불균형 데이터 집합에서의 의사결정나무 추론: 종합병원의 건강 보험료 청구 심사 사례

Decision Tree Induction with Imbalanced Data Set: A Case of Health Insurance Bill Audit in a General Hospital

허 준 (Joon Hur) SPSS Korea (주)데이터솔루션 컨설팅부문 컨설턴트

김 종 우 (Jong Woo Kim) 한양대학교 경영대학 경영학부 부교수, 교신저자

요 약

다른 산업과 달리 병원/의료 산업에서는 건강 보험료 심사 평가라는 독특한 검증 과정이 필수적으로 있게 된다. 건강 보험료 심사 평가는 병원의 수익 문제 뿐 아니라 적정한 진료행위를 하는 병원이라는 이미지와도 맞물려 매우 중요한 분야이며, 특히 대형 종합병원일수록 이 부분에 많은 심사관련 인력들을 투입하여, 병원의 수익과 명예를 위해서 업무를 수행하고 있다.

본 논문은 이러한 건강보험료 청구 심사 과정에서, 사전에 수많은 진료 청구 건 중 심사 평가에서 삭감이 될 수 있는 진료 청구 건을 데이터 마이닝을 통해서 발견하여, 사전의 대비를 철저히 하고자 하는 한 국내 대형 종합병원의 사례를 소개하고자 한다. 데이터 마이닝을 적용함에 있어, 주요한 문제점 중 하나는 바로 지도학습 기법을 적용하기에 곤란한 데이터 불균형 문제가 발생하는 것이다. 이런 불균형 문제를 해소하고, 비교 조건 중에 가장 효율적인 삭감 예상 진료 건 탐지 모델을 만들어 내기 위하여, 데이터 불균형 문제의 기본 해법인 Sampling과 오분류 비용의 다양한 혼합적인 적용을 통하여, 적합한 조건을 가지는 의사결정 나무 모델을 도출하였다.

*키워드 : 데이터 불균형, 건강보험심사평가, 의사결정나무 분석, **Sampling**, 오분류 비용*

Ⅰ. 서 론

데이터 마이닝의 지도학습 기법(supervised learning)을 수행하는 과정 중에 발생하는 문제 중 하나가 데이터 불균형(imbalanced data) 문제이다. 데이터 불균형 문제는 목표 변수(target variable)가 이분형이고, 목표변수의 범주에 속하는 데이터 수의 비율이 현격히 차이가 나는 경우를 의미한다(오장민, 장병탁, 2001). 일반적으로, 목표변수의 빈도 분포는 동등한 비율로 존재하는 것이 지도학습 기법을 통해서 패턴을 인식하는데 좋은 것으로 알려져 있다. 이는 지도학습 기법을 적용할 때, 한 쪽의 범주가 비정상적으로 큰 경우, 지도학습 모델은 전체적인 오분류를 작게 하기 위해서, 다수의 범주로 패턴 분류를 많이 하게 되고,

† 이 논문은 한양대학교 일반연구비 지원으로 연구되었음(HY-2006-G)

이 경우 소수의 범주는 다수의 범주로 취급되기가 쉽기 때문이다(Weiss and Provost, 2001). 따라서 데이터 마이닝을 이용하여, 올바른 패턴 인식 모델을 개발하기 위해서는 목표 변수의 분포가 50:50은 아니더라도, 최소한 패턴을 인식할 수 있는 수준의 비율은 유지하여야 한다. 그러나 현실에서는 그렇지 못한 경우가 많이 있다. 대표적으로 사기 적발(fraud detection)과 같은 문제가 가장 전형적인 예라고 할 수 있다(Fawcett and Provost, 1997). 예를 들어, 신용카드 회사에서 일정기간 전체 카드 사용건수 중 부정하게 다른 사람의 카드를 도용하여 사용한 건수는 매우 적으며, 거의 대다수의 경우는 정상적인 사용 건이다. 그 비율이 심지어는 99.999: 0.001의 비율로, 부정사용 건들이 적을 수 있다. 이런 상황에서 부정사용 건을 적발하기 위한 지도학습 모델을 개발 하고자 할 때, 원시 데이터(source data)에 지도학습 기법을 그대로 적용하면, 거의 분류 패턴이 나타나지 않게 된다. 이런 신용카드 부정사용 사례 이외에도, 불균형한 단백질 구조에서 서열 규칙을 찾아내는 사례(Radivojac *et al.*, 2004), 바다 표면의 위성사진을 통해서, 기름 유출이 발생되는 곳을 찾아내는 사례(Kubat *et al.*, 1998)나 부정한 사기 전화 통화 문제에 관한 사례(Fawcett and Provost, 1996), Text(문장)를 해당 문장 그룹 범주에 정확하게 분류하는 사례(Lewis and Ringuette, 1994) 등 다양한 상황에서 데이터 불균형 문제는 발생하게 된다. 이런 여러 상황 중, 본 논문에서는 국내의 한 종합 병원의 건강보험 심사청구 사례에서 나타난 데이터 불균형 문제 해결 과정을 살펴보기로 한다.

