韓國컴퓨터情報學會 論文誌
第14卷 第3號, 2009. 3.

2009-14-3-4-5

약물부작용감시시스템에서 재현성 평가를 통한 마이닝 모델 개발

이 영 호*, 윤 영 미*, 이 병 문*, 황 희 정*, 강 운 구*

Development of Mining model through reproducibility assessment in Adverse drug event surveillance system

YoungHo Lee*, YoungMi Yoon*, ByungMun Lee*, HeeJoung Hwang*, UnGu Kang*

요 약

약물부작용감시시스템(Adverse drug event surveillance system)은 약물부작용신호를 이용하여 약물의 부작용 여부를 식별하는 시스템이다. 기존의 자발적 보고나 차트리뷰 보다 효율성이 뛰어난 시스템으로 분류할 수 있다. 본 논문에서는 약물부작용감시시스템을 구현하기 위하여 임상데이터마트(CDM)를 구축하였다. 특히, 데이터 품질관리 기법을 적용하여 구축된 CDM에 지식 탐사 기법중 비교사학습 기법으로 적용하여 모델의 재현성을 평가하여 최적의 약물부작용 군집화 개수(n=4)를 도출하였다. 군집화 개수(n=4)를 이용하여 약물부작용 판별을 위한 K-means, Kohonen, two-step clustering model 알고리즘에 적용하여 분석함으로써 K-means 알고리즘이 가장 우수한 군집 효과를 나타냄을 확인하였다.

Abstract

ADESS(Adverse drug event surveillance system) is the system which distinguishes adverse drug events using adverse drug signals. This system shows superior effectiveness in adverse drug surveillance than current methods such as volunteer reporting or char review. In this study, we built clinical data mart(CDM) for the development of ADESS. This CDM could obtain data reliability by applying data quality management and the most suitable clustering number(n=4) was gained through the reproducibility assessment in unsupervised learning techniques of knowledge discovery. As the result of analysis, by applying the clustering number(N=4), K-means, Kohonen, and two-step clustering models were produced and we confirmed that the K-means algorithm makes the most closest clustering to the result of adverse drug events.

▸ Keyword : 약물부작용감시시스템(Adverse Drug Event surveillance system), 지식탐사(Knowledge discovery), 임상데이터마트(Clinical Data Mart)

• 제1저자 : 이영호 교신저자 : 강운구
• 투고일 : 2009. 2. 10, 심사일 : 2009. 2. 17, 게재확정일 : 2009. 3. 23.
* 가천의과학대학교 의료공학부 IT학과 교수

Ⅰ. 서 론

의학 관련 각종 보고서에 의하면 의료사고(Medical Error)의 많은 부분이 약물부작용에 의해 발생하며, 각종 약물과 관련한 사고는 환자의 죽음과 직결되는 문제로 그 중요성이 매우 크다. 미국 보건부(Department of Health & Human Services)의 조사에 의하면 연간 의료 사고로 인한 사망자 대부분이 약물 관련 부작용에 의한 것이라는 점에서 심각한 문제로 인식되고 있다[1].

현재 활용되고 있는 약물부작용 감시 방법에는 자발적인 보고(Volunteer reporting)방법, 의무기록조사(Chart Review) 방법, 약물부작용감시시스템(Adverse Drug Event Surveillance System) 방법 등이 있으며, 이러한 감시 방법에 의한 약물부작용 발견율을 살펴보면 자발적인보고 4%, 의무기록조사 65%, 약물부작용감시시스템 45%를 기록하였다고 보고 하였다[2].

약물부작용 감시 방법 중 자발적인 보고 방법은보고율이 저조하여 실제 부작용 탐지율이 희박하고, 차트조사는 후향적 의무기록조사(Retrospective chart review)를 바탕으로 하기 때문에 대규모의 의료진이 투입되고, 이로 인한 많은 비용이 지출되어 약물부작용을 감시하는 체계로는 적합하지 않다.

약물부작용감시시스템은 약물부작용 신호(ADE signal)를 이용하여 부작용 대상을 판별하는 시스템을 말한다. 약물부작용감시시스템은 다른 방법보다 많은 장점이 있다. 첫째, 의료진의 노동력이 아니라 전자적인 절차(electronic process)를 통해 운영함으로써 비용, 시간적인 측면의 자원 소모를 줄일 수 있다. 둘째, 단속적이고 일회성 부작용 감시가 아니라 연속적이고 병렬적인 부작용 탐지 및 추출이 가능하다. 셋째, 광범위한 데이터 조사를 통해 약물 부작용을 조기에 발견함으로써 환자의 약물 부작용에 대한 합병증을 추가적으로 예방할 수 있다. 결국 임상현장에 다양한 정보기술을 활용함으로써 약물 부작용 발견에 소요되는 비용과 시간을 현격히 감소시켜 약물부작용 탐지를 용이하게 할 뿐만 아니라 자발적 보고나 챠트 리뷰에 의존하지 않고서도 부작용을 탐지하는 것이 가능하다.

