KDD와 데이터마이닝을 이용한 바이러스성전염병 유행예측조사

윤종찬*, 윤성대**
*부경대학교 전산정보학과
**부경대학교 전산계산학과
*yjc313@hanmail.net

Virus communicable disease cpidemic forecasting search using KDD and DataMining

JongChan Yun*, SungDae Youn**
*Dept. of Computer and Information, Pukyung National University
**Dept. of Computer Science, Pukyung National University

요 약

본 논문은 대량의 데이터를 처리하는 전염병에 관한 역학조사에 대한 과정을 KDD(Knowledge Discovery in Database)와 데이터마이닝 기법을 이용해서 의료 전문인들의 지식을 데이터베이스화하여 데이터 선정, 정제, 보강, 예측과 빠른 데이터 검출을 하도록 하였다. 그리고 각 바이러스의 동향은 데이터마이닝을 활용하므로 일부분만의 데이터를 산출하지 않고 전체적인 동향을 산출, 예측하도록 한다.

1. 서론

데이터마이닝은 대량의 데이터베이스로부터 숨어있는 유용한 정보를 통계적 기법이나, 인공지능 등을 이용하여 찾아내는 일련의 과정으로 정의되며, 전자상거래, 의사결정 지원, 의료, 금융업 등의 다양한 분야에서 유용하게 활용되고 있다. 데이터마이닝의 기법으로는 인공신경망, 의사결정 나무, 사례기반 추론, 유전자알고리즘, 판별 분석, Association 등이 있다.[1, 2, 3, 4]

기존의 전염병역학조사에서는 아직까지 오프라인상에서 역학조사를 하고 있다. 전문가들의 지식을 가지고 현재 그 결과를 분석하고 매년 일어나는 바이러스 유행을 오프라인으로 서류를 통해서 그리고 전문가들의 경험을 통해서만 많은 데이터의 규칙성을 다루는 데 문제점이 있다.

데이터마이닝을 통해서 기존의 발견된 데이터를 데이터베이스화하고 KDD(Knowledge Discovery in Data- base)를 통해 빠른 검색과 다양한 형태의 지식을 발견하려고 한다.

본 논문에서는 국내(부산) 환자 중에 소화기바이러스에 대한 양성자 데이터를 가지고 연령별, 성별, 원인 바이러스별, 월별로 역학조사를 하고, 또한, 각 분석 차트를 통한 분석을 한다. 본 논문의 연구대상은 국내(부산시내)
병원환자를 대상으로 하여, 본 연구에 참여한 병원 환자 중에서 무작위 군집추출(random cluster sampling)하여 이들을 연구가능 대상자로 선별하였다.[5]

예를 들면, 기존 오프라인 데이터와 전문가들의 지식 기반 “2002년 인플루엔자 RAT(Rapid Antigen Test) & 분리 양성자 환자 리스트”를 지식탐사 절차(KDD)를 통해서 2002년도 인플루엔자 환자 중 기존 양성자 환자 테이블을 이용하여 성별, 연령별, 원인 바이러스별, 년도별 데이터 분석을 추출한다.

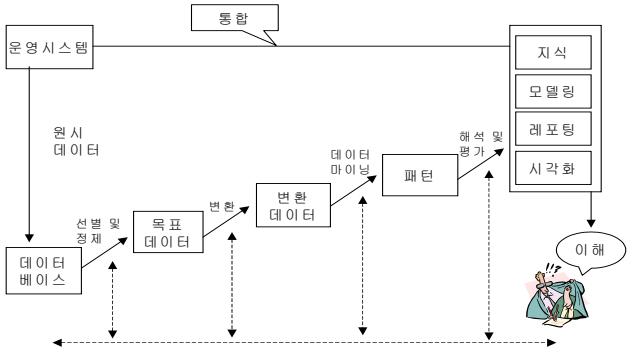
본 논문의 구성은 다음과 같다. 2장에서는 KDD절차와 데이터마이닝 분석에 대해서. 3장에서는 데이터마이닝의 개념에 대해서, 4장에서는 시스템 구현 방안 및 적용 평가를 제시, 마지막으로 4장에서는 결론 및 향후 제시를 한다.