1.1 논문의 배경 - 건강 보험 심사 평가

국내 모든 의료기관에서 행하는 의료 행위 중 거의 대다수는 건강보험에 적용을 받게 된다(건강보험관리공단, 2006). 즉 의료비의 일부는 건강보험료로 나오게 되어, 환자는 전체 의료비 중 건강보험료 적용 분을 제외하고, 나머지 차액을 병원에 지불하게 된다. 병원은 전체 의료비 중 보험료 해당 금액을 건강보험관리공단에 청구하여, 그 금액을 받게 되는데, 금액을 받기 전 의료기관(병원)은 의료행위에 대한 내역을 일정 양식 및 데이터 파일화하여, 건강보험심사평가원에서 보험금 청구 심사를 받은 후 심사가 완료된 의료/투약 행위 건에 대해서만, 보험금을 받게 된다. 이런 심사 청구 과정은 불필요한 검사나 진료, 투약 행위 등 부당 의료행위로 인한 환자의 부담을 가중시키는 것을 사전 차단하여, 올바른 의료서비스가 정착하는데 목적이 있다(건강보험심사평가원, 2006). 이 과정에서 특히 대형 종합병원은 건강보험심사평가원에 청구를 하기 전 자체적으로 보험료 청구 대상인 의료행위 전부에 대하여 사전 심의 및 정리를 하는 팀 또는 부서를 운영하여, 사전에 과잉/부당 의료행위를 방지하고, 또한 환자 및 질병의 특성에 따라 건강심사평가원의 기준에 적발될 수 있는 의료행위에 설명 및 특별 자료를 첨부하여, 합법적인 진료비의 청구를 진행하고, 가급적 청구된 보험료가 삭감되지 않도록 하고 있다(최길림, 1995). 대한병원협회에 따르면 전국적으로 이런 보험 심사 청구에만 종사하는 인력은 약 4만 여명으로 알려져 있으며(이수연 외, 2004), 또한 건강보험심사평가원에서도, 전국 병원에서 올라오는 의료보험 청구의 심사업무를 위해 의료 지식이 있는 간호학 출신 전공자들 1000여명이 업무에 참여하고 있다(장익암, 2000).

1.2 본 논문의 목적과 데이터 불균형의 문제

이런 심사 청구 업무에 대한 병원의 입장은 매우 민감하다. 심사 청구가 잘 이루어지면 문제가 없지만, 만약 삭감이 발생하는 경우 2가지 입장에서 병원들은 손실을 입는다. 첫 번째는 수익의 손실이다. 진료나 투약, 각종 검사를 통해서 비용은 이미 지출이 되었으나, 돈을 받을 길이 없기에 고스란히 그 비용 부담은 병원으로 오게 되는 것

이다. 두 번째는 이미지의 손실이다. 삭감 건수가 많다는 것은 그 만큼 해당 병원이 과잉 진료 등 올바른 의료 서비스를 하지 못한다는 것을 의미하기 때문에, 병원에서는 무형적으로 심각한 손실을 입을 수 있다. 따라서 매우 많은 보험료 청구 건이 발생하는 대형 종합병원 같은 곳에서는 어떤 경우에 건강보험심사평가원에서 보험료가 삭감이 되는지 사전에 파악할 수 있는 시스템이 있는 경우, 상당히 많은 관련 업무를 감소시킬 수 있고, 직접적인 수익을 올릴 수도 있으며, 병원의 이미지 향상에 도움을 줄 수 있다(유상진, 박문로, 2005). 본 사례의 대상이 되는 한 종합병원에서는 데이터 마이닝의 지도학습 기법 중 의사결정나무 추론(decision tree induction)을 이용하여 사전에 보험료가 삭감이 될 것 같은 의료 처방 규칙을 찾아내는 프로젝트를 수행하였다. 삭감 규칙을 찾아내는 데이터 마이닝 프로젝트를 통해서 사전에 찾아낸 진료 건이 만약 과잉 진료인 경우에는 해당 의료진에게 경고 및 조치를 지시하고, 의학적으로 불가피한 상황이라면 추가 자료를 청구서에 첨부하여, 최대한 청구된 건이 삭감되는 것을 방지하자는 목적을 가지고 있다. 이를 데이터 마이닝으로 수행함에 있어, 가장 크게 대두된 문제가 바로 데이터의 불균형 문제이다. 즉, 대다수의 의료 행위는 부당/과잉 청구 건이 아니라 의학/약학적으로 정상적인 건이기 때문이다. 본 논문에서는, 사례가 되는 국내의 한 대형 종합병원 신경외과의 CT(Computerized Tomography: 전산화 단층촬영, 이하 CT) 건 중에서 건강보험심사평가원에서 삭감이 될 수 있는 의료 행위 건을 사전 탐지하는 의사결정나무 모델을 개발함에 있어, 데이터 불균형의 문제를 해결한 방법을 사례로 소개하고자 하고, 이를 통해서 다른 유사한 사례에서도 적용할 수 있도록 하고자 한다.