본 연구에서는 다양한 약물부작용 데이터 추출을 위하여 기존 연구 등에서 제안된 실시간 기반의 부작용 신호 추출 방식이 아니라 사전에 부작용신호 유형을 추론엔진으로 정의하고 신호 감지 시점을 구분하여 구축함으로써 병원정보시스템과 전반적인 시스템 운영 안정성에 기반을 둔 약물부작용감시시스템에서 부작용 신호 탐지를 위한 기술 구조를 제시 한다. 이렇게 구축된 시스템을 통해 약물 부작용 여부로 판별 한 후 구축된 데이터 중 약 처방 결과 데이터를 중심으로 지식탐색의 비교사학습(Unsupervised Learning)방법 중 Two-step, K-means 알고리즘(algorithm)을 통해 재현성(Reproducibility)을 평가하였다. 그 결과 최적의 군집(cluster) 개수(n=4)를 도출하였고, 세그먼테이션 그룹을 각 알고리즘별로 분석하여 약물부작용 판별 여부를 조사하였다. 본 논문의 구성은 2장에서 약물부작용 탐지에 대한 관련연구를 분석하고, 3장에서는 약물부작용감시시스템에서 재현성 평가를 위한 마이닝 모델을 제안한다. 4장에서 실험을 통해 제안 모델을 검증하고, 5장과 6장에서는 실험 결과와 향후 연구를 기술한다.

Ⅱ. 관련 연구

미국 식품의약국(FDA)은 1998년부터 약물부작용보고시스템(Adverse Event Reporting System)을 가동하고 있는데, Thomas J. Moore 등의 논문에 의하면 약물부작용 중 심각한 약물부작용(Serious Adverse Drug Events)은 1998년 34,966건에서 2005년 89,842건으로 약 2.6배 증가하였으며, 생명을 위협하는 치명적인 약물부작용(Fatal Adverse Drug Events)은 1998년 5,519건에서 2005년 15,107건으로 약 2.7배 증가한 것으로 나타났다.

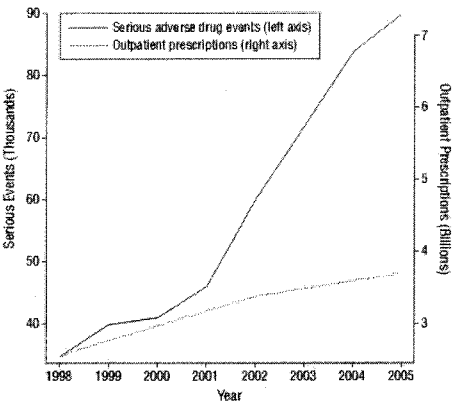

그림 1. 1998-2005년 사이 보고된 약물부작용과 외래처방 사례

Fig. 1. Reported Serious Event VS Outpatient Prescriptions 1998-2005.

또한 그림 1.같이 외래환자에 대한 약물 처방률이 증가하면서 그에 비례하여 심각한 약물부작용 또한 증가한 것으로 나타났다[3].

이와 같은 약물부작용의 증가로 의약품 안전성을 제고하기 위하여 RFID 시스템 도입 등 다양한 연구가 이루어지고 있으며[4] 특히, 약물에 대한 관심이 높아지면서 약물부작용 탐지에 대한 연구가 활발히 진행되고 있다.

약물부작용 탐지를 위한 연구를 살펴보면, Bates 등은 약물부작용탐지 시스템을 결과 측정(outcome measured)과 자동화수준(level of automation) 등을 비교하여 평가하였으며, James 등은 전자화된 시뮬레이션 모델 구축을 통해 약물부작용 감시의 효율성을 평가하여 약물부작용 감시시스템의 비용 대비 효율성을 입증하였다. 또한 Peter 등은 약 처방관련 데이터베이스 모델을 개발하여 CPOE(Computer physician order entry) 운영 단계에서 약품 유형별 처방용량의 계산, 처방단위 환산, 과용량 처방 시 전자적 경고 기능의 탑재 등의 데이터베이스 설계 모델을 제안하였으며[5], Wilson과 Kusiak등의 연구에서는 제약 산업 분야에 적용할 수 있는 데이터웨어하우징 기술 분야를 제안하였다[6][7].

최근 서울의 A병원에서는 환자의 약물 알레르기 정보의 입력 과정과 알레르기의 원인 약물을 확인 하는 과정 및 확인 과정에서 치명적인 알레르기의 원인으로 확진된 약물의 차단 과정으로 구성된 약물알레르기 경보시스템을 개발하여 운영하고 있다[8].

이상 대부분의 기존 연구 사례는 약물부작용감시시스템과 같은 의사결정지원시스템을 지원하기 위한 정보시스템 구조로 실시간 기술구조를 지원하고 있다[9][10]. 실시간 처리 구조는 병원정보시스템에 특정 이벤트가 발생 시 이를 인지하여 즉각적으로 조치를 취할 수 있는 가장 이상적인 시스템이다. 그러나 실시간이벤트(Real-time event) 감시를 위한 작업은 대상 이벤트수 또는 시스템에 발생하는 트랜잭션 양에 따라 이벤트 감시에 따르는 비용이 기하급수적으로 증가한다[11]. 본 연구에서 제안하는 약물부작용감시시스템의 기술구조는 OLTP, OLAP(On-Line Analytical Processing)의 이중 구조(Dual-mode)의 운영을 통해 병원정보시스템의 안정성과 약물부작용감시시스템의 본연의 기능 확보라는 목표를 달성하고자 하며 이 기반에서 시스템 개발을 추진하여 결과 데이터를 분석하고자 한다.

