데이터마이닝을 이용한 의료사기 탐지 시스템

이준우[A], 지원철[B], 박하영[C], 신현정[A1)]

[A]아주대학교 산업정보시스템공학부, [B]홍익대학교 정보컴퓨터공학부, [C]서울대학교 공과대학

Medical Fraud Detection System Using Data Mining

Junwoo Lee,[A] WonChul Jhee Ph.D.[B] HaYoung Park, Ph.D.[C] Hyunjung Shin Ph.D.[A]

[A] Dept. of Industrial Engineering, Ajou University, Suwon 443-749, {neo100in, shin}@ajou.ac.kr

[B] Dept. of Industrial Engineering, Hongik University, Seoul 121-791, jwc118@naver.com

[C] Dept. of Industrial Engineering, Seoul National University, Seoul 151-742, hayoungpark@snu.ac.kr

Abstract: 본 연구는 데이터마이닝 기법을 이용하여 건강보험청구료에 있어서 이상정도가 심한 요양기관을 탐지하고, 실제 의료영역에 적용하기 위한 시스템 개발을 목적으로 한다. 현재 건강보험 심사평가원의 이상탐지시스템은 평가대상이 되는 항목을 개별적으로 평가하고, 탐지된 기관의 선정 이유에 대한 근거제시가 부족한 단점을 가지고 있다. 따라서 본 연구에서는 항목을 종합적으로 평가할 수 있는 정량적 지표를 설계하고, 항목들의 상대적 중요도를 파악할 수 있도록 항목들에 대한 가중치 부여한다. 또한 지표에서 얻어진 값으로 등급을 구분하고, 의사결정나무기법(decision tree)를 이용하여 해석력을 높이는 방법을 제시한다.

Keyword: *Bill Claim, Health Insurance Review Agency, Fraud Detection, Degrees of Anomaly, Health care*

1. 서론

최근 우리나라 국민의료비는 급속하게 증가하고 있고 2002년 요양급여비용[2] 총액은 13조 8천억 원에서 2007년 32조 2,590억 원으로 5년 동안 2.5배나 상승하였을 정도로 그 상승폭이 매우 크다. 또한 전년도에 비교하여 요양급여비용 총액이 13% 증가하였는데 이러한 경향은 소득수준의 향상과 노인인구의 증가, 그리고 의학시술의 발전 및 의료비 상승 등으로 지속적으로 상승할 것으로 예상된다. 하지만 사회발전이라는 측면이 아닌 다른 부정적인 요인에서도 의료비 상승의 이유를 찾을 수 있는데 요양기관[3]의 진료 오·남용과 요양급여비의 부당청구가 이에 해당한다. NHCAA(National Health Care Anti-fraud Association)의 보고에 따르면 미국 전체 의료비용 중 3~10(600억 달러 ~1600억 달러)%가 의료사기에 의한 손실로 추정하고 있다. 이에 비추어볼 때, 우리나라의 요양급여비 부당청구에 의한 손실 규모 또한 상당 수준에 이를 것으로 예상할 수 있다.

2. 방법론

2.1 이상치 지표(Degree of Anomaly)

개별 항목 위주의 문제점과 순위 위주의 문제점을 개선하기 위하여 정상에서 벗어난 이상 정도까지도 측정할 수 있는 정량지표와 각 개별 항목들을 통합할 수 있는 종합지표가 필요하다. 식(1)은 정상에서 벗어난 이상정도를 측정할 수 있도록 다음과 같은 변수 변환식을 제안한다.

---

1) 교신저자: 신현정 아주대학교 산업정보시스템공학부 shin@ajou.ac.kr

2) '요양급여비용'이란 정부가 국민건강보험공단을 통하여 요양기관 및 국민에게 제공하는 의료비용을 의미함.

