전자의무기록으로부터 진료경로 추출을 위한 연관규칙마이닝 접근 방법

배인호*, 김진상*, 최상열*, 김윤년**
*계명대학교 정보통신대학
**계명대학교 의과대학 내과학 및 의료정보학교실
e-mail:inobae@gmail.com, jsk@kmu.ac.kr,
el2idea@hanmail.net, ynkim@dsmc.or.kr

An Association Rule Mining Approach to Extract Clinical Pathways from EMR

Inho Bae*, Jin Sang Kim*, Yoon Nyun Kim**
*College of Information and Communication, Keimyung University
**Medical School, Internal Medicine, Keimyung University

요 약

본 논문에서는 임상의들의 진료데이터를 토대로 진료경로를 동적으로 생성하는 방법을 기술한다. 각 진료단계에서 추출된 규칙들을 토대로 진료경로를 생성하는데, 이를 위해 전자의무기록으로 구성된 임상 데이터를 기반으로 연관규칙마이닝을 이용하여 진료단계별 규칙을 추출하였다. 신뢰성 있는 진료경로의 추출이 이루어지면 의료 서비스의 질을 높이고, 병원 경영의 효율성 증대에 도움을 줄 수 있다.

키워드 : 연관규칙마이닝, 전자의무기록, 진료경로

1. 서론

초창기 의료 분야의 전산화 방향은 원무 시스템을 중심으로 의사의 진료 영역은 배제된 채 행정업무에 그쳤으나, 현재는 진료 영역을 포괄하는 전반적인 전산화 시스템을 구축하고 있다. 처방전달시스템(OCS, Order Communication System) 중심의 진료를 지원하는 유틸리티 성격의 초기 소프트웨어에서 현재는 병원 내에 전자의무기록(EMR, Electronic Medical Record)을 통해 차트가 없어지는 디지털 병원의 형태를 추구하고 있다. 그러나, 이러한 병원정보시스템(HIS, Hospital Information System)을 통해 쌓여지는 방대한 양의 데이터들에 대한 활용은, 진료 기록에 대한 열람, 통계 기록 등과 같은 단순한 저장과 검색의 수단으로만 사용되는데 그치고 있다. 의료 정보 시스템을 통해 축적된 임상 데이터나, 연구관련 데이터에 대한 가공을 통한 재활용은 의료의 질과 서비스 만족도를 높이고 의료기관의 경영 효율성을 높이며, 합리적 의사결정에 대한 지원을 가능하게 할 수 있다.

진료경로(CP, Clinical Pathway)는 1980년대 미국의 New England Medical Center에서 DRG (Diagnosis Related Group), PPS(Prospective Payment System)을 보완하기 위하여 도입된 후 의료의 각 분야에서 급속히 확산되고 있다. DRG란, 미국의 예일 대학팀에 의해 1960년대 말부터 10여년에 걸쳐 병원경영 개선을 위해 개발된 입원환자 분류체계로 1980년 미국 노인대상의료보험(Medicare)의 병원 진료비 지불방식이 포괄수가 제도로 바뀌면서 이 제도의 지불단위로 사용되기 시작했다. DRG 분류체계에서는 모든 입원환자들이 주 진단명 및 부상병명, 수술명, 연령, 성별, 진료 결과 등에 따른 질병군으로 분류되는데 이때 각각의 질병군을 DRG라고 부른다. 이는 곧 환자가 병원에서 어떤 치료를 받든지 입원 일수와 질병의 정도에 따라 미리 정해진 보험급여를 병원에 지급하는 제도를 말한다.

1950년대 미국의 의료비 지출은 GNP의 4%정도에 불과했다. 그러나 1995년에는 인구의 노령화로 인해 의료비 지출은 GNP의 14%로 상승하게 되었고, 노인을 위한 공적 의료제도인 Medcare의 비용도 1965년 60억불에서 1993년 1,100억불로 상승하였다. 이후에도 의료기술의 진보와 고령자의 증가로 인한 의료비는 계속 증가되고 있다. 마찬가지로, 우

리나라에서도 의료 기술의 향상, 인구의 노령화 등으로 인해 의료비 상승이 문제시 되고 있다. 통계청 발표 자료[1]에 의하면 우리나라도 2026년 초 고령화 사회가 되며, 2050년에는 전체 인구의 약 37%가 65세 이상의 노인이 차지하게 될 것이라고 보고된 바 있다. 이와 같은 이유로 인해 보험을 지불하는 쪽에서는 상승하는 의료비를 억제하기 위해 진료비 지불 절차를 까다롭게 하는 조치를 취하게 되었고, 의료기관은 삭감으로 인한 적자를 예방하고 적절한 이익을 얻을 수 있는 방법으로 입원 기간을 단축하면서 검사나 처치 등 필요한 서비스는 중복이나 과잉 서비스가 공급되지 않도록 요구되어진다. 여기에 CP의 활용이 도움을 줄 수 있다.