III. 약물부작용감시 시스템

일반적으로 약물부작용감시시스템에서 부작용 감지의 출발은 약물부작용의 가능성이 있는 환자를 찾아내기 위한 방법에서 시작되며 이때 Clinical event monitor(이하 CEM)라고 부르는 정보기술구조를 활용한다. 그림 2와 같이 CEM에서 약물부작용을 시사 하는 조건이 컴퓨터 알고리즘으로 내장되어 환자의 의무기록이 병원정보시스템 또는 전자의무기록 시스템에 기록될 때마다, 해당하는 환자의 정보에 알고리즘을 적용하여 사전 정의된 조건에 해당하는 환자에 대하여 약물부작용 경고를 발생한다. 본 연구에서는 약물부작용감시시스템의 결과를 CEM에 내장된 알고리즘으로 뿐만 아니라 데이터 마이닝의 클러스터링 기법을 통해 약물부작용을 판별하였다.

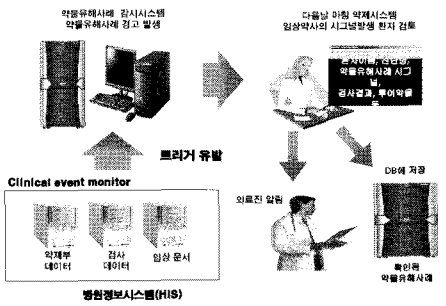

그림 2. 악물부작용감시시스템 처리 과정
Fig. 2. Adverse Drug Event Surveillance System Process

1. 구축 환경

구축 환경은 ER-win4.1을 이용하여 데이터베이스를 설계하였고, HIS의 데이터복제를 원활하게 지원하기 위하여 CPU 4way, Memory 4G, Storage 750G의 환경으로 데이터베이스는 HIS 동일한 환경의 Oracle 10g 환경으로 구축하였다. 데이터의 추출 및 변환 환경은 ETL(Extract Transformation Loading)을 적용하기 위하여 MS-SQL Server의 DTS(Data Transfer Service)를 활용하였다. CPU는 2way, Memory 2G, Storage 1.2T 환경으로 분석을 위한 실험환경을 구성하였다. 데이터 전처리 및 데이터 마이닝을 위한 Tool은 Clementine 12.0을 활용하였다.

2. 데이터전처리

약물부작용감시시스템의 기반이 되는 병원정보시스템의 특징은 의료진, 연구자, 행정 직원 등 여러 사용자의 접근이 가능하여 결국 다양한 입력 사용자로 하여금 데이터 원천 발생 시점에서부터 많은 문제점을 내포하게 하고 있다. 본 실험에서는 이 문제점을 극복하기 위하여 데이터품질관리 방법론을 적용하여 데이터의 전처리를 수행하였다. 본 연구에서는 데이터품질관리를 처리하는 프로세스 중 Data Cleansing(정제), Data Deriving(파생), Data Screening(검토)과정을 수행하였다.

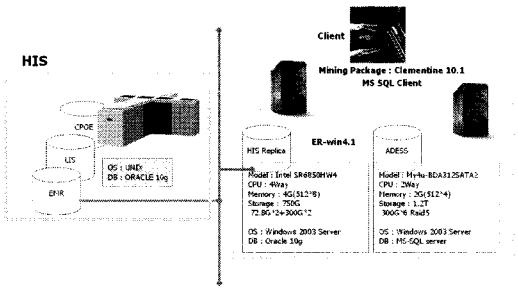

그림 3. 실험 환경
Fig. 3. Experimental environment

2.1 Data Cleansing

본 연구에서는 데이터 품질관리의 첫 번째 과정으로 Data Cleansing(데이터정제) 기준을 적용 중복항목과 데이터의 누락 항목의 제거 과정을 수행하였다. 우선 약 처방 전체 데이터 중에서 데이터 없는 추가포수, 처방반복구분, 투약개시일자 필드 3개를 제거하였다. 또한 필드의 충실도 50% 기준으로 정의하고 그 이하의 17개 필드 제거하였다(투약시간 1~6, Repeat월요일~일요일, 조정량 1~4). 검사 처방 전체 데이터 중에서 데이터 없는 필드 검사결과단위 1개를 제거하였다. 데이터정제 이후의 ERD(Entity Relationship Diagram)는 다음과 같다.

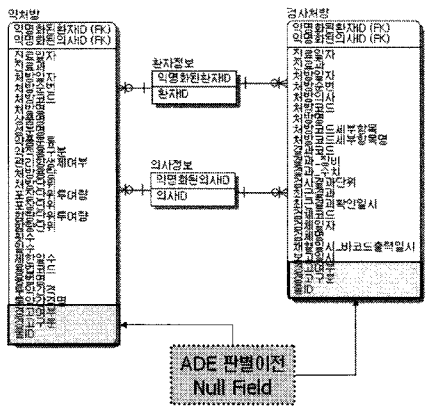

그림 4. 데이터마트 논리 ERD
Fig. 4. Data Mart Logical ERD.

