순차패턴 마이닝을 이용한 상병의 연관성 분석

진종식[a] 박희준[b] 이정현[b] 김윤년[b] 윤경일[b] 엄홍섭[b]

[a] *Biomedical Informatics Technology Center, Keimyung University, Daegu, 700-712*
Tel: 053-428-7951, Fax: 053-428-7953, E-mail: make1s@naver.com

[b] *Department of Medical Informatics, School of Medicine, Keimyung National University, Daegu, 704-701*
Tel: 053-428-7951, Fax: 053-428-7953

Abstract

데이터 마이닝 기법 중 순차 패턴 마이닝(Sequential Pattern Mining)은 연관 규칙에 시간의 개념을 추가하여 시간의 흐름에 따른 항목(item)들의 상호 연관성을 찾아내는 것이다. 본 연구의 목적은 순차적인 상병의 발생 가능성이 높은 상병 군의 패턴을 찾아내어 이를 모형화함으로써 차후에 발생될 상병을 예방하고 이를 통하여 환자와의 관계를 관리하여 보다 나은 의료서비스를 제공하는데 있다.

Keywords:

데이터 마이닝, 순차패턴, 연관성 분석, 병원 경영정보시스템, 환자관계관리

서론

지난 20세기 동안 의료패러다임은 상병의 예측 불가능성을 전제로 하여 많은 기구를 할당하고 급성 질환을 관리하기 위한 역할을 위주로 진행이 되어 왔다. 하지만 의료분야에서도 정보화가 진행이 되고 전자의무기록(EHR: Electronic Health Record) 과 개인건강기록(PHR: Personal Health record) 등의 개념이 등장함에 따라 의료정보가 분석 가능한 형태의 데이터로 저장이 되어 왔다. 그리고 의료패러다임이 진단 및 치료중심에서 예측 및 관리 중심으로 변화함에 따라 의료기관은 이를 이용하여 질병의 예방과 건강 증진뿐만 아니라 환자의 의료요구를 파악하고 의학적으로 필요한 시기에 적절한 의료서비스를 제공하여야 할 것 이다.

이에 본 연구는 순차패턴 마이닝 기법을 이용하여 상병의 순차적인 패턴을 알아 봄으로써 의료의사결정과 의료의사결정 지원시스템을 개발하는데 도움을 주고자 하였다.

연구의 배경

1. 병원경영정보학

정보화 사회가 진행됨에 따라 일반 관리 분야에서도 정보기술의 사용이 증가 되었다. 병원 분야의 조직관리 또한 타 분야의 조직 관리에서처럼 정보기술을 널리 사용하게 되었는데 특히 보건의료 분야는 방대한 데이터를 신속, 정확하게 처리해야 함으로 경영정보시스템의 필요성이 강조 되었다. 하지만 경영정보시스템이 보건의료분야에 접목했을 때 기대한 만큼의 좋은 성과를 가져오기 힘들었는데 이는 병원이라는 느슨한 조직을 일반적인 경영정보 시스템의 적용으로는 효율적으로 활용하기가 힘들었기 때문이다. 이러한 이유로 병원의 특성에 부합되는 새로운 학문 분야인 병원경영정보학이 등장 하게 된 것이다[1].

2. 의료 환경의 변화와 환자관계관리

현재까지 의료행위는 생명을 위협하는 급성질환을 관리 하기 위한 역할을 위주로 진행이 되어 왔다. 즉 지금까지의 의료행위는 가능한 폭 넓은 위험 요소를 위한 비용을 분산 시키기 위해 건강 보험이라는 의료 재원을 마련해 왔으나 이것은 상병의 예측 불가능성을 전제로 하여 많은 비용을 할당하고 관련 기구를 운영 하게하였다. 하지만 향후의 미래에는 상병의 예측 가능성이 비중을 둔 의료체계로의 변화가 일어날것이다. 즉, 상병에 대한 증상의 발현 이전에 만성상병의 위험을 예측하고 그 질환에 대한 위험요인에 대하여 관리함으로써 처방 중심에서 예방 중심으로 의료 패러다임이 변화 할 것이다[1].