Ⅱ. 관련연구

과거부터 데이터 불균형 문제를 해결하고자 하는 연구들은 많이 있었다. 보편적으로 불균형 데이터를 극복하기 위한 가장 기본적인 방법으로 sampling을 이용한 방법이 있으며, 다른 방법으로는 오분류 비용을 조정하거나 분류의 결정 기준(decision thresholds)을 조정하는 방법이 있다.

먼저, sampling 방법에는, 다수 범주 집단에서 임의적 sampling을 하여, 소수 범주와 균형(balance)을 이루도록 하는 "Under Sampling" 방법과 소수 범주 집단을 반복적으로 복사하여 다수 범주 집단과 균형을 이루게 하는 "Over Sampling" 방법이 있다. Japkowicz(Japkowicz, 2000)는 가상 데이터를 생성하여, 이 2가지 sampling 방법을 비교하여 데이터 불균형을 해결하고자 하는 연구를 수행하였다. 또한 단순히 1개의 sampling 방법이 아닌 다양한 다른 방법을 결합하여, sampling 효과를 높이고자 하는 연구의 하나로 Chawla 등(2002)은 k-NN(k-Nearest Neighbour) 기법을 이용하여, 소수 범주 집단의 데이터를 over sampling하는 방법인 SMOTE(Synthetic Minority Over-Sampling Technique)를 제시하였다. SMOTE는 over sampling을 수행할 때, 소수 범주의 값을 반복하여 추출하는 것이 아니라, 소수 범주들과 이들 이웃들(neighbours) 사이에서 새로운 값을, 보간법을 이용하여 over sampling 하는 방법이다. 즉, 소수 범주를 그대로 반복하는 것이 아니라, 소수 범주들과 가까운 주변의 소수 범주들 사이의 값을 sampling하여, over sampling에서 나타날 수 있는 과적합(over fitting) 문제를 해결한 sampling 방법이다. 또한 Guo and Viktor (2004)의 경우에는 DataBoost-IM이란 알고리즘을 사용하여, 데이터를 생성시키는 새로운 sampling 방법을 이용하여, 데이터 불균형의 문제를 해소하고자 하였다. 그 외에도 Jo and Japkowicz(2004)는 데이터 불균형 문제를 Small disjuncts라는 소수의 오분류 건에 초점을 맞추고, 이를 군집분석을 이용하여, 데이터 불균형 문제를 해결하는 방안을 제시하였으며, 또한 Su 등(2005)은 동질성 지수(homogeneity index)와 비구분율(undistinguisable ratio)이라는 지표를 이용한 KAIG

(Knowledge Acquisition via Information Granulation)라는 방법을 개발하고, 불균형 데이터에서 SVM (Support Vector Machines)과 C4.5를 이용하여, 불균형 데이터 하에서의 효율성에 대한 연구를 수행하였다. 또한 Laurikkala(2001)는 NCL(Neighbourhood CLeaning rule)이라는 것을 고안하였는데, 이는 소수 범주와 유사한 다수 범주를 제거하는 cleaning 작업 후 sampling을 하여 모델을 개발하는 것으로 이상치 및 분류에 문제가 되는 데이터 제거 후 sampling하여, 좀 더 패턴이 뚜렷하게 나타나도록 하였다. 이와 비슷한 개념으로 Hart(1968)는 CNN(Condensed Nearest Neighbor rule)이라는 것을 이용하여, 다수의 범주의 중심에서 떨어진 레코드를 제거 후 sampling을 하여, 데이터 불균형 문제를 해결하고자 하였다. 국내에서는 강필성 등(2004)이 SVM 앙상블 기법을 적용한 over sampling을 통해서 데이터 불균형 문제를 해소하고자 하였다.