3) '요양기관'이란 종합병원·일반병원·의원(치과, 한의원 포함)·약국 등을 통칭함.

$$e^{\left[\frac{\max(x_{ij}-\mu_j,0)}{\sigma_i}\right]^2} \quad (1)$$

식(1)에서 평균이상의 값을 가지는 요양기관은 평균에서 멀어지는 정도에 따라 큰 값을 부여받고, 평균 이하의 값을 가지는 요양기관은 1의 값을 갖도록 한다. i는 record index로서 요양기관을 나타내고, j는 심사에 사용되는 평가지표들을 나타낸다. 즉 x_{ij}는 요양기관 i의 평가지표 j값을 의미한다. μ_j와 σ_j는 각각 평가지표의 평균과 표준편차를 나타낸다. 따라서 평균까지는 1의 값을 가지지만 평균보다 클 경우에는 평균에서 멀어질수록 매우 큰 값을 가지게 되고, [그림3]의 분포를 통하여 그 수는 매우 적음을 알 수 있다.

또한 개별적 지표가 아닌 종합적 지표로서 식(2)를 제안한다.

$$D_A(x_i) = \frac{\sum_{j=1}^{|J|} w_j e^{\left[\frac{\max(x_{ij}-\mu_j,0)}{\sigma_j}\right]^2}}{\sum_{j=1}^{|J|} w_j} \quad (2)$$

제안된 지표를 통하여 나오는 값을 D_A 값이라 명명하고 있는데, 식(2)는 식(1)에서 얻은 값과 입력변수 가중치와 곱하여, 그 합을 통하여 얻을 수 있다. 입력변수 가중치(W_j)에 대한 설명은 4.2절에서 하기로 한다.

2.2 입력변수 가중치

입력항목이 모두 같은 중요도를 가진다고 가정하고 부정현황을 탐지하는 것은 현실적으로 많은 문제점을 가지고 있다. 따라서 입력항목의 중요도를 나타내는 가중치를 얻기 위한 분석은 의료 부정탐지 영역에서 매우 중요한 부분이다. 사용된 분석은t-test, 상관분석, 판별분석, 로지스틱 회귀분석, 불순도 검정, 카이제곱검정 등을 개별적으로 사용하거나, 또한 여러 가지 분석의 합을 합하여 도출할 수 있다. 다음의 식(2)는 j 번째 입력변수에 대하여, 여러 분석에서 얻은 가중치들을 통합하는 과정을 단순화하여 보여준다.

$$w_j = w_j^{\rho} + w_j^{R} + w_j^{t} + w_j^{E} + w_j^{D} + w_j^{\chi} \quad (2)$$

3. 분석

3.1 데이터

본 연구에 사용된 데이터는 건강보험심사평가원의 2007년 하반기 의료보험진료비 청구자료 중 요양현황정보, 청구내역정보, 환자관련정보, 조정관련 정보를 사용하였다. 데이터는 부정 의료시술 공급자(service provider fraud) 탐지에 관련되었다고 보아지는 187개의 변수와 45,700건 정도의 청구현황으로 구성되어 있다. 또한 결측값과 이상치가 다수 포함되어 있으며, 한쪽으로 크게 편향되어 있는 분포로서 명목형 변수와 연속형 변수 모두를 가지고 있다.

3.2 가중치

내과에서 선택된 입력항목과 그에 대한 가중치는 [표1]과 같다. [표 1]에서 보는 바와 같이 38개의 입력항목 중에서 부당청구기관으로 선정하는 지표로서 CI 지표와 제1,2질병 CI는 높은 가중치를 가지고 있어 가장 중요한 지표로 분석되지만, PET CI와 정신요법 CI는 낮은 가중치를 가지고 있어 다른 지표에 비해 중요하지 않은 변수로 분석된다

3.3 등급화

등급은 0등급부터 그 심각성에 따라 4등급까지 나누어 총 5개의 등급으로 세분화 하였고, 각 등급에 해당하는 D_A 등급과 D_A 경계값, 경계로그 값, 해당 기관의 수, 등급별 비율 등이 [표 2]에 나타나 있다.