의료 패러다임이 "진단 및 치료 중심"에서 개인의 유전정보와 건강 행태를 기반으로 개인의 질환을 예측 한 후 고 위험도 집단에서 이를 예방할 수 있도록 관리 해주는 "예측 및 관리 모형"으로 변화 함에 따라 병원의 조직 및 운영 체계, 마케팅 전략 등에서 많은 변화가 있어야 할 것 이다.

이를 위해 일반기업에서 이미 많이 사용되고 있는 고객 관계관리 기법을 활용한 병원 마케팅 전략의 개발이 필요하다.

의료기관의 마케팅개념은 기업의 마케팅 개념과 많은 차이가 있으나 조직의 효율적인 이용을 통하여 마케팅적인 사고를 이용한다는 점은 같다.

의료기관에서 마케팅의 고객관계관리(CRM: Customer Relationship Management) 개념을 적용 시켜보면 "환자와의 관계를 통해 상병의 예방과 건강 증진뿐만 아니라 환자의 의료요구를 파악하고

의학적으로 필요한 시기에 적절한 서비스를 제공함으로써 환자의 충성 도를 유지, 증대 시키고 지역 사회 및 국민의 건강증진과 향상이라는 사회가치를 구현한다" 라고 말할 수 있다.

하지만 기업의 고객관계관리(CRM) 기법을 보건의료분야에 적용하는데 중요한 문제는 일반 기업에서는 고객의 수익성을 중심으로 고객을 분류하고 관리하는데 반하여 보건의료 분야에서는 환자의 건강정보를 필요한 관점에 따라 분류한다는 점이다.

즉 건강 정보가 필요한 계층에게 적절한 건강정보를 제공하여 바람직한 건강 행위를 유도하여 건강증진 및 치료가 이루어지도록 환자를 관리하는 것이라고 할 수 있는데 의료분야에서의 고객관계관리는 이러한 특성을 반영한 환자관계관리(PRM: Patient Relationship Management)라는 개념을 사용 할 수 있다[1].

3. 보건의료 분야의 데이터 마이닝

데이터 마이닝은 방대한 데이터 속에 숨겨진 패턴이나 관계를 찾아내기 위해 데이터를 선택하고 탐 색 및 모델링하는 과정이다.

데이터 마이닝은 마케팅과 은행, 고객관계관리, 공학 등 다양한 분야에서 성공적으로 적용이 되어 왔지만 높은 열망에도 불구하고 의료분야에서는 많은 제한이 있었다.

의료분야의 데이터 마이닝의 목표는 견고한 예측 모델의 생성과 신뢰성 높은 예측을 통하여 의료실무자들이 병의 예후와 진단에 도움을 주는 것이다[3]. 때문에 데이터의 정확성과 높은 수준의 신뢰성을 요구한다.

4. 순차패턴 마이닝

데이터 마이닝 기법 중 순차패턴(Sequential Pattern) 마이닝은 연관 규칙에 시간의 개념을 첨가하여 시간의 흐름에 따른 항목들의 상호 연관성을 찾는 것 이다. 즉 순차패턴 마이닝 기법은 사용자가 지정한 최소지지도를 만족하는 빈도가 높은 시퀀스(1)들을 추출하고 이들 가운데 최대 시퀀스를 찾는 것이다.

식(1) 트랜잭션 데이터베이스에서의 최소지지도

$$Support = \frac{\#\ of\ transaction\ all\ the\ item\ in\ X \cup Y}{total\#\ of\ transaction\ in\ the\ database}$$

본 연구에서 순차연관규칙 X⇒Y 는 "상병X 가 발생하면 일정한 시간이 경과한 다음 상병 Y 가 발생한다." 라고 해석한다. 연관규칙과 순차패턴의 차이점은 연관규칙은 X⇒Y, Y⇒X 가 성립하지만 순차패턴에서는 X⇒Y 만 성립한다는 것이다.