2.2 Data Deriving

데이터 파생은 데이터 품질관리 작업에서 중요한 과정 중에 한 단계이다. 이는 분석에 활용 가능한 속성으로 현재 필드를 추가 및 변형, 가공하는 작업이다. 우선 약 처방, 검사처방 테이블에서 생년월일, 입원일자, 처방일자, 결과보고일자 등 날짜 데이터 분석 가능한 형태의 문자열로 속성을 변경하였고 경고구분, 경고여부, 물ID의 NULL값 "0"값으로 파생시켜서 약물부작용의 최종판별 결과가 입력되도록 하였다.

2.3 Data Screening

데이터파생, 데이터정제를 통해 결정된 약 처방 테이블 29개 필드는 다시 데이터 필터 과정을 적용하여 참고정보 성격인 처방명, 상품명, 용법명 등 12개 필드를 분석 대상 필드에서 제거하고, 다음 9개 필드를 분석 필드로 결정하였고, 검사 처방데이터는 24개 필드 중에서 8개를 분석 필드로 확정하였다.

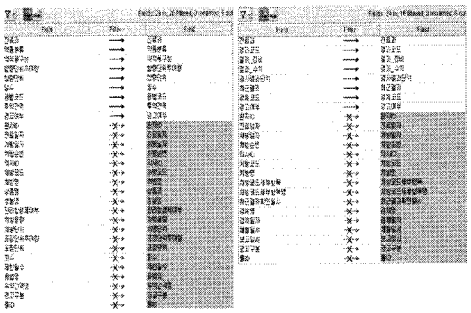

그림 5. 데이터 필드 추출
Fig. 5. Data Field Extraction.

IV. 실 험

지식탐사(Knowledge Discovery)프로세스의 핵심적인 역할을 담당하는 기법을 분류하는 방법은 데이터 형태에 따라 교사학습(Supervised Learning)과 비교사학습(Unsupervised Learning)으로 구분하는 것이다. 현재 가용한 데이터가 환자(혹은 처방)단위의 각 레코드 및 여러 가지 입력변수(Inputs)와 주관심이 되는 목표 혹은 결과변수(Target)로 이루어져 있는 경우 교사학습을 수행할 수 있다. 이에 반해 비교사학습 기법은 각 레코드가 환자(혹은 처방)단위로 여러 가지 입력 변수들만으로 이루어진 데이터로 구성되어 있을 경우를 학습을 수행 할 수 있다.

K-means, Kohonen, Two-step algorithm은 대표적인 비교사학습 기법이다. 비교사학습 기법은 기술(descriptive) 분석과 같은 탐색적 데이터 분석(exploration data analysis)을 통해 자료를 이해하는 것이 1차적인 목표이며 이중 비교사학습의 대표적 분석 방법인 군집분석(cluster analysis)은 많은 변수들을 일정한 특징에 따라 몇 개의 군집으로 분류하여, 같은 군집에 속한 개체들의 유사성과 특이성을 규명하는 방법이다. 비교사학습 기법의 군집분석에서

K-means와 Kohonen 알고리즘은 사용자가 군집의 수를 정한 후 이를 반복적으로 수행하여 최종 군집을 형성하며 , Two-step 알고리즘에서는 사용자가 정한 범위에서 군집화를 수행한다. 그러나 비교사 학습과 같은 군집분석 시 주의 할 점은 군집의 크기는 어떻게 할 것이며, 군집의 개수를 몇 개로 할 것이고, 그 군집결과의 타당성은 어떻게 제공할 것인가에 대한 사전 정의가 반드시 이루어져야 한다. 본 연구에서는 기존 약물부작용의 결과를 미리 판별하여 그 결과와 신뢰성 있는 군집갯수의 재현성 평가 알고리즘을 통해 각 클러스터별 특징을 분석하였다. 다음은 신뢰성 있는 군집 개수 확보를 위한 방법을 기술한다.

1. 재현성 평가 모델 검증

일반적으로 모델 검증을 위하여 교사학습 기법에서는 자료를 몇 개 데이터로 분할(partitioning)한 후 훈련(training) 데이터와 시험(test) 데이터로 구분하여 모델에 적용함으로써 대상 모델의 적합성 및 타당성을 검증할 수 있다. 잘 알려지지는 않았지만 이를 응용하여 비교사학습 기법에서도 자료를 분할하여 동일한 군집화 방법의 반복을 통해 재현성을 평가할 수 있다. 몇몇 연구에서 비교사학습 기법에 대한 cluster의 평가 방법이 제시되었다

① 주어진 자료를 임의로 2개로 분할한다. 그중에서 하나의 데이터 그룹을 자료 1이라 부르고 다른 하나를 자료 2라 정의한다.
② 자료 1에 대한 군집화(clustering)을 수행한다. 이때 생성된 모델을 모델1이라 하고 자료 2의 각 개체를 모델1에 따라 분류한다.
③ 자료 2를 동일한 방식으로 군집화를 수행하여 모델 2를 생성한다. 이에 따라 자료 2의 각 개체를 분류한다.
④ 자료 2의 개체들에 대한 모델 1과 모델 2의 분류 결과를 교차분류표로 나타내어 적용된 군집의 분포가 대응한다면(이때 순서는 상관없다.) 관련 행, 열이 강한 상관성을 보일 것이다. 그렇다면 적용된 군집의 분포를 적절한 것으로 판단하여 세그먼테이션 모델을 구축하여도 모델의 재현성이 확보되었다고 할 수 있다.

본 연구에서는 K-means와 Two-step 알고리즘의 재현성를 실험한다.

