연관규칙을 이용한 의료데이터 마이닝

임준호*, 조태원**, 강재우** +
*고려대학교 컴퓨터정보통신대학원 디지털정보미디어공학과
**고려대학교 정보통신대학, +교신저자
e-mail:{troubless, twjoh, kangj}@korea.ac.kr

Mining Association Rules From Medical Records

Junho Lim*, Taewon Joh**, Jaewoo Kang** ***
*Dept of Digital Information & Media Engineering, Graduate School of Computer & Information Technology, Korea University
**College of Information and Communications, Korea University
+corresponding author

요 약

정보화 시대를 거치면서 모든 산업분야에서 대량의 데이터가 생성되고 관리되고 있다. 최근에는 비즈니스 환경의 변화로 인하여 의사결정을 지원할 수 있는 고급 정보에 대한 필요성이 대두되었으며 IT 기술의 발전과 더불어 데이터마이닝에 대한 많은 연구가 활발히 이루어졌다. 데이터마이닝은 금융, 정부, 제조, 유통 등 다양한 분야에서 활용되고 있다. 한편 의료데이터는 다른 산업분야의 데이터와 구별되는 특징이 있는데, 데이터의 이질성과 복잡성, 부정확성과 오류가능성, 불완전성과 윤리 및 법적인 문제, 개인정보보호, 특징 선택의 제한, 모델의 투명성과 설명력에 대한 높은 요구도 등이 그것이다. 이와 같은 이유로 의료데이터에 대한 접근은 제한적일 수 밖에 없다. 그럼에도 병원 전산화를 통해 발생하는 의료데이터의 양은 기하급수적으로 증가하고 있으며, 임상정보를 포함하는 의료데이터는 데이터 자체로도 가치가 매우 크다.

이에 본 논문은 국내 제 3차 의료기관의 2년간 내원환자에 대한 진단데이터를 사용하여 데이터마이닝의 연관법칙을 이용, 상병간의 관계를 연구하고자 하였다. 이를 통해 잠재고객에게는 객관화된 의료지표를 제공하고, 의료기관은 예측 가능한 정보를 종합의료시스템에 활용하여 고객만족도를 높이는 효과를 볼 수 있을 것으로 사료된다.

1. 서론

비즈니스 환경의 변화에 따라 기업은 의사결정 지원을 위한 고급정보를 필요로 하게 되었으며, 비즈니스 우위를 위해 기존의 대용량 데이터베이스의 조회방법보다 우수한 분석모델을 통한 예측 데이터를 필요로 하게 되었다. 데이터마이닝은 대량의 가공하지 않은 데이터로부터 알려지지 않은 새로운 정보나 유용한 패턴과 상관관계를 추출하여 의사 결정에 이용하는 작업으로써 최근 H/W와 S/W를 비롯한 IT의 기술적 발전과 더불어 많은 연구가 이루어져 왔다. 데이터 마이닝은 보험사기 색출[1], 이탈고객 모델링[2], 타겟마케팅, 교차판매, 상승판매, 상품 디스플레이, 위험관리, 웹마이닝(Web Mining), 동적웹페이지 구성 등 다양한 산업분야에서 연구 및 활용되고 있다.

그러나 의료데이터는 다른 산업분야의 데이터와는 구별되는 특징을 가지는데, 데이터의 이질성과 복잡성, 부정확성과 오류 가능성, 불완전성과 윤리 및 법적인 문제, 개인정보보호 등이 그것이다. 또한 분석시에는 특징 선택의 제한, 모델의 투명성과 높은 설명력을 필요로 한다.[3] 의료서비스 분야의 데이터마이닝은 의료데이터 자체가 갖는 기술적 문제로부터 윤리 및 사회학적인 문제까지 다양한 복잡한 문제를 안고 있어서 데이터에 대한 접근 자체가 제한적일 수 밖에 없다. 그럼에도 병원 전산화를 통하여 발생하는 의료데이터의 양은 기하급수적으로 증가하고 있으며, 환자에 대한 임상정보를 포함하는 의료데이터는 임상지식의 보고로서 그 가치가 매우 크다.

본 연구의 목적은 국내 제 3차 의료기관의 상병정보를 포함한 의료데이터를 통해 데이터마이닝의 기법인 연관규칙을 이용해 상병간의 연관관계를 찾아내고 이를 모형화함으로써 잠재고객에게는 객관적인 의료지표를 제공하는 한편 향후 발생될 상병을 예방하여 국민 건강 증진에 기여하고, 의료기관에서는 예측 가능한 데이터를 종합의료시스템에 활용하여 고객만족도를 높이는데 있다.