2. 지식 탐사 절차(KDD)

KDD는 지식이 데이터를 기반으로 하는 발견(Data-driven Discovery)의 최종 생산물로, 특히 데이터마이닝의 고급 수준의 응용과 데이터에서 지식을 발견하는 넓은 의미의 프로세스를 지칭한다. 즉 KDD는 데이터로부터 유용한 지식을 발견하는 전체적인 프로세스를 의미한다.[6]

2.1 지식 탐사 절차(KDD)의 절차 과정

그림1은 데이터베이스 내에 있는 원시 데이터를 KDD의 절차과정을 통해 산출하는 과정이다. 이 과정에서 산출된 데이터들의 규칙 패턴을 데이터마이닝을 통해 결과를 내고 해석 및 평가 후 시각화나 레포팅, 모델링한 결과 데이터를 산출한다.



자료원 : Fayyad. 1998

(그림1) KDD의 절차과정

(1) 데이터 선정(data selection)

데이터 선정은 표1과 같이 기존의 데이터베이스에서 사용하려는 원본데이터를 선정한다.

<표 1> 2002년도 소화기계 바이러스 양성자 결과

월	가검물수	검출건수	Calici	Rota	Echo group	Coxsackie group	Adeno	Polio Pool	Geno-typing 중	Entero	Non-entero	Untypable
1	11	1		1								
2	6											
3	13											
4	19	5			2			2				1
5	101	10			8			1				
6	205	23			21	1						2
7	144	32			26							1
8	82	9			7	5						2
9	39	6			4							2
10	21	5			4	1						
11	48	2				2						
12	14											
합계	703	93	0	1	72	9	0	3	0	0	0	8

(2) 정제(cleaning)

정제는 표2와 같이 원본 데이터에서 보고서에서 사용할 잡음을 제거하고, 데이터에 있는 불일치를 바로 잡는다.

<표 2> 2002년도 소화기계 바이러스 양성자 결과

월	Calici	Rota	Echo group	Coxsackie group	Adeno	Polio Pool	Geno-typing 중	Entero	Non-entero	Untypable
1		1								
2										
3										
4			2			2				1
5			8			1				
6			21	1						2
7			26							1
8			7	5						2
9			4							2
10			4	1						
11				2						
12										
합계	0	1	72	9	0	3	0	0	0	8

(3) 보강(enrichment)

보강은 원본 데이터 정보에 추가 자료를 보강하고, 기존의 정보와 조인을 통해 자료를 보강한다.

(4) 코딩(coding)

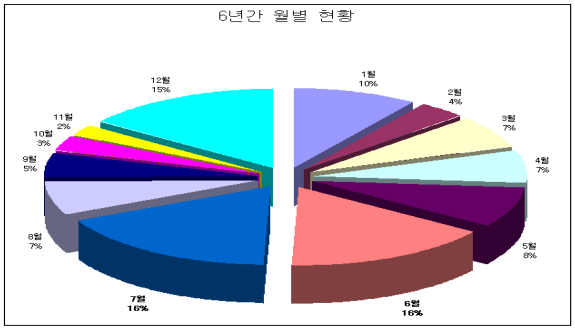
코딩은 데이터에 대한 변환과 누락된 정보를 가진 레코드는 삭제한다. 또한 병원별, 년도별, 발병일자별로 발생한 데이터에 대한 현황 및 예측을 파악할 수 있다.

(5) 보고서 작성(reporting)

보고서작성은 표3과 같이 6년 간 발생한 소화기계 바이러스 양성자현황 결과이다.