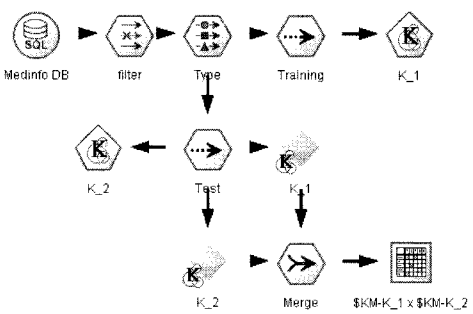

그림 6. 재현성평가를 위한 클레멘타인 노드
Fig. 6. Clementine Node For Reproducibility.

① 데이터의 분할

전체자료 중 절반을 뽑아 훈련데이터(training data)로 쓰기로 하였다. 훈련데이터에서는 체계적으로 짝수 번째 케이스를 삭제하였다. 시험데이터(test data)에서는 훈련데이터와의 중복을 피하기 위해 짝수 번째 케이스만 선택한다.

② 데이터의 군집화

우선 훈련데이터에 군집의 개수 6을 적용하고 여기에서 생성된 모형을 K-1로 정의하였다. 시험데이터에도 마찬가지로 군집의 개수 6을 적용하고 K-2 모델 군집화를 수행하였다. 계속해서 군집 개수를 5, 4로 적용하여 군집화를 수행한 후 비교분류표를 생성한다.

③ 비교분류표 생성

K-1, K-2로 2개의 군집결과를 병합하여 비교분류표를 만들어 각 개체를 분류한다. 군집의 개수를 6, 5, 4개 적용하였다.(표 1)

표 1. K-means 학습을 통한 재현성평가
Table 1. K-means Learning Reproducibility

# of cluster = 6

K=6		K-2						Total
		clus1	clus2	clus3	clus4	clus5	clus6	
K-1	clus1	53392	0	0	0	0	0	53392
	clus2	0	40033	0	0	1602	0	41635
	clus3	0	0	2036	0	0	0	2036
	clus4	0	0	8	5472	0	0	6215
	clus5	230	0	31	0	8	0	2339
	clus6	0	0	0	0	0	1216	1216
	Total	55692	40033	2075	5472	2345	1216	106833

# of cluster = 5

K=5		K-2					Total
		clus1	clus2	clus3	clus4	clus5	
K-1	clus1	53392	0	0	0	0	53392
	clus2	0	40033	0	0	1602	41635
	clus3	0	0	3252	0	0	3252
	clus4	0	0	0	5480	735	6215
	clus5	2331	0	0	0	8	2339
	Total	55723	40033	3252	5480	2345	106833

# of cluster = 4

K=4		K-2				total
		clus1	clus2	clus3	clus4	
K-1	clus1	55731	0	0	0	55731
	clus2	0	41635	0	0	41635
	clus3	0	0	3252	0	3252
	clus4	0	0	0	6215	6215
	Total	55731	41635	3252	6215	106833

K-means algorithm에 K-1과 K-2가 군집의 개수를 5개로 적용하여 재현성을 평가하였을 때는 cluster2 (40033:1602), 4(5480:735), 5(2331:8)에서 여러 개체가 주 경향에서 벗어 나 있다. 마찬가지로 군집의 개수를 6개로 적용하여 재현성을 평가하였을 때는 cluster2, 4, 5에서 여러 개체가 주 경향에서 벗어 나 있고 특히, cluster2에서는 상당히 일치성이 떨어지고 있다. 반면 군집의 개수를 4개(# of cluster = 4)로 정의하여 평가한 결과 전체 데이터는 전체가 대각선으로 주 경향에 일치하였고 있음을 알 수 있다.

따라서 K-means algorithm을 적용하여 모델1(K-1)과 모델2(K-2)의 분류 결과를 비교분류표로 생성하였을 경우 순서에 상관없이 군집 개수는 4개 일 때 관련 행, 열이 가장 강한 상관관계를 보임을 알 수 있다. 따라서 군집의 개수가 4개일 경우 적용된 군집 분포가 적절한 것으로 판단하여 세그먼테이션 모델을 구축하여도 모델의 재현성이 확보되었다고 할 수 있다.

표 2. Two-Step 학습을 통한 재현성평가
Table 2. Two-Step Learning reproducibility

# of cluster = 6

cluster		Two-step 2						Total
		clus1	clus2	clus3	clus4	clus5	clus6	
Two-step 1	clus1	10	0	0	0	0	0	10
	clus2	0	10283	15929	55	0	0	26267
	clus3	0	9583	15881	3748	0	0	29212
	clus4	0	0	0	3502	0	0	3502
	clus5	0	0	0	0	21794	7811	29605
	clus6	0	0	0	2	14272	3963	18237
	Total	10	19866	31810	7307	36066	11774	106833

# of cluster = 5

cluster		Two-step 2					Total
		clus1	clus2	clus3	clus4	clus5	
Two-step 11	clus1	10	0	0	0	0	10
	clus2	0	10283	15929	55	0	26267
	clus3	0	9583	15881	3748	0	29212
	clus4	0	0	0	3503	0	3503
	clus5	0	0	0	0	47841	47841
	Total	10	19866	31810	7306	47841	106833