본 논문의 구성은 다음과 같다. 제 1장에서는 본 연구를 수행하게 된 배경과 목적을 살펴보고 본 연구의 구성을 제시한다. 제 2장에서는 연관규칙의 기본 개념, 그리고 관련연구에 대해서 알아본다. 제 3장에서는 데이터마이닝 수행절차와 분석하는 과정을 기술한다. 제 4장에서는 데이터마이닝 결과를 분석하며, 제 5장에서는 결론 및 향후 연구 방향에 대하여 기술한다.

2. 배경 및 관련 연구

데이터마이닝 기술은 다양한 의학 분야에서 적용되고

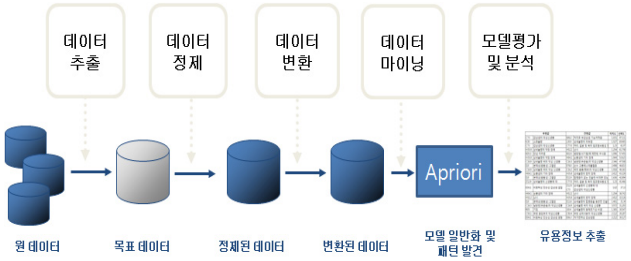
있으며, 다음과 같은 예가 있다. 바이오메디컬 및 임상기록으로부터 질병-약에 관한 지식을 자동으로 획득하는 것에 대한 연구[4], 당뇨병 치료에 관한 전산기록의 마이닝을 통해 일반의가 시행한 치료의 적절성 여부에 대한 판단[5], 개인용 건강 기록 시스템을 통한 데이터 분석 및 저장 등이 그 예이다. 그리고, 많은 연구들이 환자에 대한 진단 데이터를 데이터셋으로 사용하고 있다. 그 예도 다양한데, 환자의 상태를 체크해 주는 진단 데이터를 의사결정 트리 기법으로 마이닝하여 환자의 상태에 대한 진단 패턴을 추출하는 연구[6], 불임환자의 내원기록을 기초로 베이지안망 기법을 사용하여 불임의 주요원인을 추출해 낸 연구[7] 등을 그 예로 들 수 있다.

본 연구에서도 환자에 대한 진단 데이터를 데이터셋으로 사용하여 상병 간의 상관관계를 추출하고자 하였다. 사용한 데이터마이닝 기법은 연관규칙(Association Rule)이다.

연관규칙은(Association Rule)은 대용량 데이터베이스에서 아이템들 간의 유용한 연관 패턴을 찾아내는 것으로 연관규칙을 찾는 대표적인 기법으로는 Apriori 알고리즘이 있다.[8] 데이터베이스는 여러 트랜잭션으로 이루어져 있는데, 각각은 아이템의 집합으로 이루어져 있다. 이때, 아이템셋 X와 Y에 대한 연관규칙 R은 조건부와 결과부로 구성되며 아이템셋인 X와 Y에 대하여 '어떤 트랜잭션이 아이템셋 X를 포함할 때 아이템셋 Y 또한 함께 포함한다'는 의미로 다음과 같이 표현할 수 있다. R : X ⇒ Y 여기서 X, Y ⊆ I(전체아이템셋)이고, X∩Y = Φ 이다. 이때 최종적으로 찾아낸 연관 규칙은 사용자에 의해 주어진 최소지지도(Minimum Support)와 최소신뢰도(Minimun Confidence)를 충족한다.

3. 설계 및 구현

본 연구는 데이터 추출, 데이터 정제, 데이터 변환, 데이터마이닝, 모델평가 및 분석의 순으로 진행되었다.



(그림 1) 데이터마이닝 수행절차

연구에 사용된 데이터는 국내 제 3차 의료기관인 A병원의 2008년 9월 1일에서 2010년 8월 31일까지 2년간의 81,746명의 환자에 대한 진단 정보로써 전체 835,418건의 트랜잭션 데이터 추출을 통해 목표데이터를 구성하였다. 추출된 목표데이터에서 테스트 데이터를 포함한 결측값은 전체 데이터의 5% 미만으로 Case Deletion 방법을 이용하여 데이터 정제과정을 거쳤다. 본 연구는 상병간의 상관관계를 밝혀내는 것이 목적이므로 환자의 식별 가능한 ID를 기준으로 상병정보의 정렬을 위한 데이터 변환과정을 수행하였다. 위의 과정에서 얻어진 변환데이터는 4,461개의 상병코드를 포함하는 전체 70,975명의 환자에 대한 레코드로 구성된 데이터이다. 데이터마이닝 과정은 SPSS사의 PASW Modeler13의 Apriori 알고리즘을 이용하여 수행하였다. 최소지지도 1.03%, 최소신뢰도 30.10% 를 적용한 결과 20개의 유효한 규칙이 생성되었다.