<표 3> 1998~2002년 소화기계 바이러스 양성자 결과

월	가검물수	검출건수	Calici	Rota	Echo group	Coxsackie group	Adeno	Polio Pool	Geno-typing 중	Entero	Non-entero	Untypable
1	68	26	16	6		1		2				1
2	87	10	5	5								
3	101	18	4	10				1		1	1	1
4	143	18		1	6	6		2			2	1
5	265	23			11	6		1		1	4	
6	463	43			29	5		1			6	2
7	428	47		1	28	13		2			1	2
8	185	19			10	2	1				4	2
9	255	14			6	3		1			1	3
10	108	8		1	4	2					1	
11	96	6			1	2					2	1
12	211	41	20	14	1	4	1				1	
합계	2410	273	45	38	96	44	2	10	0	2	23	13



(그림 2) 데이터 처리된 결과 차트

그리고 그림2는 6년간(1998-2003년) 발생한 바이러스 중 어느 월에 가장 바이러스 발생률이 큰가를 확인할 수가 있다. 6년간 데이터를 처리한 결과 6~7월에 바이러스 발생률이 큰 것을 알 수 있다.

3. 데이터마이닝

데이터마이닝이라는 것은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정이라고 할 수 있다. 예를 들어, 정보화 사회에 따라 병원에서는 처방전달시스템, 의무기록전산화로 방대한 양의 데이터가 축적되어 있는데, 이를 데이터 속에서

의학연구나 병원경영에 유용한 정보를 찾아내는 것을 보건분야에 있어서의 데이터마이닝이라 할 수 있다.[7]

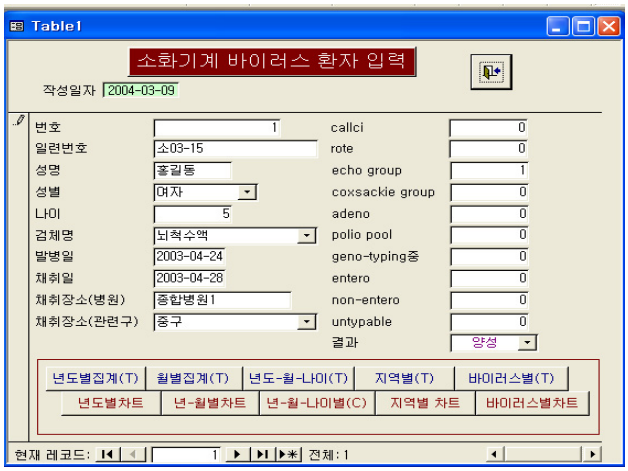
본 논문에서는 데이터마이닝 기법 중 질의 도구, 온라인 분석처리, 사례-기반 학습, 연관 규칙 등을 참고로 분석된 자료를 산출하였다.

4. 시스템 구현 방안 및 적용 평가

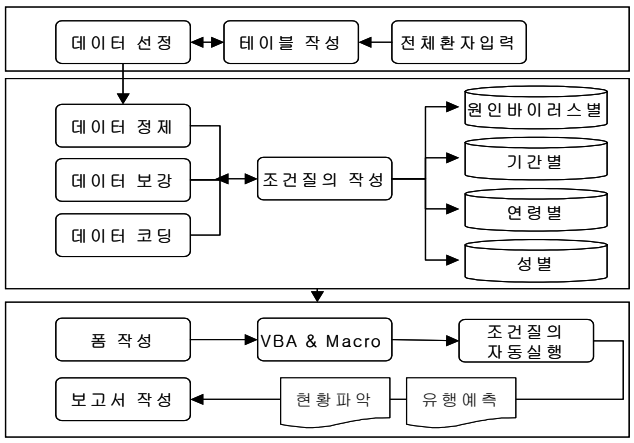
4.1 예측조사 시스템 모델

본 논문에서 설계된 그림3의 예측조사 데이터 처리 시스템은 환자 문진표 입력 데이터 분석표로 구성된다. 환자 문진표 입력은 기본 환자 입력 사항 등록, 검색을 한다.

또한 기존에 없던 판단 예측은 데이터 처리를 통해 분석된다. 분석표에서는 예측조사를 한 결과를 성별, 나이별, 바이러스별로 구별해서 차트로 구분해서 나타낸다.