두 번째 방법은 오분류 비용을 조정하거나 각종 가중치를 줌으로써, 데이터의 불균형을 해소하고자 하는 방법이다. 이는 원 데이터 구조를 그대로 유지하면서, 오분류에 가중치를 주어, 데이터의 불균형을 해소하는 방법이다. 오분류 비용의 조정은 의사결정나무 추론 기법에서만 사용이 가능하며, 그 외 로지스틱 회귀분석 등의 기법에서는 목적변수에 가중치를 다르게 주어서, 분류가 되는 기준을 변화시켜서 불균형한 데이터 문제를 해결할 수 있다. 먼저 Christianini and Shawe-Taylor(2000)은 SVM과 커널 기반학습 방법에서 오분류 비용의 조정에 따른 불균형 데이터의 해결 방법을 제시하였다. 또한 Huang 등(2004) 은 Biased Minmax Probability Machine이라는 방법을 고안하여, 오분류 기각역의 변화에 따른 불균형 데이터의 해결을 연구하였으며, 김지현과 정종빈(2004)은 각종 오분류 비용을 통한 복원 sampling 방식이나 소수 범주에 가중치 적용을 통해서, 단순한 sampling 방법보다 더 좋은 효율을 내었다고 하였다. 다양한 방법들에 대한 소개 및 개발 이외에도 여러 기법을 비교 연구하며, 상황에 맞는 가장 좋은 기법을 찾고자 하는 연구들이 있었는데, Batista 등(2004)은 다양한 불균형한 데이터들에서 SMOTE를 비롯한 다양한 sampling 방법 등을 이용한 데이터 불균형 해소 방법들의 비교를 통해서, 가장 좋은 방법을 찾고자 하였다. 또한 Huang 등(2005)은 계좌 정보를 중심으로 은행 고객의 신용 위험도 데이터에서 나타난 불균형의 문제에 대하여 여러 데이터 마이닝 기법을 비교하여, 효율적인 모델을 찾고자 하는 연구를 하였다. 그리고 Chawla(2004) 등은 현재까지 기계학습에서 발생하는 데이터 불균형 해결의 주요 연구 결과들과 연구 방향을 정리하는 연구를 수행하기도 하였다.

Ⅲ. 본 논문 제안 방법과 사례의 개요

본 논문에서 소개되는 사례는 국내의 대형 종합병원의 신경외과 CT 검사에 대한 보험료 청구 사례이다. 지도학습 기법을 적용하기 위해, 본 자료를 훈련용(training) 데이터와 테스트용(testing) 데이터로 분리하였다. 데이터의 개요는 <표 1>과 같다. 본 자료는 총 7개월 동안 CT 촬영 결과 데이터로, 훈련용 데이터로는 앞의 5개월 데이터를 이용하였으며, 나머지 2개월 데이터를 테스트용 데이터로 이용하였다. 사례병원의 전체적인 보험료 청구 건에 대한 실질 삭감 비율은 약 1.5% 정도이다. 그러나 다음의 <표 1>에서 보는 바와 같이 신경외과 CT의 경우에는 평균 3~4%로 전체와 비교해서 보험료 삭감이 매우 높은 분야라고 할 수 있다. 이는 CT 분야의 보험료 삭감과 비삭감의 데이터 불균형도 심각하지만, 다른 진료과로 갈수록 더욱 더 데이터 불균형이 심각한 상황이라는 것을 의미한다.

3.1 모델 성능 평가 기준

다음의 <표 1>에서 보는 것과 같이 매우 심각

〈표 1〉 사례 병원에서 분석에 활용된 데이터

훈련용 데이터의 수	삭감	412건 (4%)	10,310건	총 합계 : 14,751건
	비삭감	9,898건 (96%)		
테스트용 데이터의 수	삭감	191건 (3.7%)	4,441건	
	비삭감	4,250건 (95.7%)		

한 불균형 데이터일 때, 개발된 모델의 성능 평가기준을 오분류율 또는 정확도율(= 1-오분류율)을 사용하는 것은 적절한 성능 평가 기준이 되지 못 한다(김지현, 정종빈, 2004). 위의 <표 1>에서 보는 바와 같이 별다른 모델을 만들 것도 없이 "*무조건 삭감이 되지 않는다.*" 라는 규칙만으로도 96%는 맞출 수 있기 때문이다. 따라서 이런 데이터 불균형의 상태에서는 상황에 맞는 다른 성능 평가기준이 필요하다. 본 논문에서는 2가지의 성능 평가 기준을 설정하였다. 첫째는 "소수범주 오분류율"이 최소가 되는 것이고, 둘째는 전체에서 차지하는 "심사업무 축소율"[1]이 최소가 되는 것이다.