기존의 CP에 대한 연구는 임상 전문의들의 지식을 토대로 문서화 중심의 연구가 진행되어 왔다. 그러나, CP는 병원의 특색에 따라 틀려질 수 있고, 시간이 지남에 따라 내·외부적인 요인에 의해 변화될 수 있다. 따라서, 기존의 문서화 중심의 CP는 변화하는 환경에 적응하기 힘든 문제점이 생기게 된다.

본 연구에서는 기존의 진료 데이터들을 토대로 의료 정보 시스템의 한 부분으로서 동적으로 CP를 생성할 수 있는 시스템에 대해 연구하였다. 진료의 단계를 세분화하여 각 단계에 대해서 규칙을 생성하고, 이 규칙들을 세분화된 진료단계에서의 CP를 위한 기초 데이터로 보았다. 각 단계의 규칙을 생성하기 위해서 연관규칙학습 알고리즘[2]을 이용하였다.

본 논문의 구성은 2절에서 알고리즘에 대해 설명하고, 3절에서 연구방법을 설명하고, 4절에서 실험결과 보여주고, 5절에서 결론을 맺는다.

2. Apriori 알고리즘

<표-1>과 <표-2>의 Apriori[3]알고리즘을 이용해서 빈발항목 집합을 만들게 되고, 최소 신뢰도를 이용해 연관규칙을 생성한다.

<표-1> Apriori-gen 알고리즘

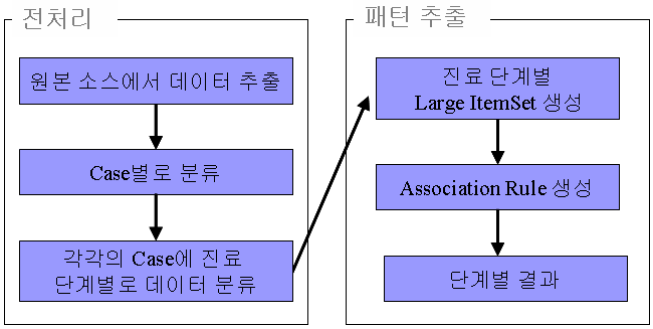
```
Input :
  L_{i-1}    // i-1 크기의 Large Itemset
Output :
  C_i        // i크기의 후보 집합
Apriori-gen algorithm :
  C_i = 0;
  for each I∈L_{i-1} do
    for each J≠I∈L_{i-1} do
    if i-2 of the elements in I and J equal then
      C_k = C_k∪{I∪J};
```

<표-2> Apriori 알고리즘

```
Input :
  I      // 아이템 집합
  D      // 트랜잭션들의 데이터
  s      // 지지도
Output :
  L      // Large itemsets
Apriori algorithm :
  k = 0;   // 스캔 번호
  L = 0;
  C_1 = I  // 초기 후보 집합
 repeat
   k = k+1;
   L_k = 0;
  for each I_i∈C_k do
    c_i = 0;
   for each t_j∈D do
    for each I_i∈C_k do
     if I_i∈t_j then
       c_i = c_i + 1;
    for each I_i∈C_k do
     if c_i≥(s×|D|) do
       L_k = L_k∪I_i;
    L = L∪L_k;
    C_{k+1} = Apriori-Gen(L_k)
 until C_{k+1} = 0;
```

3. 연구방법

본 논문에서 구현된 시스템은 크게 전처리와 패턴 추출을 위한 알고리즘 부분으로 나누어진다. 그 구조는 (그림-1)과 같다.



(그림-1) 시스템 구조

3.1 데이터 전처리

전처리 단계는 크게 3가지 부분으로 나누어 볼 수 있다. 원본 소스에서 데이터 추출하는 단계에서는 데이터에서 환자와 관련된 정보를 익명화 하고, 불필요한 데이터들을 삭제하는 작업을 수행한다. 그 이후 상별별로 데이터들을 분류하게 된다. 상병별로 데이터를 분류하는 이유는 각각의 상병들에 대해서 검사 패턴을 찾기 위해서 이다. 최종 단계에서는 각각의 상병들에 대해 진료 단계별로 데이터를 분류하는 단계를 거쳐 실제 알고리즘이 이용하는 데이터의

형태로 데이터가 만들어진다. 데이터 셋은 아래와 같다.

(환자 구분, 검사나 처방에 대한 아이템 집합)
ex) (1, “J656541,J656551,J10101B,..”)