(그림 3) 환자관리 시스템



(그림 4) KDD이용한 시스템 설계 구조

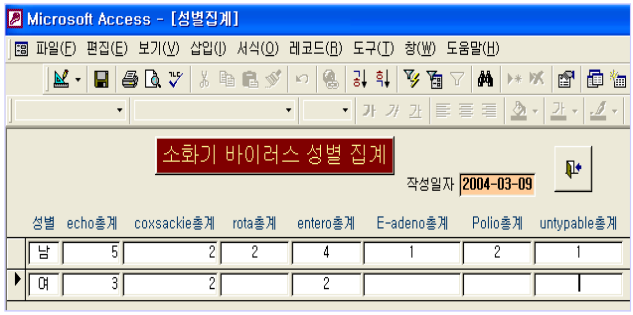
그림4는 환자관리 시스템의 데이터의 입력에서 질의를 해석하여 현황파악 및 유행예측 결과를 나타내는 시스템설계 구조이다.

4.2 시스템 사양

본 연구의 시스템 사양은 운영체제는 Windows98, CPU는 팬티움III이상이고, HDD는 10GB이상, RAM은 128M이며, 사용된 프로그램은 MS-OFFICE2000의 패키지 중 ACCESS2000으로 작성했다.

4.3 시스템의구현 결과

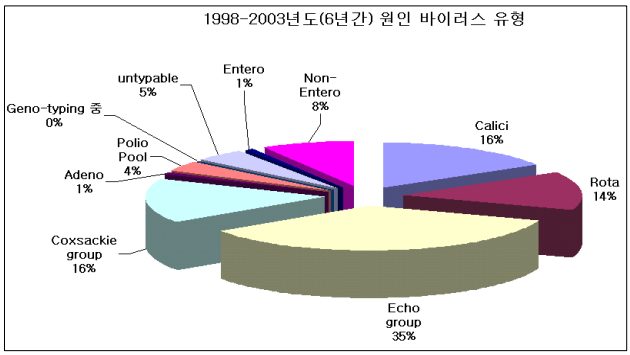
의학 전문가들 중 전산 지식이 없는 것을 대비해서 현재 대중화되어 있는 MS-OFFICE 패키지 중 액세스를 가지고 프로그램을 작성했다. 조작법도 VBA(Visual Basic For Application)와 매크로(macro)를 이용해서 명령단추를 만들어서 데이터를 입력한 후 관련 명령단추만 눌러도 그림5(2003년도 바이러스별 성별 집계)나 표4(1998~2003년도 바이러스 현황테이블)와 같이 나타나거나 차트가 나오도록 구성되어 있다.



(그림 5) 2003년도 소화기 바이러스 성별 집계

<표 4> 1998~2003년도 바이러스 현황테이블

년도	Calicl	Rota	Echo group	Coxsackie group	Adeno	Polio Pool	Geno-typing중	Entero	Non-entero	Untypable
1998	8		7	13		1			1	1
1999	28	22	15	17	2	1			10	4
2000	9	14	1	1		5				
2001				4					12	
2002		1	72	9		3				8
2003		1	1					2		



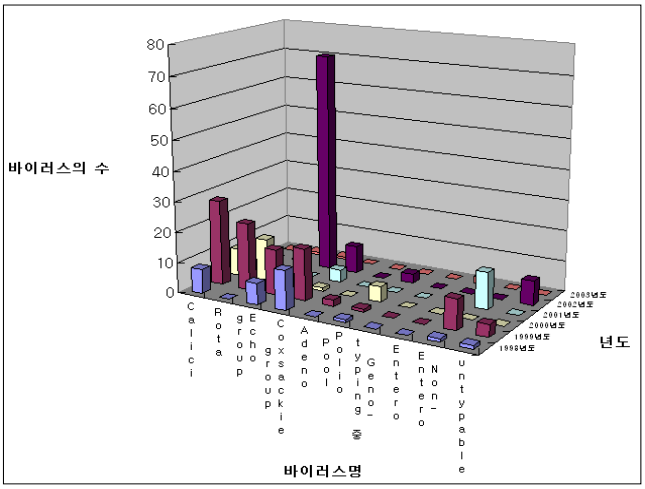
(그림 6) 데이터 처리를 통한 결과 차트

그림6의 데이터 처리된 결과 차트를 보면 1998년부터2003년까지의 소화기 계통의 바이러스별 발생 빈도

를 분석할 수 있다.