4. 결과

<표 1> 상병간의 연관관계 규칙 (%)

후항값		전항값		지지도	신뢰도
C73	갑상샘의 악성신생물	E89.2	처치후 부갑상샘 기능저하증	1.031	97.131
K30	소화불량	L30.9	상세불명의 피부염	1.527	69.465
C73	갑상샘의 악성신생물	C77.0	머리, 얼굴 및 목의 림프절속발성	1.23	61.97
H35.9	상세불명의 망막 장애	H52.2	난시	1.254	61.798
K05.3	만성 치주염	K02.0	법랑질(사기질)에 제한된 우식증	1.096	57.455
H35.9	상세불명의 망막 장애	H04.1	눈물샘의 기타 장애	1.944	53.623
C16.9	상세불명 위의 악성 신생물	C16.3	날문방(유문동)의 악성신생물	2.346	47.508
I10	본태성(원발성) 고혈압	E78.0	순수 고콜레스테롤혈증	2.498	46.813
C16.9	상세불명 위의 악성 신생물	C16.2	위의 몸통(체부)의 악성신생물	1.918	46.363
H04.1	눈물샘의 기타 장애	H35.9	상세불명의 망막 장애	2.412	43.224
I10	본태성(원발성) 고혈압	E11.9	합병증이 없는 인슐린-비의존 당뇨	1.636	42.894
Z12.9	상세불명의 신생물에 대	C77.0	머리, 얼굴 및 목의 림프절속발성	1.23	41.466
E04.1	비중독성 단순성 갑상샘 결절	Z12.9 C73	상세불명의 신생물에 대 갑상샘의 악성신생물	1.62	37.13
H04.1	눈물샘의 기타 장애	H52.2	난시	1.254	36.742
H52.2	난시	H35.9	상세불명의 망막 장애	2.412	32.126
I10	본태성(원발성) 고혈압	E11.8	상세불명의 합병증을 동반한 인슐	1.461	31.34
C16.3	날문방(유문동)의 악성신생물	C16.9	상세불명 위의 악성 신생물	3.572	31.203
R05	기침	J30.4	상세불명의 알레르기성 비염	1.365	30.547
C50.1	유방 중앙부의 악성신생물	C50.4	유방 상외사분의 악성신생물	2.112	30.287
E04.1	비중독성 단순성 갑상샘 결절	E06.3	자가면역성 갑상샘염	2.212	30.127

첫 번째의 E89.2(처치후 부갑상샘 기능저하증) ⇒ C73(갑상샘의 악성신생물) 연관규칙은 지지도 1.031%, 신뢰도 97.131로 나타났다. 이는 변환된데이터의 전체 70,795명의 환자 중 E89.2(처치후 부갑상샘 기능저하증)과 C73(갑상샘의 악성신생물)이 함께 진단될 확률이 1.031% 임을 의미하며, E89.2(처치후 부갑상샘 기능저하증)를 진단받은 환자의 경우 C73(갑상샘의 악성신생물 진단)이 함께 진단될 확률이 97.131% 임을 의미한다. 그 원인은 다음과 같다. C73(갑상샘의 악성신생물(갑상선암))의 치료는 종양을 제거하기 위해 갑상샘을 제거하는 수술을 하게된다. 부갑상샘의 해부학적 위치는 갑상샘의 뒤편에 밀착하여 위치하므로 갑상샘을 제거하는 수술과정에서 분리되지 못하고 같이 제거되는 경우가 많으며 그 결과 부갑상샘에서 분비되는 부갑상샘 호르몬의 결핍으로 인해 기능저하증이 발병하게 된다. 따라서, 갑상샘의 악성신생물로 갑상선 절제술을 시행받은 환자들의 대부분에서 처치후 부갑상샘 기능저하증이 발생하게 된다.[9][10]

두 번째의 L30.9(상세불명의 피부염) ⇒ K30(소화불량) 연관규칙은 지지도 1.527%, 신뢰도 69.465% 를 나타내는데, 이는 피부염의 주된 치료제인 스테로이드가 위염, 위궤양 등의 소화기 질환을 유발하기 때문이다. 기존 연구에서 스테로이드가 투여되면 24시간 이내 위산 분비가 50~100% 증가하여 위궤양 발생 위험이 높아진다고 보고되었다. 그래서, 임상에서는 피부염 환자에 스테로이드 약물을 투여 시 위궤양 등의 소화기 질환을 예방하기 위해 위산