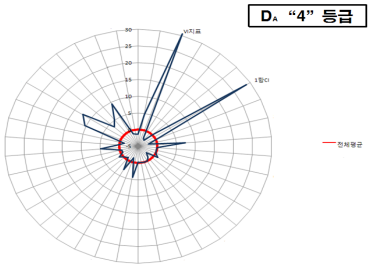
[표 1] 선정된 입력항목과 가중치

	입력항목	가중치
1	약품목수	2.90
2	CI 지표	4.91
3	VI 지표	2.48
4	A_1	3.04
5	A_2	2.29
6	CMI 지표	1.55
7	진찰료 CI	2.67
8	A_3	0.66
9	A4	2.77
10	주사투약료 CI	2.96
11	마취료 CI	1.28
12	A_5	1.48
13	A_6	0.66
14	처치 및 수술료 CI	1.86
15	검사료 CI	1.83
16	A_7	1.64
17	A_8	0.99
18	MRI항 CI	0.65
19	PET항 CI	0.24
20	A_9	2.01
21	A_{10}	2.38
22	투약일당 약품비	2.44
23	고가약 처방비중	1.60
24	A_{11}	2.37
25	A_{12}	2.06
26	소화기용약 처방비율	1.48
27	부신피질-호흡기계처방률	1.30
28	A_{13}	0.76
29	A_{14}	2.50
30	건당 내원일수	1.09
31	건당 투약일수	2.55
32	A_{15}	2.86
33	A_{16}	2.43
34	제 1질병 CI	3.46
35	제 2질병 CI	3.10
36	A_{17}	2.48
37	A_{18}	2.77
38	제 5질병 CI	2.30

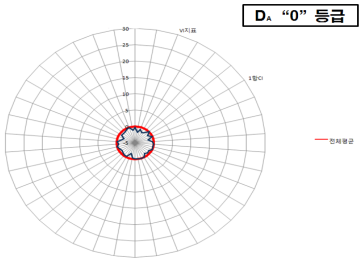
[표 2] D_A 등급에 따른값

등급	D_A	Log(D_A)	빈도수
4	1000+	6.92	236 (6.37%)
3	100+	4.61	154 (4.16%)
2	10+	2.30	570 (15.38%)
1	5+	1.61	403 (10.88%)
0	~5	1.00	2342 (63.21%)

[그림 1]은 D_A 각 등급에 속하는 요양기관들의 입력항목 값과 입력항목 평균값과의 차이를 보여주는 그림이다. 각 등급에 해당하는 요양기관의 입력항목 패턴의 예를 확인 할 수 있는데 도형의 가운데 조그만 원이 평균값을, 그 위의 불규칙한 다각형 모양이 해당 요양기관의 입력항목 각각의 D_A 값을 나타낸다. 원의 격자가 하나의 입력항목을 나타내는데, 위에 언급한 것과 마찬가지로 항목은 비공개한다.



(a) D_A 4등급의 예



(e) D_A 0등급의 예

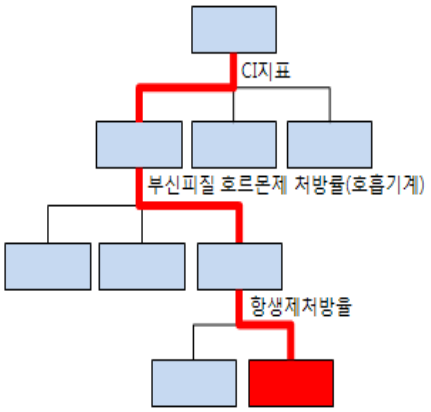
[그림. 1] 등급별 입력변수의 D_A모형

[그림 1(a)]는 D_A 4등급으로서 평균에 비해 VI 지표, 1항CI 등의 항목이 평균에 비교하여 월등히 높음을 알 수 있는데, 이상 정도가 크기 때문에 높은 D_A 값을 받게 되고 4등급에 선정이 되었다. [그림 1(e)]는 D_A 0등급으로 거의 모든 입력항목의 D_A값이 평균보다 작은 값을 가지고 있어, 다각형 모양의 도형이 가운데의 조그만 원 안에 들어있다.