소수범주 오분류율이 최소가 된다는 것은 <표 2>를 통해서 보면, $\frac{FP}{FP+TN}$ 가 최소가 되는 것을 의미한다. 이는 개발한 모델에서 소수범주에 대하여, 예측력이 좋아야 한다는 것을 의미한다. 다시 말해서, 다른 데이터 마이닝 예측 모델과 다른 점은 전체적으로 잘 맞추는 것이 중요한 것이 아니라, 특히 비삭감은 고정한 상태에서, 삭감 규칙의 정확도가 특히 더 중요하다는 것을 의미한다. 다음, 심사업무 축소율이 최소가 된다는 것은 <표 2>에서 보면, 전체 건수 중에서 예측된 삭감 건수를 의미하는 $\frac{(FN+TN)}{(TP+FN+FP+TN)}$ 이 최소가 되는 것이다. 즉, 예측 모델에서 삭감으로 예측한 것이 전체 건수와 비교하여 최소가 되는 것이다. 이는 병원 내부 심사자들이 사전 청구 심사를 할 때 심사 건수를 줄여주어야 한다는 것을 의미한다. 즉, 최소의 노력을 통해서, 최대의 효과를 발생시켜야 하는 경제적인 논리가 적용된 평가기준이라고 할 수 있다. 그러나 이 2개의 기준은 한 쪽이 높으면 다른 한 쪽이 낮아지는 상충관계(trade-off)를 가지고 있다. 따

〈표 2〉 보험료 청구 삭감의 오분류 표의 구조

구분		예측	
		비삭감	삭감
실제	비삭감	참(비삭감을 잘 맞춘 경우): TP	거짓(비삭감을 삭감으로 예측): FN (제 1종 오류)
	삭감	거짓(삭감을 비삭감으로 예측): FP (제 2종 오류)	참(삭감을 잘 맞춘 경우): TN

* 전체 Test 데이터의 수(TP+FN+FP+TN) : 4,441건

1) 소수범주오분류율과 심사업무 축소율은 본 논문에서 사용하는 2개 지표의 대표 명칭으로 사용하기로 한다.

라서 본 논문의 사례 병원에서는 이 2개 평가기준을 동시에 만족할 수 있도록 사전에 다음과 같은 2개의 목표를 설정하였다.

① 기본적으로 심사 업무량을 30% 이하로 줄일 것[2)]

② 기본 심사 업무량이 30% 이하로 줄어 있는 경우에서 심사 업무량과 소수범주 오분류율이 동시에 가장 축소되도록 할 것

첫 번째 목표는 심사업무량이 전체적으로 30% 이하로 감소하는 것을 의미한다. 여기서 30%라는 값은 시스템 개발비 및 용역비 등을 비용으로 하고, 심사 업무량의 감소를 통한 비용 절감을 수입으로 계산하여 ROI가 1년 안에 흑자로 될 수 있는 지점을 잡아서 결정을 하였다.[3)] 두 번째 목표는 기본적으로 30% 이하로 업무가 줄어야 하고, 그 중에서 최대한 소수범주 오분류율과 심사업무 축소율이 동시에 줄여야 한다는 의미이며, 이를 수식으로 정리하면 다음과 같다.

두 번째 목표 =

$$\alpha\frac{FP}{TP+TN}+(1-\alpha)\frac{FN+TN}{TP+FN+FP+TN} \quad (\text{단}, \alpha=0.5) \qquad (\text{식 } 1)$$

위의 식에서 $\alpha=0.5$는 심사업무 축소율과 소수범주 오분류율 모두 동등한 가중치를 가진다는 것을 의미한다. 또한 (식 1)은 실험 후 모델을 평가하는 최종 결정 기준이기도 하다. 여기서 상대적인 가중치 결정은 현업 담당자와의 협의를 통해 균등하게 부여되었다.