(그림 7) 병원별 소화기바이러스 환자 현황

그림7은 바이러스 검사 대상물을 채취한 병원별 현황으로 병원별 특성에 따라 소화기 계통의 바이러스별 발생 빈도를 표현하여, 소아들이 있는 산부인과병원에서 바이러스의 다양한 유형이 나타남을 볼 수 있었다.



(그림 8) 년도별 바이러스 현황 차트

위의 그림8에서 볼 수 있듯이 기존 테이블의 데이터 처리된 결과를 보면 향후 발생할 바이러스의 년도 예측이 가능하다.

<표 5> 2003년도 나이별 바이러스 현황

	데이터						
나이	개수:echo	개수:coxsackie	개수:rota	개수:entero	개수:adeno	개수:Polio	개수:untypable
1-5	6	4	1	5		1	
6-10	2		1		1	1	1
36-40				1			
총 합계	8	4	2	6	1	2	1

표5의 결과는 사용한 패키지에 VBA(Visual Basic for Application)와 매크로를 이용해 작성된 명령단추를 누르면 각 년도별 소화기별 바이러스 나이별현황과 예측이 가능한 표 화면이다.

이 표에서는 2003년도 나이별 현황을 볼 수 있는데, 나이가 어릴수록 소화기 바이러스 유형별이 고르게 분포되는 것을 볼 수 있었다.

5. 결론

본 논문은 역학조사에 있어서의 오프라인의 데이터 구현 방식을 데이터 처리 기법과 데이터마이닝을 이용해서 환자들의 예측조사를 최대한 반영하고 결과의 만족도를 극대화하여 그 데이터를 사례베이스로 구축하여 차후에 발생하는 결과치를 빠르게 판단하는 시스템을 구축하였다. 그러므로 기존의 오프라인 역학조사에서 빠르게 볼 수 없는 성별, 나이별, 원인 바이러스별, 년도별로 역학조사를 예측하는 시스템도 설계, 구성되어 년도별 각 바이러스성 전염병 유행예측 조사를 볼 수 있다.

그리고 제안된 시스템은 관측자의 판단 선례를 비교 분석할 수 있도록 하였으며, 앞으로의 바이러스성 전염병의 유행 예측성을 관찰할 수 있으며, 관측자들의 빠른 판단과 예측성을 가질 수 있도록 하였다.

향후 과제는 구현 시스템을 네트워크를 통해서 전문가들과 실시간 데이터를 연결하여 빠른 역학조사의 예측을 알아 볼 수 가 있고, 담당의사와도 실시간으로 대화하여 서로의 의견이나 상담이 이루어질 수 있도록 시스템을 구축하는 것이다. 그리고 환자들과의 실시간 상담도 이루어질 수 있도록 다양한 자연어 처리의 데이터베이스가 필요하다.

<참고문헌>

[1] Pieter Adriaans, Dolf Zantinge(著), 용환승(譯), "DataMining & KDD", 그린(出), 1998.
[2] 박우창, 승현우, 용환승, 최기헌(譯), "데이터 마이닝 개념 및 기법", 자유아카데미(出), 2003.
[3] 김진현, 윤성대 "다중데이터베이스 마이닝의 전처리를 위한 가중치 거리 기반 클러스터링", 2004년 2월 정보처리학회논문지, 2004.
[4] 이준옥, 이용준, 류근호, "시간 데이터마이닝 프레임워크", 2002년 6월 정보처리학회논문지, 2002.
[5] 조경순, 부산보건환경 연구원 역학조사과 "바이러스성전염병 유행예측조사 결과 모음집", 1998-2003.
[6] http://iems.net/k/dir/mt/related_tern/dm_k.html
[7] 최국렬 외 8명 공저, "데이터마이닝 이론과 실습, 보건의료데이터 중심", 청구문화사(出), 2001.