3.4 해석

의사결정나무는 위에서 아래로 가지를 분할하면서 마디를 생성하고, 마디에는 일정한 규칙을 가진 요양 기관들이 분포하게 된다. 아래 [그림 2]에서 붉은 색의 마디는 내과에서 D_A 4등급에 해당하는 요양기관을 보여준다. 요양기관을 분류하는데 가장 중요한 변수로 선택된

'CI지표' 에 의하여 세 개의 마디로 나누어지고, 그 다음은 부신피질 호르몬, 마지막으로 항생제 처방율의 변수로 마디를 분할 한 것을 확인 할 수 있다. 오른쪽으로 이동할수록 높은 등급이 분포하는데, 이 예에서는 CI지표가 낮더라도 다른 특정항목이 동일 과목 내에서 현저히 높을 경우 그 요양기관들은 높은 D_A 등급을 받을 수 있는 경우를 보여준다. 즉 부신피질호르몬제와 항생제의 과다 처방이 문제가 되어 D_A 4등급으로 선정되었다.



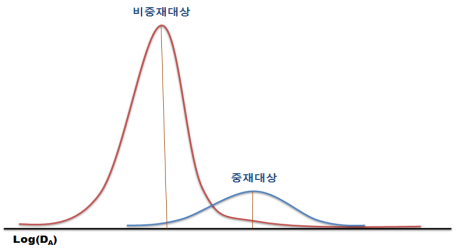
[그림 2] 의사결정나무의 예(1)

4. 검증

제안할 방법은 내과 전체 2007년 4사분기 요양기관들의 중재대상 선정 자료를 참조하여 비중재대상과 중재대상 기관으로 나누고 D_A 값의 분포를 비교 하였을 때 중재대상이었던 기관들의 평균 D_A값이 훨씬 높은 것을 [그림 3]에서 확인할 수 있다. 그림에서 곡선의 높이는 해당 요양기관의 수를 나타내고 가로축은 Log(D_A)값을 의미하는데, 왼쪽에 위치한 곡선이 비중재대상의 곡선이고, 오른쪽에 위치한 곡선이 중재대상을 나타내는 곡선이다. 곡선의 평균 D_A값은 [표 3]에 있다. 중재대상 기관들의 D_A 값이 더 크게 나타나는 것은, 새로 제안된 지표값인 D_A값이 기존 중재대상선정과 많은 부분 일치한다는 사실을 보여주고 있다.

[표 3] 중재 / 비중재대상의 평균 D_A

등급	비중재대상	중재대상
N	3,346	359
D_A 평균	$e^{2.23}$	$e^{3.97}$



[그림 3] 중재 / 비중재대상의 평균 D_A 분포

5. 결론

본 연구에서는 국내 적정의료비 심사제도를 효율화하기 위하여 최신 데이터마이닝·기계학습기법을 사용하여 다양한 형태의 부당청구 및 허위청구 수법을 탐지 · 관리할 수 있는 지표 및 방법론을 제시하였다. 제안된 모형은 분석적 측면에서 기존의 유사지표에 대하여 다음과 같은 이점을 갖는다. 첫째 각 지표별 데이터 분포가 로그정규분포를 따르고 있음을 감안할 때 D_A는 이를 자동적으로 보정해 주는 역할을 하고, 둘째 각 지표별로 데이터가 평균으로부터 이격된 정도를 차등화 하여 이상치 정도를 수치화하여준다. 또한 탐지에 유효한 변수 선정 방법론 개발하여 입력항목에 대한 가중치를 설정하였다.

<참고문헌>

[1]P. Bartlett, S. Ben-David, S. Kulkarni (2000) Learning changing concepts by exploiting the structure of change. Machine Learning, 41: 153-174