3.2 본 사례에서 사용한 불균형 해소 방법

관련 연구에서도 살펴보았듯이 의사결정나무 추론에서 불균형 데이터를 처리하는 방법은 몇 가지 방법이 있었다. 이를 크게 2가지로 정리하면, 하나는 sampling을 활용하는 방법이고, 다른 하나는 오분류 비용을 이용한 방법이다. 본 사례에서는 현재 소개되어져 있는 이 2가지 주요한 불균형 해소 방법들을 다양하게 조합하여 사용하고자 하였다. 즉, 주요한 방법들을 조합하는 다양한 경우를 생성하고, 이들을 동일한 테스트용 데이터를 이용하여 평가한 후 가장 효율적인 방법을 선택하고자 하였다. 본 사례에서 사용한 방법들은 다음과 같다.

<표 3>의 1번과 2번 방법의 경우는 일반적인 임의적 under sampling과 over sampling을 이용한 방법이고, 3번의 경우는 Chawla(2002) 등이 제안한 SMOTE 방법에서, kNN이 아닌 K-평균 군집분석을 이용하여, 다수 범주의 거리보다 소수 범주와 가장 유사한 거리를 가지는 건을 소수 범주로 바꾸는 방법을 이용하여(허명회, 2005), over sampling을 하였다.[4)] 또한 4번째 방법은 Laurikkala(2001)가 사용한 NCL 방법과 같이, 다수 범주(본 사례에서는 비삭감 건)에서 소수범주와 유사한 성향을 가지는 이상한 건을 제외하고 under sampling하는 방법을 이용하였다. 이상치 제외 방법으로는 K-평균 군집거리를 이용하여, 다수 범주의 군집 중심보다는 소수 범주의 군집 중심에 가까운 다수 범주 건들을 제외하는 방법을 이용하였다. 이렇게 다양한 방법의 조합을 통해서, 본 사례에 가장 적합한 데이터 불균형 해소 방법을 찾는 것이 본 논문의 목적이라고 할 수 있다. 또한 1개의 방법만을 이용한 것보다, 2개의 방법을 조화롭게 사용하는 것이 더 효율적인지 검정하고자 한다. 본 목적을 달성하기 위하여 첫 번째 과정으로 <표 3>에 포함된 4가지 sampling 방법(under sampling, over sampling,

2) 30%로 고정이 아니라 그 이하로 더 줄일 수 있으면, 업무량이 더 축소되어 더욱 효율적임.

3) 인건비 등이 공개될 수 있어, 해당 병원의 정보 보호차원에서 정확한 수식을 공개하지 못함.

4) 본 사례가 된 프로젝트에서는 사례가 된 병원에서 향후 분석자들이 손쉽게 데이터 마이닝 패키지를 이용해야 하므로, 패키지가 제공하는 알고리즘을 이용하기 위하여 일부 변형된 방법을 사용하였다.

〈표 3〉 본 사례의 적용방법

방법	방법 설명	경우의 수	sampling 방법 약칭	반복수
1	(under sampling을 통한 범주 균형 변화 조건 5가지) × (오분류 비용의 변화 조건 3가지)	15개	RUS (Random Under Sampling)	10회
2	(over sampling을 통한 범주 균형 변화 조건 5가지) × (오분류 비용의 변화 조건 4가지)	20개	ROS (Random Over Sampling)	10회
3	(k-평균의 군집거리와 SMOTE 통한 범주균형 변화 조건 5가지) × (오분류 비용의 변화 조건 4가지)	20개	SMOTE	10회
4	(k-평균의 군집거리를 통해서, 다수 범주 중 이상치를 제외한 under sampling을 이용한 범주 균형 변화 조건 5가지) × (오분류 비용의 변화 조건 3가지)	15개	USWO (Under Sampling Without Outlier)	10회