# of cluster = 4

cluster		Two-step 2				total
		clus1	clus2	clus3	clus4	
Two-step 1	clus1	10	0	0	0	10
	clus2	0	50977	4499	0	55476
	clus3	0	0	3506	0	3506
	clus4	0	0	0	47841	47841
	Total	10	50977	8005	47841	106833

Two-step model(표 2)에서도 군집의 개수가 4개 이면 해당 군집화가 충분한 재현성을 가지고 있음을 비교분류표에서 확인 할 수 있다

2. 비교사 학습 모델의 적용

본 연구에서는 절대적 군집의 크기에 대한 고려사항은 배제 하였고 군집에 개수, 군집의 타당성 확보에 무게를 두었다. 우선 한 달간의 약 처방 결과인 213,666건의 데이터를 기준으로 재현성 평가를 통해 확보된 군집수 4개가 가장 최적 군집의 개수임을 확인하였다. 군집화 방법을를 통한 결과 도출 시 특히 주의할 점은 사용되는 개체들은 원칙적으로 연속형 변수(continuos) 이여야 한다는 점이다. 만일 입력변수가 Yes, No와 같은 이항형인 경우 0과 1(혹은 1과 2)로 재 코딩하고, 다항 범주형인 경우에는 더미(dummy) 코드화하여 사용하여야 각 개체들 간에 거리 계산이 가능하다. 현재 앞서 사용된 변수들을 입력변수로 사용하기 위해서는 함량단위투여량, 투여횟수, 투여일수, 일일투여량(투여횟수*함량단위투여량)의 연속형 변수와 경고여부(이항형)를 입력방향으로 지정하여 알고리즘을 수행하였다. 이상과 같은 작업으로 통해 군집을 형성하였을 경우 각 군집의 특징이 확연히 구별될수록 좋은 학습기법이라 할 수 있다.

Clementine의 K-means Algorithm, Two-steps, Kohonen SOM, 에 다음과 같이 군집 개수의 재현성 평가 결과를 토대로 4개로 지정하여 실험을 수행하였다.

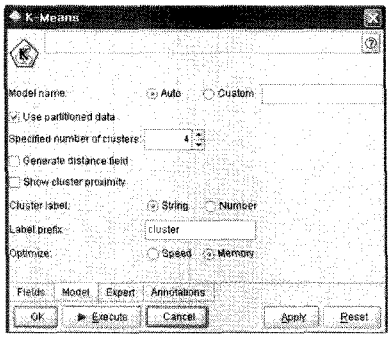

그림 7. K-means 클러스터링 마이닝 모델
Fig. 7. K-means clustering mining model

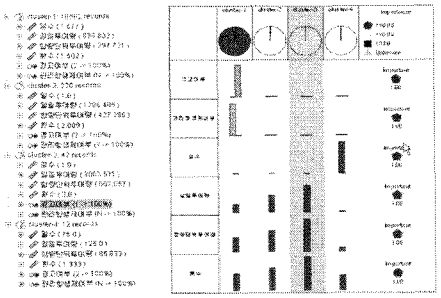

그림 8. K-means 모델 결과
Fig. 8. As K-means model results

V. 실험 결과

데이터의 분석 일관성 확보를 위해 전체 처방 데이터를 투여단위(AMP, VIA, TUB등)로 분류하여 가장 일반적인 투여단위인 밀리그램(mg) 데이터를 군집화 하였다. 전체 약처방 데이터(213,666)중에서 약 8%에 해당하는 데이터가 밀리그램으로 구분되는 18,856건이 군집 대상 데이터 이며 지식베이스를 사용하여 과처방 여부를 확인 하였을 경우 약 처방에서 과 처방 비율은 전체의 213,666건의 0.1%에 해당하는 222건 이였고, 이중 투여단위가 밀리그램으로 확인 된 데이터는 18%인 42건 이였다.

1. K-means 결과

결과 군집의 특성은 다음과 같다. 대부분의 데이터(18,582건 - 98%)가 cluster-1에 위치하고 있어서 가장 군집의 규모가 크다. cluster-1은 경고여부가 정상인 케이스(100%)로 분류되어질 경우 관리항생제 사용대상에서 제외된 데이터(N=0, Y=1)는 일일투여량과 함량단위투여량이 다른 군집에 비하여 상대적으로 낮게 투여되었음을 알 수 있다. 그 외에 나머지 cluster-2,3,4는 투여량이 상대적으로 과도하게 많았으며, 특히 cluster-3은 투여량이 cluster-1에 비하여 과도하게 투여되었음을 확인 할 수 있다. 따라서 K-means clustering을 통해 각 군집의 특성을 확인할 경우 cluster-1을 제외한 나머지 그룹군은 투여량이 군집을 형성하는 주요한 원인임을 알 수 있다.

2. Two-step clustering 결과

Two-step clustering은 K-means나 Kohonen algorithm 보다 대량의 자료를 군집화 하는데 가장 효과적으로 적용 될 수 있는 방법이다. 다른 알고리즘이 자료를 반복처리하여 읽는데 반하여 Two-step clustering에서는 1개체를 1회 반복하여 읽기 때문에 상당히 컴퓨터 자원의 활용을 효율적으로 사용할 수 있다.