3.2 패턴 추출

패턴 추출 과정은 전처리 단계에서 생성된 각각의 데이터 셋을 기반으로 패턴에 사용될 아이템 셋을 추출하고 이를 이용해 규칙을 생성하며, 각각의 진료 단계별로 생성된 규칙을 구분 짓는 역할을 한다. 본 연구에서는 Apriori[3] 알고리즘을 이용하여 빈발 항목집합들(Large Itemsets)을 추출하였다. 환자들의 집합을 하나의 트랜잭션으로 보고, 그 환자들의 검사나 처방과 같은 진료내역에 대한 아이템 집합을 Apriori 알고리즘을 적용하여 돌려 보았다. 데이터 전처리 과정에서 분류한 상병 확증과 상병확증 후 진료 단계의 2가지 단계에서 빈발 항목집합들을 추출하고 이를 이용하여 연관 규칙을 생성하였다.

4. 실험 및 결과

4.1 진료단계 분리

환자가 병원에 찾은 시점부터 병이 완치된 시점까지의 진료 흐름을 (그림-2)와 같이 몇 가지 단계로 나눠 보았다. 실제 진료 흐름은 다양한 조건에 따라 더 세분화 되지만, 본 연구에서는 진료 흐름을 아래와 같이 5가지의 단계로 보고 진행하였다.

(그림-2) 진료 흐름

본 실험에서는 상병 확증된 시점의 단계와 상병이 확증된 후 병을 치료하는 단계의 2단계로 나누어 실험을 진행하였다.

4.2 데이터

실험에는 D의료원의 특정 상병에 대한 2004년 1년 동안의 Order 데이터들을 토대로 진행 하였다. 내과에 외래진료 데이터 중에서 3가지 상병에 대한 1600건의 데이터를 이용해 실험을 진행했다. 실험에 사용된 상병은 <표-3>와 같다.

<표-3> 실험에 사용된 상병

코드	영문 병명(한글 병명)
I10	Essential (primary) hypertension [본태성(원발성) 고혈압]
I21.3	Acute transmural myocardial infarction of unspecified site [상세불명 부위의 급성 뇌경색]
I64	Stroke, not specified as haemorrhage or infarction [중풍 또는 경색으로 분류되지 않은 뇌졸중]

(그림-3)진료 데이터들을 추출해 낸 화면이다. 환자의 등록번호는 임의의 숫자로 대체하여 환자에 대한 개인 정보를 감췄다. Order Code항목은 실제 진료 항목에 대한 코드이다. J로 시작하는 Order Code는 일반적인 검사를 나타내고 K로 시작하는 Order Code는 방사선 검사를 나타낸다. 그 외에 Order Code는 처방이나 기타 검사, 증상 등을 나타낸다.

상병코드	등록번호	진료일(외래)	진료과		Order Name	Order Code
I64	4	2/28/2004	CV	8	(T) Neck	R22,1-9
I64	4	2/28/2004	CV	9	Lungs	J98,4-1
I64	4	2/28/2004	CV	10	Heart	I51,9-P
I64	4	2/28/2004	CV	18	Chest pain	R07,4-H
I64	4	2/28/2004	CV	19	Electrocardiography	J656541
I64	4	2/28/2004	CV	20	Gated Myocardial SPECT	J60292E
I64	4	2/28/2004	CV	23	Tenormin tab 25mg(현대약품)	TENOR25
I64	4	2/28/2004	CV	24	Elroton SR Cap, 50mg/C(현대약품)	ELROT50
I64	4	2/28/2004	CV	25	Nitroglycerin 0.6mg/t(하나)	NITRO,6
I64	4	3/26/2004	CV	12	Chest pain	R07,4-H
I64	4	3/26/2004	CV	15	Tenormin tab 25mg(현대약품)	TENOR25
I64	4	3/26/2004	CV	16	Elroton SR Cap, 50mg/C(현대약품)	ELROT50

(그림-3) 진료 데이터

4.3 실험 및 결과

실험은 3가지 상병을 대상으로 진행되었다. 아래 표은 I64에 대한 후보 항목집합들(Candidate Itemset)을 나타낸다. 아래는 최소 지지도(Minimum Support)를 30%로 두고 상병 확증 단계와 상병 처치 단계에서의 빈발 항목집합들(Large Itemsets)을 구한 것이다. 실험에서는 최소 지지도(Minimum Support)를 50%와 30%로 두고 결과를 보았다. 최소 지지도(Minimum Support)를 너무 높게 잡으면 실제 유용한 항목들에 대한 데이터들까지도 사라지게 되고, 너무 낮게 잡으면 항목의 수가 많아져 빈발

항목집합을 탐색하는데 많은 시간이 걸리게 된다.