억제제 및 소화제 등을 같이 처방하고 있다.[11][12][13]

여덟 번째의 E78.0(순수 고콜레스테롤혈증) ⇒ I10(본태성(원발성) 고혈압)의 연관규칙은 지지도 2.498(%), 신뢰도 46.813(%)으로 빈번하게 발생하며 높은 상관관계가 있음을 보여준다. 고콜레스테롤혈증과 고혈압의 상관관계는 이미 잘 알려진 사실이며, 실제 연구에서 고콜레스테롤혈증을 가진 그룹에서 고혈압 발생위험이 1.3배 (p-value<0.0001) 높은 것으로 보고되었다.[14][15]

실험결과 도출된 상병간의 연관관계 규칙의 유의성 평가를 위하여 트랜잭션 원데이터로부터 A 병원에서 빈번하게 발생하는 상위 20개의 상병리스트를 구성하여 비교분석 하였다.

<표 2> 상위 20개의 상병리스트

상병코드	상병명	발생건수	발생비율
C73	갑상샘의 악성신생물	41,076	0.049168201
C50.1	유방 중앙부의 악성신생물	22,162	0.026528037
D24	유방의 양성신생물	20,186	0.024162754
I10	본태성(원발성) 고혈압	19,762	0.023655224
C16.9	상세불명 위의 악성 신생물	17,870	0.02139049
E04.1	비중독성 단순성 갑상샘 결절	17,824	0.021335427
C34.9	상세불명의 기관지 또는 폐의 악성신생물	17,688	0.021172635
C50.4	유방 상외사분의 악성신생물	16,930	0.020265304
C53.9	상세불명 자궁목의 악성신생물	15,888	0.019018025
C20	직장의 악성신생물	14,282	0.017095634
C16.3	날문방(유문동)의 악성신생물	12,750	0.015261821
Z12.9	상세불명의 신생물에 대	11,068	0.013248458
C16.2	위의 몸통(체부)의 악성신생물	10,832	0.012965964
C50.9	상세불명 유방의 악성신생물	9,060	0.01084487
C50.2	유방 상내사분의 악성신생물	8,866	0.010612651
C22.0	간세포 암종	8,715	0.010431904
E78.0	순수 고콜레스테롤혈증	8,445	0.010108712
K30	소화불량	8,022	0.009602379
E05.0	미만성 갑상샘종을 동반한 갑상샘중독증	7,218	0.008639986
E04.2	비중독성 다결절성 갑상샘종	7,165	0.008576545

<표 1>에서 C73(갑상샘의 악성신생물) 진단이 전체 트랜잭션 835,418건 중 41,076회 발생하여 발생비율 4.9%로 가장 빈번하게 발생하였음을 보여준다. <표 1> 상변간의 연관관계 규칙과 <표 2> 상위 20개의 상병리스트를 비교한 결과 최상위 C73(갑상샘의 악성신생물 진단), C50.1(유방 중앙부의 악성신생물), I10(본태성(원발성) 고혈압), C16.9(상세불명 위의 악성 신생물), E04.1(비중독성 단순성 갑상샘 결절) 등을 비롯한 <표 2>의 대부분의 상병이 <표 1>의 상병간의 연관관계 규칙에도 나타남으로써 A 병원에서 빈번하게 발생하는 상병들에 대한 연관규칙을 다수 발견하였다. 이는 본 연구의 목적이 이미 발생한 상병에 이어서 발생할 수 있는 다른 상병들의 예측을 위한 분석 시스템이라고 볼 때, 보다 빈번한 발생빈도를 가지는 상병들에 대한 경우를 예측할 수 있으므로, 효용성이 비교적 높다고 생각할 수 있겠다.

하지만 희귀질병 같은 빈도수가 적은 상병들은 낮은 지지도와 낮은 신뢰도를 가진 상병은 분석에서 제외될 수 밖에 없다는 단점을 발견하였다. <표 2>의 세 번째로 빈번하게 발생한 D24(유방의 양성신생물) 진단은 <표 1>에 나타나지 않는 데, 이는 D24(유방의 양성신생물)를 포함하는 연관규칙의 지지도가 최소지지도를 충족하지 못하거나 신뢰도가 최소신뢰도에 미치지 못함을 의미한다.