Two-step clustering 결과는 앞서 살펴본 K-means와는 다른 결과를 나타내고 있다. cluster-1이 전체 데이터(18,582)의 64%인 12,007건을 차지하여 가장 많은 비중을 차지 하지만 K-means algorithm과 같이 특별히 경고구분값을 구분하여 과 처방 데이터 군집을 식별하지는 못하고 있다. 다만 cluster-4 그룹이 처방횟수나 투여량 측면에서 다른 군집보다는 확연히 많이 처방되어 있음을 알 수 있다.

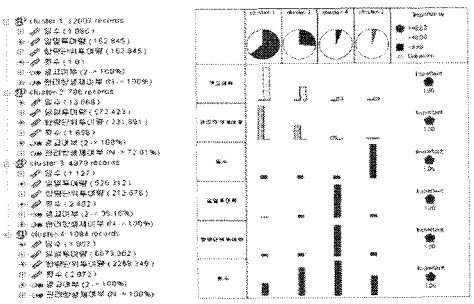

그림 9. Two-steps 모델 결과
Fig. 9. As Two-steps model results

3. Kohonen SOM 결과

Kohonen algorithm은 수많은 반복적 순회 계산을 하게 되어 K-means보다 많은 자원이 시간이 소모된다. 하지만 네

트워크상의 노드에 대한 격자(grid)를 통해 Self-Organizing Feature Map(SOFM)을 얻게 되는데 이 결과는 비선형적인 차원축소에 의한 다변량 개체들의 위상순차화(Topological ordering)로 확인 할 수 있다. 그림 10은 SOM을 형성하는 과정을 나타내고 있다.

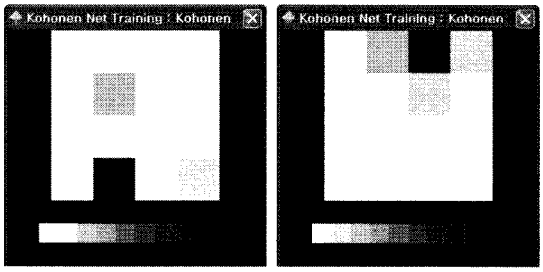

그림. 10. Kohonen 자기조직화지도
Fig. 10. Kohonen Self-Organizing Map(SOM)

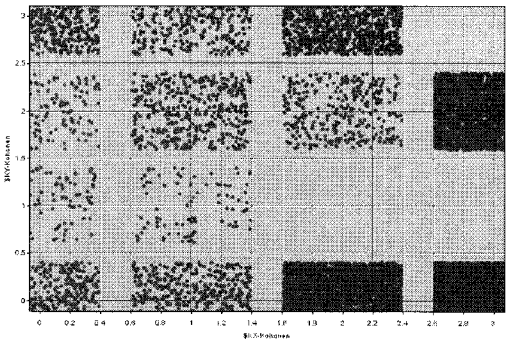

그림. 11. Kohonen 자기조직화지도 그리드
Fig. 11. Kohonen SOM Grid

Kohonen의 결과는 Grid와 결과표(그림 11)에서 보듯이 전체 대상 데이터(18,582)중에서 44%인 8,231개 케이스가 첫 번째 군집화(X=3, Y=0) Grid에 집중되어 있다. 이 군집은 투여횟수와 투여량이 경미하여 경고여부와는 상관이 없는 군집으로 분석할 수 있고 투여량이 다른 군집에 비하여 과다한 군집인 5 번째 군집은(X=0, Y=3) 그림 12처럼 경고여부를 대표할 수 있는 의미 있는 특징을 보이지는 않는다.

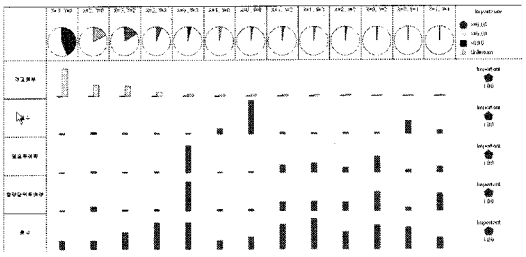

그림 12. Kohonen 모델 뷰어를 통한 결과
Fig. 12. As the result of Kohonen model viewer

VI. 결 론

약물부작용감시시스템은 약물부작용을 경험했을 환자들을 대상으로 정해진 기준인 약물부작용신호(adverse drug event)를 임상전문가의 지식을 바탕으로 지식베이스를 구축하여 판단 한 후 결과를 실제 의무기록과 함께 검토하여 판별하는 시스템으로 정의할 수 있다.

바람직한 약물부작용감시시스템 체계는 임상전문가의 의학적 판단을 통해 병원정보시스템에 기록된 환자의 정보 및 진단, 처방정보와 각종 의무기록지에 정의된 처치 자료 등을 종합하여 부작용 여부나 투약의 오류를 발견하는 것이 가장 정확한 과정이라 할 수 있다. 그러나 현실적으로 대형종합병원에서 하루에 처방되고 검사되는 각종 데이터 건수는 수천 건에 이르고 부작용 감시를 위해 이를 일일이 임상의사 나 약리학 전공자가 확인하기에는 무리가 있다. 따라서 다소 그 정확성이 떨어지더라도 합리적으로 약물부작용을 발견하고 감시하는 체계는 일반 병원에서 뿐만 아니라 공중의료분야에서도 시급히 선행되어야 하는 연구과제 일 것이다.

