야 한다. 이 때, 고혈압 고위험군을 탐지하는데 유효한 사례가 매 시점 별로 증가하거나 감소하는지에 대한 변화를 분석할 수 있어야 하며, 분석을 통해 얻어진 시사점을 바탕으로 일반적인 수준에서 사례 관리에 대한 방침을 수립할 수 있도록 해야 한다. 현재 연구 결과에서는 이러한 내용이 반영되지 못하였으므로, 추가 데이터 수집을 통하여 문제가 개선되어야 한다. 그리고 방법론 측면에서는, 개발된 시스템이 제공해주는 의료정보가 유용한 것인지 전문가 및 담당기관 실무자로부터 충분한 검증을 받지 못한 점이 문제점으로 지적될 수 있다. 이를 개선하기

위해서는 전문가에 의한 정성적 평가가 이루어져야 할 것이며, 실제로 지역 주민을 대상으로 서비스한 결과가 효과적인 것으로 나타나는지 실증해보아야 한다.

참 고 문 헌

- [1] 김용권, 허제연, 박은경, 진영수, “사례기반추론을 활용한 운동처방지원 시스템의 구축”, 『한국체육학회지』, 제41권, 제3호(2002), pp.351-358.
- [2] 김태수, 채영문, 김도마, 조승연, 윤진희, “고혈압 관리를 위한 의학적 의사결정지원 시스템의 데이터마이닝 접근”, 『한국지능정보시스템학회학술대회』, 제2권(2002), pp.203-212.
- [3] 박지우, 지원철, “사례기반추론을 이용한 데이터 웨어하우스 설계방법론”, 『대한산업공학회/한국경영과학회 공동학술대회』, 춘계학술대회논문집, (1998), pp.1-6.
- [4] 박한샘, 조성배, 이흥규, “베이지안 네트워크를 이용한 대사증후군 예측 모델링”, 『한국정보과학회학술대회』, (2005), pp.292-294.
- [5] 호승희, 채영문, 조승연, 최동훈, 송용욱, 박충식, 조경원, 송지원, “데이터마이닝 기법을 활용한 고혈압 관리를 위한 의사결정 지원시스템의 개발”, 『한국지능정보시스템학회 학술대회』, 제1권(2001), pp.271-281.
- [6] 홍태호, 박지영, “사례기반추론을 이용한 다이렉트 마케팅의 고객반응예측 모형의 통합”, 『정보시스템연구』, 제18권, 제3호(2009), pp. 375-399.
- [7] Aamodt, A. and E. Plaza, “Case-Based Reasoning : Foundational Issues, Methodological Variations, and System Approaches”, *AI Communications*, Vol.7 No.1(1994), pp.39-59.
- [8] Geuse D. A. and N. B. Miller, “Strategies for Medical Knowledge Acquisition”, *Medical Informatics*, Huston, Springer, (1997), pp.

- 277-292.
- [9] Inmon, W. H., *Building the Data Warehouse*, John Wiley and Sons, Inc, 1996.
- [10] Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. *The Sixth Report of the Joint National Committee on Prevention, Detection, and treatment of High Blood Pressure (JNC VI)*. Arch Intern Med, Vol.157(1997), pp.2413-2446.
- [11] Kahn, C. E., *Case-based Selection of Diagnostic Imaging Procedures*, In Leake, D. B., ed., *Case-Based Reasoning : Papers from the 1993 Workshop*, AAAI Press, Menlo-Park, CA, 1993.
- [12] Kahn, C. E. and G. M. Anderson, "Case-Based Reasoning and Imaging Procedure Selection", *Investigative Radiology*, Vol.29(1994), pp.643-547.
- [13] Katia G. C., L. T. Maria, W. Rosina, M. Alejandro, and M. R. Barcia, *Designing Nutritional Programs with Case-Based Reasoning*, IEEE Computer Society Press, Los Alamitos, CA, 1998.
- [14] Kolodner, J. L., *Case-based Reasoning*, Morgan Kaufmann, Los Altos, CA, 1993.
- [15] Robert, J., "Clinical Decision Support System", *Electronic Health Records : Chaining the Vision*, Philadelphia, WB, Saunders, (1999), pp.305-315.
- [16] Turban, E., *Expert Systems and Applied Artificial Intelligence*, Macmillan, New York, N.Y., 1992.
- [17] Wang, R. Y. and D. Strong, "Beyond Accuracy : What Data Quality Means to Data Consumers", *Journal of Management Information Systems*, Vol.12, No.4(1996), pp.5-34.
- [18] Watson, I. and F. Marir, "Case-Based Reasoning : A Review", *The Knowledge Engineering Review*, Vol.9, No.4(1994).

◆ 저 자 소 개 ◆

**박 성 혁 (heypark@business.kaist.ac.kr)**

한국과학기술원(KAIST) 수리과학부에서 학사를 취득하였고, 현재 한국과학기술원 경영대학(KAIST Business School) 경영공학과에서 박사과정 중에 있다. IEEE ICDE(International Conference on Data Engineering), HICCS (Hawaiian International Conference on System Sciences) 국제학술대회 논문발표 및 한국경영과학회지, CRM연구 등의 국내학술지에 논문을 게재한 바 있다. 주요 관심분야는 비즈니스 인텔리전스, 소셜 네트워크 서비스, CRM(Customer Relationship Management), 데이터마이닝 등이다.

**양 근 우 (kwyang@kmu.ac.kr)**

현재 계명대학교 전자무역학과 조교수로 재직 중이며, 한국과학기술원 경영대학에서 경영공학 박사를 취득하였다. 한국경영과학회지, 경영정보학연구, 통상정보연구, 해운물류연구, 인터넷전자상거래연구, 한국지능정보시스템학회논문지, Journal of Database Management, Expert Systems with Applications 등 국내외학술지에 논문을 게재한 바 있다. 주요 관심분야는 웹 2.0, 무역 프로세스 자동화, KMS 등이다.