5. 결론 및 향후과제

본 연구결과를 토대로 얻어진 상병간의 관계를 병원 홈페이지의 전면부에 배치하고 지속적으로 업데이트되는 동적인 웹페이지를 구성하여 잠재고객에게는 객관화된 의료 지표를 제공할 수 있을 것이다. 또한 CRM(Customer Relation Management)과 연계하여 특정 상병 진단을 받은 고객에게 연관성 높은 상병에 대한 정보를 제공하고, 그에대한 예방책을 알려줌으로써 고객만족도를 높일 수 있을 것으로 사료된다.

본 연구의 문제점 및 향후 과제는 다음과 같다. 첫째, 대용량의 진료데이터에서 상병간의 연관성 규칙을 찾아보고자 하였기에 지지도와 신뢰도가 낮은 규칙들의 관계는 배제되었다. 둘째, 높은 지지도와 신뢰도를 가지는 규칙들은 주 질병과 그에 따른 합병증으로 나타난 것이 대부분이다. 향후의 연구에서는 몇 가지 대표적인 질병군을 분류하여 질병군 별로 연관관계 패턴을 알아봄으로써 보다 상세한 규칙에 대한 연구가 필요하다. 셋째, 본 연구는 환자에 대한 진료데이터만 가지는 변환된데이터를 이용하여 상변간의 연관관계를 확인하였는 데, 환자의 기초데이터와 질병의 특성 정보를 포함하여 다차원 연관관계 분석이 필요하다.

참고문헌

[1] 성태경, 배문준, "Data Mining Tool을 활용한 보험사기 적발", 한국경영정보학회 추계학술대회, 2002.

[2] 홍태호, 전성용, "데이터마이닝을 이용한 고객이탈등급에 기반한 고객 세분화", 한국정보시스템학회 추계학술대회, 2005.

[3] 이선미, 박래웅, "상에서의 데이터 마이닝 개념과 원칙", 대한의료정보학회, 제15권 제2호, pp. 175-189.

[4] Elizabeth S. Chen, PhD, George Hripcsak, MD, MS, Hua Xu, MS, Marianthi Markatou, PhD, and Carol Friedman, PhD., "Automated Acquisition of Disease-Drug Knowledge from Biomedical and Clinical Documents: An Initial Study," Journal of the American Medical Informatics Association, 2008, 15 (1), pp. 87-98

[5] Jaco Voorham and Petra Denig., "Computerized Extraction of Information of the Quality of Diabetes Care from Free Text in Electronic Patient Records of General Practitioners," Journal of the American Medical Informatics Association, 2007, 14 (3), pp. 349-354

[6] 권은희, 이승철, 이주창, 김응모, "개인 맞춤형 의료진단 서비스 제공을 위한 효율적인 데이터마이닝 기법", 한국정보과학회 추계학술발표회, 2007

[7] 정용규, "의료데이터마이닝을 위한 특징 축소와 베이지안망 학습", 한국정보과학회 춘계학술발표회, 2004

[8] Bing Liu, "Web Data Mining," Springer

[9] Testini M, Gurrado A, Lissidini G, Nacchiero M., "Hypoparathyroidism after total thyroidectomy," Minerva

Chir. 2007 Oct;62(5):409-15
[10] Godlewska P, Kaniewski M, Stachlewska-Nasfeter E, Bisz D, Lyczek J., "Parathyroid hypofunction after total thyroidectomy for differentiated thyroid carcinoma-perspectives after long term observation and treatment," Wiad Lek, 2001;54 Suppl 1:398-404. Review. Polish.
[11] Usatine RP, Riojas M., "Diagnosis and management of contact dermatitis," Am Fam Physician, 2010, Aug 1;82(3):249-55.
[12] Bandyopadhyay U, Biswas K, Bandyopadhyay D, Ganguly CK, Banerjee RK., "Dexamethasone makes the gastric mucosa susceptible to ulceration by inhibiting prostaglandin synthetase and peroxidase-two important gastroprotective enzymes," Mol Cell Biochem, 1999, Dec;202(1-2):31-6.
[13] Bleichner G, Mignon F, Mignon M., "Current clinical and physiopathological aspects of gastroduodenal lesions due to corticotherapy," Ann Med Interne (Paris), 1974, Dec;125(12):903-12. French.
[14] Gotto AM Jr., "Hypertriglyceridemia: risks and perspectives," Am J Cardiol, 1992, Dec 14;70(19):19H-25H.
[15] Yao XG, Frommlet F, Zhou L, Zu F, Wang HM, Yan ZT, Luo WL, Hong J, Wang XL, Li NF., "The prevalence of hypertension, obesity and dyslipidemia in individuals of over 30 years of age belonging to minorities from the pasture area of Xinjiang," BMC Public Health, 2010, Feb 24;10:91.