본 연구에서는 지식탐사 기법 중 비교사학습의 대표적인 K-means, Kohonen, Two-step 알고리즘을 적용을 재현성 평가를 통한 최적 군집수를 도출하였고 그 결과를 바탕으로 부작용 여부에 대한 군집이 분류 되는지는 확인하였다. Kohonen, Two-step 알고리즘에 비하여 별도의 부작용 군집의 특징이 확연히 구별될 정도로 K-means 알고리즘은 가장 성공적인 군집 결과를 나타내었다. 때로는 오래된 기술이 최신의 기술을 제압할 수 있는 것처럼 K-means 알고리즘은 다른 알고리즘에 비하여 가장 전통적이고 단순한 군집분석 방법 이지만 본 연구에서는 가장 효율적인 군집 결과를 도출하였다.

향후 연구는 본 연구에서 제시한 기술구조를 기반으로 다른 부작용신호의 적용을 검증하고 지식분류 기술을 활용하여

환자의 안전성 확보를 위해 필수적인 약물부작용감시시스템에 대한 확장 연구를 진행할 것이다.

참고문헌

[1] 범희승, 박성희, 최진욱, 김춘배, "임상의사결정지원시스템의 약제부작용 감소 효과에 관한 메타분석," 대한의료정보학회지, 제8권, 제2호, 55-60쪽, 2002년 11월.

[2] Ashish K. Jha, Gilad J. Kuperman, Jonathan M. Teich, Lucian Leape, Brian Shea, Eve Rittenberg, Elisabeth Burdick, Diane Lew Seger, Martha Vander Vliet, and David W. Bates, "Identifying Adverse Drug Events: Development of a Computer-based Monitor and Comparison with Chart Review and Stimulated Voluntary Report," J Am Med Inform Assoc Vol. 5, No. 3, pp. 305-314, May 1998.

[3] Homas J. Moore, AB et al, "Serious Adverse Drug Events Reported to the Food and Drug Administration," Arch Intern Med, Vol. 167, No. 16 pp. 1752-1759, September 2007.

[4] 최경아, 정인성, 유현선, 윤은실, 이영호, 강운구, "의약품 창고관리를 위한 RFID 시스템의 인식률에 관한 연구," 한국컴퓨터정보학회 학술발표논문집, 제16권, 제2호, 249-254쪽, 2009년 1월.

[5] Peter Martin, Walter E. Haefeli, and Meret Martin-Facklam, "A Drug Database Model as a Central Element for Computer-Supported Dose Adjustment within a CPOE System," J Am Med Inform Assoc Vol. 11, No. 5, pp. 427-432, June 2005.

[6] Kusiak A,Shah S, "Data Mining and Warehousing in Pharma Industry In J.Wang(ed.): Encyclopedia of Data Warehousing and Mining," Idea Group., Hershey, PA, pp. 239-244, Apr. 2006.

[7] Andrew M.Wilson, Lehana Tabane, Anne Holbrook, " Application of Data mining techniques in pharmacovigilance. British Journal of Clinical Pharmacology," Blackwell Publishing Ltd, 57: 2 pp. 127-134, Feb. 2003.

[8] 이재호, 손유동,오범진, 김원, 임경수, "서울아산병원의 약물알레르기 경보시스템 초기적용 경험," 대한의료정보학회지 제 12권, 제2호, 133-140쪽, 2006년 6월.

[9] David W. Bates, R. Scott Evans, Harvey Murff, Peter D. Stetson, Lisa Pizziferri, and George Hripcsak, " Detecting Adverse Events Using Information Technology," J Am Med Inform Assoc Vol. 10, No. 2, pp. 115-128, Mar. 2003.

[10] James G. Anderson, Stephen J. Jay, Marilyn Anderson, and Thaddeus J. Hunt, "Evaluating the Capability of Information Technology to Prevent Adverse Drug Events, A Computer Simulation Approach," J Am Med Inform Assoc Vol. 9, No. 5, pp. 479-490, Sept. 2002.

[11] 박희경, 최진욱, 황재준, 하승민 "MLMPlus : Arden Syntax 기반의 실시간 의사결정지원시스템의 개발," 대한의료정보학회지, 제 12권, 보완본 2호, 167-172쪽, 2005년 4월.

저 자 소 개

이 영 호

2005: 아주대학교 의료정보학과 이학 박사

현재: 가천의과학대학교 의료공학부 교수

관심분야: 데이터마이닝, 의료정보, u- 헬스케어,

윤 영 미

2005:연세대학교 컴퓨터과학과 공학 박사

현재:가천의과학대학교 의료공학부 교수

관심분야:바이오인포마틱스, 데이터베이스, u-헬스케어

이 병 문

2005: 인천대학교 컴퓨터공학과 공학 박사

현재: 가천의과학대학교 의료공학부 교수

관심분야: 네트워크프로토콜, 센서네트워크, u-헬스케어

황 희 정

2005: 인천대학교 컴퓨터공학과 공학 박사

현재: 가천의과대학교 의료공학부 교수

관심분야: SW공학, 유비쿼터스컴퓨팅, u-헬스케어

강 운 구

1998: 인하대학교 컴퓨터공학과 공학 박사

현재: 가천의과학대학교 의료공학부 교수

관심분야: SW공학, RFID/USN, u-유헬스케어