순차패턴 마이닝 알고리즘에는 AprioriAll, AprioriSome, SPADE(Sequential Pattren Discovery using Equivalence classes), PSP(Prefix tree for Sequential Pattern)등이 있다. 이러한 순차패턴 알고리즘을 이용하여 본 연구에서는 환자의 상병에 대한 순차적인 발생가능성 패턴을 찾고 가장 빈번하게 발생하는 규칙에 대한 분석을 시도하였다.

연구의 설계

1. 자료

연구의 자료는 대구에 위치한 D병원의 1998년 1월에서 2006년 12월까지의 8년간의 환자의 진단정보를 포함한 트랙잭션 데이터를 추출하고 환자의 식별 가능한 ID를 기준으로 진단데이터를 정렬하였다.

추출된 자료는 10개의 변수를 포함하고 있고 전체 환자의 수는 1,656,850 명이다.

2. 방법

본 연구의 방법은 크게 네 단계로 나누어 분석을 하였다.

첫째, 트랜잭션 데이터베이스에서 환자의 진단 정보를 포함한 상병 데이터를 추출 하였다. 추출된 데이터는 환자ID, 입원날짜, 퇴원날짜, 진단과, 상병코드, 상병명, 상병의 순서, 상병의 중요도, 주 상병, 입원 및 외래 의 필드를 포함하고 있다.

둘째, 추출된 데이터를 보다 빠른 순차패턴 마이닝 알고리즘을 적용하기 위해 환자의 ID 별로 데이터를 정렬 하고 알고리즘을 적용 시켰다.

셋째, 최소지지도와 신뢰도를 만족하는 연관성 규칙을 찾아내었다.

넷째, 발견된 순차패턴 연관성 규칙을 웹링크 분석을 통하여 분석하고 평가 해보았다.

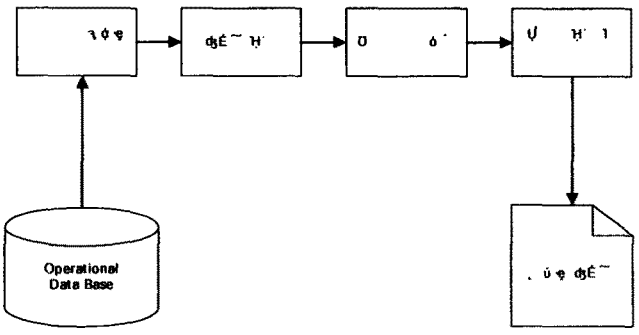


그림 1 - 순차패턴 마이닝 다이어그램

분석 결과

A 병원의 8년간의 진단 데이터를 상병의 비율과 빈도로 분석한 결과 가장 빈번하게 발생한 상위20개의 상병은 당뇨병, 고혈압, 뇌색경증, 협심증, 만성신장기능상실, 급성기관지염 등의 순으로 나타났다.

표 1에서 나타난 결과를 보면 전체상병 8,611개 가운데 가장 높은 빈도를 차지한 상병은 인슐린비의존 당뇨병(E11) 로 나타났고 이 상병은 전체 상병에서 0.036% 의 비율을 나타내고 있으며 332,818 번의 빈도를 나타내었다.
이것은 A병원의 전체 환자 중 인슐린비의존 당뇨병의 진단을 받은 환자가 가장 많은 것을 보여준다. 또한 전체 결과를 보면 당뇨병과 당뇨의 합병증에 의한 상병이 가장 많은 것으로 나타났다. 이러한 결과를 보면 A병원에서 입원 및 외래 진료를 받는 환자 가운데 가장 많은 환자가 당뇨병과 관련된 환자라고 할 수 있다. 그리고 두 번째로 많은 환자는 고혈압 환자라고 할 수 있는데 고혈압 환자 또한 당뇨병 환자를 제외한 타 상병을 가진 환자보다 배 이상으로 높은 빈도를 보여 주고 있다. 다음으로 많은 비율을 차지하는 것이 뇌색경증과 협심증, 만성신장기능상실의 순이었다.
분석의 결과 특이한 것은 대부분의 상병이 만성질환에 의한 것이었으나 위의 악성 신생물 즉, 위암이 높은 순위를 차지하고 있다는 것이다.