SMOTE를 이용한 over sampling, 이상치를 이용한 under sampling)을 통해서 다양한 비율로 sampling을 수행하였다. 여기서 다양한 비율의 sampling이란, 다수 범주와 소수 범주의 수를 50:50으로 sampling 하는 것뿐 만 아니라, 65:35, 75:25, 80:20, 90:10 등으로 다양하게(5가지 경우로) sampling 하는 것을 의미한다. 두 번째 과정으로, 오분류 비용의 변화 조건은 바로 의사결정나무 추론의 특징 중 하나인 오분류 비용을 조정하여, 오분류가 된 진료 건에 가중치를 주어, 데이터 불균형 문제를 해결하고자 하는 것이다. <표 2>의 FN:FP 비가 1:1이 아닌 다른 값을 주면 오분류 비용의 가중을 주는 것을 의미한다. 본 논문에서의 관심은 삭감을 비삭감으로 예측하는 것(FP)을 최대한 줄이려는 것이므로, FP의 오분류 비용을 변화시켜서 실험한 경우의 수 중에서는 가장 좋은 조건을 찾고자 하였다. under sampling의 경우에는 아무런 가중을 주지 않는 1:1, 1:3 그리고 1:5 등 3개의 오분류 비용 조건을 정의하였고, over sampling의 경우에는 1:1, 1:5, 1:10, 1:20 등 4가지 오분류 비용 조건을 주었다. 조건을 이와 같이 설정한 것은 만약 비용의 비율을 1씩 증감하면, 전 단계 경우의 수와 곱해져 기하급수적으로 실험 수가 증가하기 때문에, 최대 24배 정도의 차이를 보간(interpolation)으로 파악할 수 있도록 적정한 거리를 두어 파악하여 보았다.

세 번째 과정은 이렇게 해서 나온 70개(under sampling 30개, over sampling 40개)의 모델에서 본 논문의 측정 지표인 소수범주의 오분류율과 심사업무 축소율을 산출하는 것이다.

네 번째 과정은 이들 70여 가지의 경우를 10번씩 동일하게 반복하여, 지표 산출의 평균을 구하였다. 반복 실험을 통해서 실험의 재현성 및 신뢰성을 향상하였다.

다섯 번째 과정은 10번의 반복에서 구해진 소수범주 오분율의 평균과, 심사업무 축소율의 평균을 이용하여 산점도를 그리고, 심사업무축소율 30%를 만족시키면서 (식 1)을 최소화하는 조합을 최종 선정하였다.

Ⅳ. 사례 문제의 해결 결과

4.1 실제 데이터의 현황 및 설명

본 사례 데이터 구성은 다음의 <표 4>와 같다. 본 데이터는 사례 병원에서 2004년 1월부터 7월

〈표 4〉 사례 데이터의 구성

변수명	변수설명	비고
ID	환자의 고객번호	
내원일수	환자가 병원에 내원을 한 일수의 합계	
투약일수	환자가 투약을 실질적으로 시작한 일수의 합계	
진료일수	환자를 진료한 날들의 합계	
총진료비	(보험)청구액 +본인부담액	
(보험)청구액	건강보험관리공단에 청구할 진료금액	
본인부담액	보험료를 제외하고 환자가 부담해야 할 진료비 액수	
가산율	해당 진료의 가산율	단위 : %(백분율)
본인부담율	전체 진료비 중 본인이 부담을 해야 하는 비율	단위 : %(백분율) 100-본인부담율 = 보험부담율
연령	환자의 나이	
성별	환자의 성별	
외래 촬영 여부	환자가 다른 병원에서 CT 촬영한 것이 있는지 여부	병원을 이전하는 경우 전 병원에서 가지고 오는 CT
보험의 종류	보험이 건강보험인지, 산재 보험인지, 지역보험인지 등을 구분하는 변수	
명세서당 CT 종류	1개의 명세서 당 CT의 종류 수 합계	
총 CT 청구수	명세서당 CT 종류 수의 합계	환자 1명당 여러 개의 명세서가 존재 가능
입원횟수	환자가 사례 병원에 입원한 횟수	
상병코드	해당 환자의 상병 종류를 표시한 코드	
조영제 사용 여부	CT 촬영에 있어 조영제를 사용했는지 여부	조영제 : CT 촬영에 필요한 검사 시약의 종류
삭감여부	건강보험심사평가원에서 최종 삭감 판정을 받았는지 여부	목표 변수(Yes / No)

까지 7개월 간 신경외과 CT촬영을 하고, 건강보험심사평가원의 심의 평가를 받은 자료이다.

본 연구에서는 SPSS[5]사의 데이터 마이닝 솔루션인 클레멘타인 버전 10.0의 의사결정나무 추론 알고리즘인 C5.0을 이용하였다. 클레멘타인 버전 10.0에서는 총 4가지, C5.0(Quinlan, 1992), C&RT (Brieman *et al.*, 1984), CHAID(Kass, 1980), QUEST (Loh and Shin, 1997)의 의사결정나무 추론 알고리즘을 지원한다. 4개의 알고리즘에 대하여, 모수 변화 없이 실험을 한 경우[6] C5.0을 제외하고는 전혀 의사결정나무가 확장하지 못하였다. 따라서 본 연구에서는 C5.0을 사용하여 실험을 수행하였다.