<표-4> I64에 대한 단계별 추출된 진료 빈발 항목집합(Minimum Support 30%)

상병 확증 단계	처치 단계
Palpitations, Dyspnea on exertion, Chest pain, Holter monitoring, Transthorax	Acertil Tab 4mg/T, Aspirin Protect 100/T, Plavix Tab 75mg/T
Dyspnea on exertion, Chest pain, Triiodothyronine, Thyroxine,TSH	

여기서 추출된 빈발 항목집합은 각각의 진료 단계에서 행해지는 각종 검사, 처방등을 나타낸다. <표-4>를 보면 중풍에 대해 추출된 빈발 항목 집합을 보여주고 있다. 상병 확증 단계를 보면, 주로 증상과 검사에 대한 항목들이 나타나고 있고, 처치단계에서는 약 처방이 나타나고 있다.

연관 규칙은 <표-4>에 보여 지는 빈발 항목집합과 최소 신뢰도(Minimum Confidence)를 통해 생성된다. 각각의 규칙들은 진료 항목들에 대한 연관도를 나타내준다.

I64의 빈발 항목집합을 통해 나타난 연관 규칙 중에 “Palpitations, Dyspnea on exertion, Chest pain -> Holter monitoring, Transthorax(50%, 100%)”와 같은 규칙이 나타났는데, 이는 Palpitations와 Dyspnea on exertion, Chest pain 라는 증상이 있을 때 Holter monitoring, Transthorax라는 검사가 지지도 50%에서 100%의 신뢰도를 나타낸다는 것을 보여주고 있다.

실험에서 출력된 자료들은 각 단계에서 의사들의 실제 처방을 토대로 작성된 병원의 임상 데이터를 이용한 각 단계의 CP로 볼 수 있다. 사용된 데이터들의 신뢰도와 더 많은 데이터가 확보된다면 보다 나은 결과를 얻을 수도 있을 것이다.

5. 결론

HIS를 통해 축척되는 방대한 양의 데이터들이 다시 재가공 되어 활용되는 경우에 대한 연구는 거의 없었다. 본 연구에서는, 이 데이터들을 활용하는 방법 중 하나로서 CP를 동적으로 추출할 수 있는 시스템에 다뤘다. 각 병원마다 사용되는 CP는 환경적 요인에 따라 달라질 수 있다. 따라서, 실제 병원에 축척된 임상 데이터를 바탕으로 CP를 추출하고 이를 활용하는 방안에 대한 연구를 하게 되었다. CP는 포괄수가제의 적용에 대한 대비, 의료 서비스 향상, 경영 효율성 증대에 효과적인 방향을 제시해 줄 수 있는 대안이 될 수 있다.

CP를 추출하기 위해 연관규칙 추출 알고리즘을 이용하였고, 각각의 상병에 대해 진료 단계를 두고 접근하였다. 각 단계에서 추출된 필수적인 검사나 처방에 대한 파악을 통해 생성된 연관규칙들을 CP로 활용 할 수 있는 시스템을 구현하였다.

현재 구현된 시스템은 처방이나 검사에 대한 이름만을 대상으로 데이터를 처리한다. 그러나, 이러한 방법은 여러 가지 문제점을 발생시킬 소지가 있다. 검사의 경우를 보면, 결과에 따라 재진시 검사나 처방이 바뀌는 경우가 발생할 수 있다. 이러한 경우를 대처하기 위해서 검사의 결과에 따른 경로가 재 설정 되어야 할 필요가 있다. 이에 대한 연구가 진행된다면 실제 HIS와 통합하였을 때 좀더 신뢰성 있는 시스템이 될 수 있을 것으로 기대된다.

Acknowledgment

본 연구는 산업자원부 지역연구개발클러스터 구축사업 중 경북대학교 첨단 진단/예측 의료기술 클러스터 사업단의 연구비 지원을 받아 수행되었음.

참고문헌

[1] “장래인구 특별추계 결과”, 통계청, 2005.

[2] R. Srikant and R. Agrawal, "Mining Generalized Association Rules", VLDB, 1995.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for mining association rules", VLDB, pp.487-499, Sept, 1994.

[4] M. Houtsam and A. Swami, "Set-Oriented mining for association rules", IBM Research Report RJ 9567, IBM Almaden Research Center, Oct, 1993.

[5] R. Srikant and R. Agrawal, "Mining Generalized Association Rules", VLDB, 1995.

[6] M. H. Dunham, “Data Mining Introductory and Advanced Topics”, Prentice Hall, 2002.

[7] T. Mitchell, Machine Learning, McGraw-Hill, 1997.