상병 이름	비율	빈도
인슐린 비의존 당뇨병 (E11)	0.036	332818
본 태성 고혈압 (I10)	0.025	228116
뇌색경증 (I63)	0.011	102094
협심증 (I20)	0.009	86765
만성신장 기능 상실 (N18)	0.008	76156
급성 기관지염 (J20.9)	0.008	71817
위의 악성 신생물 (C16)	0.008	71258
만성 위염 (K29.3)	0.007	68517
소화불량 (K30)	0.007	66491
합병증 동반 당뇨병 (E11.4+)	0.007	65408
위궤양 (K25)	0.006	58752
전립샘의 염증성질환 (N40)	0.006	57462
만성 신장기능상실 (N18.0)	0.006	57454
신장 합병증 당뇨병 (E11.2+)	0.006	53454
눈 합병증 당뇨병 (E11.3+)	0.006	53143
위염 및 십이지장염 (K29)	0.006	51878
말초 순환장애 합병 당뇨병 (E11.5)	0.006	51106
당뇨 망막병증 (H36.0*)	0.006	50998
상세불명의 폐렴 (J18.9)	0.005	48230

표 1. A병원의 상위 20개의 상병의 비율과 빈도

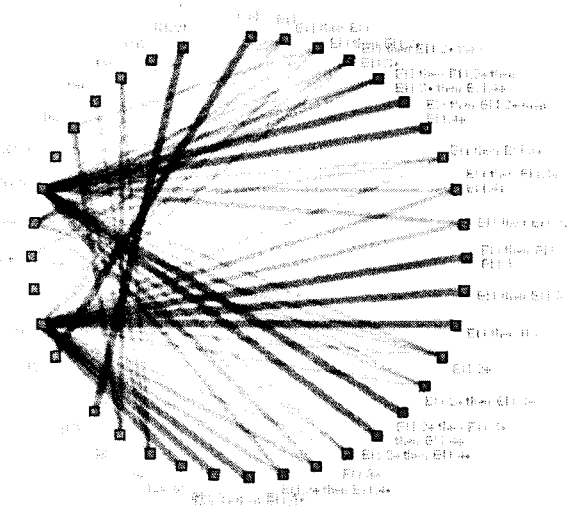
본 연구에서는 SPSS사의 Clementine 10.1을 이용하여 순차적인 연관성 규칙 모델을 생성하였다.
전체 데이터에서 연관규칙의 최소 지지도는 1.9 최소 신뢰도 10.4 를 적용한 결과 57개의 유효한 규칙이 생성 되었다.
생성된 57개의 규칙을 다시 후항 값으로 정렬한 결과 후항 값을 기준으로 하여 전체 12개의 상병이 나타났다.
본 연구의 순차규칙 후항 값은 환자들의 최종 상병을 의미한다. 최종상병을 기준으로 후항 값을 정렬한 결과 순차 규칙 적용전의 결과와 비슷하게 당뇨병(E11) 환자가 가장 높은 빈도를 나타내었다. 하지만 순차규칙 적용이전과의 차이를 보이는 결과는 말초순환장애 합병당뇨병(E11.5) 과 합병증동반 당뇨병(E11.4+)이 순차규칙 적용 이전보다 높게 나타 난 것인데 이것은 초기에 당뇨병 진단을 받은 환자가 결과적으로 당뇨합병증을 가지게 된다는 결과를 나타내는 것이다.

순서	값	%	빈도
1	E11	31.58	18
2	E11.5	26.32	15
3	E11.4+	14.04	8
4	I10	7.02	4
5	I63	5.26	3
6	E11.3+	5.26	3
7	I20	1.75	1
8	N18	1.75	1
9	H36.0*	1.75	1
10	E11.2+	1.75	1
11	N18.0	1.75	1
12	C16	1.75	1

표 2. 후항 값의 분포

후항값의 결과로 나타난 12 개의 상병을 기준으로 웹 분석을 시도하였다. 웹 분석은 전항값을 가지는 연관규칙과 후항값을 가지는 연관 규칙 사이의 연결관계를 보다 쉽게 알아보고자 한 것이다.
전체 연관 규칙 57개를 상병들 사이의 연관성 비율로 웹분석을 시도한 결과 강한 링크를 보이는 12개의 연관 규칙을 찾을 수 있었다.
규칙의 특징을 보면 첫 번째 규칙 전항값 "E11.2+ then E11.3+ then E11.4+" 와 후항값 "E11.5"를 가지는 규칙은 신장합병당뇨병에 걸린 환자는 눈합병당뇨병에 걸리고 합병증동반당뇨병의 결과를 가지며 최종적으로 말초순환장애합병 당뇨병에 걸린다. 는 결과를 얻을 수 있었다. 그리고 다

셋번째 규칙에서는 당뇨망막병증에 걸린 환자가 최종적으로 인슐린비의존당뇨병에 걸린다는 규칙을 얻을 수 있었다. 마지막 12번째 규칙에서는 당뇨병을 제외한 규칙이 나타났는데 이 규칙은 뇌색경증이 있는 환자가 본태성고혈압에 걸린다는 것을 의미한다.
하지만 이러한 결과가 반드시 뇌색경증이 있는 환자가 본태성고혈압에 걸린다는 것을 의미하지는 않는다. 분석에서 사용된 규칙의 비율은 순차패턴 마이닝 분석을 통하여 일정한 지지도와 신뢰도를 기준으로 찾아낸 것이기 때문이다.
전체 빈도가 높은 상병 가운데 위와 관련된 상병이 위의 악성 신생물, 만성위염, 소화불량, 위궤양, 등으로 높게 나타났다. 하지만 순차 규칙에서는 최소지지도와 신뢰도를 만족하는 결과가 나타나지 않았다. 이러한 결과는 A병원이 종합병원 즉, 3차 병원이기 때문에 위암과 같은 중증 질환은 1차와 2차 병원에서 대부분 초기진단을 받고 다시 3차 병원으로 와서 재 진단 후에 치료를 받기 때문이다.



랭크	필드 1	필드 2
100%	전항값 = "E11.2+ then E11.3+ then E11.4+"	후항값 = "E11.5"
100%	전항값 = "E11 then E11.2+ then E11.3+ then E11.4+"	후항값 = "E11.5"
100%	전항값 = "E11 then E11.2+ then E11.4+"	후항값 = "E11.5"
100%	전항값 = "E11.2+ then E11.4+"	후항값 = "E11.5"
100%	전항값 = "H36.0"	후항값 = "E11"
100%	전항값 = "E11.5"	후항값 = "E11"
100%	전항값 = "E11 then E11.5"	후항값 = "E11"
100%	전항값 = "C16"	후항값 = "C16"
100%	전항값 = "E11.4+ then E11.5"	후항값 = "E11"
100%	전항값 = "E11 then E11.4+ then E11.5"	후항값 = "E11"
100%	전항값 = "E11 then I10"	후항값 = "E11"
100%	전항값 = "N18"	후항값 = "N18.0"
50%	전항값 = "E11 then E11.2+ then E11.3+"	후항값 = "E11.4+"
50%	전항값 = "E11 then E11.3+ then E11.4+"	후항값 = "E11.5"
50%	전항값 = "E11.3+ then E11.4+"	후항값 = "E11.5"
50%	전항값 = "E11 then E11.4+"	후항값 = "E11.5"
50%	전항값 = "E11.4+"	후항값 = "E11.5"
50%	전항값 = "E11 then E11.2+ then E11.3+"	후항값 = "E11.5"
50%	전항값 = "E11.4+"	후항값 = "E11"
50%	전항값 = "E11 then E11.4+"	후항값 = "E11"
50%	전항값 = "E11.3+ then E11.4+"	후항값 = "E11"
50%	전항값 = "E11 then E11.3+ then E11.4+"	후항값 = "E11"
50%	전항값 = "I63"	후항값 = "I63"
50%	전항값 = "I63"	후항값 = "I10"

그림 2-순차규칙의 웹링크분석

결론

본 연구에서는 순차패턴 마이닝을 이용하여 상병의 연관성을 분석하였다. 연관성 분석의 결과 A병원의 주 상병은 당뇨병과 그에 따른 합병증이 가장 많은 것으로 나타났으며 고혈압, 뇌색경증, 협심증, 만성신장기능상실, 급성기관지염의 순으로 나타났다. 또한 대부분의 상병이 만성질환인데 반하여 위암에 대한 비율이 높게 나타난 것이 특이한 결과로 나타났다.
분석의 결과에서 가장 지지도와 신뢰도가 높았던 순차규칙인 당뇨병과 그에 따른 합병증을 나타내는 규칙을 보면 현재 A병원의 환자들 가운데 가장 많은 환자들이 당뇨병에 대한 질환으로 재입원하는 경우가 많다는 것을 의미한다.
이러한 결과를 볼 때 A병원의 환자관계관리는 다음과 같은 방법을 제안 할 수 있다.

첫째, 당뇨병에 걸린 환자들의 측면에서 전문 당뇨클리닉의 운영 등으로 당뇨병 치료 및 당뇨병관리, 당뇨합병증 예방 등에 대한 교육을 실시하여 합병증의 재발을 줄여 나가야 할 것이다.

둘째, 전체 상병에서 빈도가 높고 재입원 규칙도 높은 위암환자들에 대한 관리가 필요 할 것이다. 즉 위암에 대한 처치와 치료를 특성화 시켜 위암의 치료효과를 증대 시키고 병의 조기발견과 예방에 노력해야 할 것이다.

연구의 한계와 향후 연구 계획

본 연구의 한계와 향후연구계획은 다음과 같다.

첫째, 대용량의 데이터에서 순차적인 연관성 규칙을 찾아보고자 하였기에 지지도와 신뢰도가 높은 규칙들은 대부분 평 이한 결과를 보여 주었다. 즉 가장 높은 지지도와 신뢰도를 가지는 규칙들은 주 질병과 그에 따른 합병증으로 나타난 것이다. 하지만 가장 높은 신뢰도와 지지도를 가지는 규칙들을 제외한 나머지 규칙들에서 몇 가지 유용한 규칙들이 나타났는데 이를 통하여 주 질병 다음으로 발생 가능성이 높은 몇 가지 질병을 확인 할 수 있었다. 향후의 연구에서는 몇 가지 대표적인 질병군을 분류하여 질병 군 별로 순차적인 패턴을 알아봄으로써 보다 상세한 순차패턴 규칙이 생성되어야 할 것이다.
둘째, 본 연구는 질병과 관련된 데이터를 통하여 순차적인 발생 패턴을 알아보기 위해 순차패턴 마이닝 알고리즘을 적용하였다. 즉, 질병에 대한 데이터만 분석 하였기에 환자에 대한 정보와 질병의 상세한 분석이 부족하였다. 때문에 향후 연구에서는 환자의 개인정보와 질병의 특성 정보를 포함하여 여러 차원에 걸친 다차원 순차패턴 마이닝 분석이

유용할 것으로 예상 된다.

참고문헌

[1] 강성홍 외(2002). *병원경영정보관리*, 고려의학, pp. 337-340.

[2] Riccardo, B., and Blaz, Z. (2006). "Predictive data mining in clinical medicine: Current issues and guideline," *International Journal of Medical Informatics*, pp. 76-87.

[3] Agrawal, R., Srikant, R.(1995). "Mining Sequential Patterns", *In Proceedings of the 11th International Conference on data Engineering*, pp3-14