4.2 실험 결과

먼저, 별도의 sampling없이 오분류 비용만 조정하여 본 결과는 <표 5>와 같다. 오분류 비용의

5) SPSS Inc.(http://www.spss.com)
SPSS Korea(http://www.spss.co.kr)

6) 기본 설정값(default) 상태를 의미함

〈표 5〉 오분류 조정만 한 경우의 분석결과

오분류비용 배수	구분		예측		소수범주 오분류율	심사업무 축소율
			비삭감	삭감		
1	실제값	비삭감	4,241	9	**91.1%**	**0.6%**
		삭감	174	17		
3	실제값	비삭감	4,057	193	**55.5%**	**6.3%**
		삭감	106	85		
5	실제값	비삭감	3,977	273	**50.3%**	**8.3%**
		삭감	96	95		
10	실제값	비삭감	3,943	307	**44.1%**	**9.3%**
		삭감	84	107		
17	실제값	비삭감	3,999	251	**50.3%**	**7.8%**
		삭감	96	95		
24[7]	실제값	비삭감	3,939	311	**49.2%**	**9.2%**
		삭감	94	97		

조정은 <표 2>에서 초기 오분류 비용의 조정 전 FN과 FP의 비율이 1:1이 되는데, 여기서 FP의 오분류 비용을 증가시켜보는 것을 의미한다. 이는 FP의 비용을 증가시켜, 소수 범주의 오분류율을 낮추고자 하는 의미가 있다. 이 방법은 분포 불균형 문제를 해결하는 기초적인 방법도 되기 때문에(허명회, 이용구, 2003) 후에 본 사례에서 분석한 결과와 비교를 위해서 우선 제시되었다.

위의 <표 5>에서 소수범주 오분류율은 전체 삭감 건 191건 중 예측을 잘못한 비율로써, 모델의 정확성을 의미하고, 심사업무 축소율은 전체 4,441건을 100%로 두었을 때, 심사 업무 축소율 만큼만 진료 건을 검토하면, 100-소수범주 오분류율(%) 만큼 정확한 삭감 건을 찾아낼 수 있다는 것을 의미한다. 또한 심사업무축소율이라는 지표는 병원에서 관련된 심사 건의 축소를 파악할 수 있는 지표로써, 낮으면 낮을수록 적은 수의 심사를 하는 것을 의미한다. 다시 <표 5>를 보면, 오분류 비용만 조정을 한 경우 소수 범주 오분류율에서, 오분류 비용을 3으로 준 이후부터는 소수범주의 오분류율이 4~50% 대를 계속 유지하고 있는 것을 알 수 있다. 그리고 심사건수의 축소를 나타내는 심사업무축소율 역시 오분류 비용에 어떤 값을 주어도 큰 변화가 없는 것으로 나타났다. 이는 단순히 오분류 비용만 조정을 하는 경우에는 어떠한 오분류 비용을 선택하여도, 유사한 결과를 보이는 것을 알 수 있다.

다음의 <표 6>부터 <표 9>까지는 sampling의 조건 변화와 오분류 비용 조건 변화에 따른 본 논문의 성능 평가 기준인 소수범주 오분류율과 심사업무 축소율을 나타낸 것이다. 다음의 <표 6>은 RUS(Random Under Sampling) 방법에 오분류 비용을 변화시켜서 나온 결과이다.

<표 6>에서 오분류 비용을 1로 한 경우는, 오분류 비용을 적용하지 않고, 단순히 under sampling만한 것을 의미한다. 다음의 <표 7>은 ROS (Random Over Sampling) 방법에 오분류 비용을

7) 24를 최대로 한 것은 삭감과 비삭감의 비율이 24배가 나타나기 때문이다(허명회, 이용구, 2003).

〈표 6〉 RUS X 오분류비용 방법의 실험 결과(10회 반복 수행)

sampling 방법	오분류 비용	sampling 비율	소수범주오분류율				심사업무축소율			
			평균	표준 편차	최소값	최대값	평균	표준 편차	최소값	최대값
RUS	1	50:50	**15.5%**	2.1	11.5%	18.8%	**20.9%**	2.8	17.6%	25.4%
	1	65:35	**20.0%**	3.9	15.2%	27.2%	**16.5%**	1.9	13.4%	20.2%
	1	75:25	**25.5%**	5.5	20.4%	35.1%	**14.4%**	2.1	10.5%	17.1%
	1	80:20	**31.1%**	7.1	23.0%	48.2%	**12.0%**	1.9	7.7%	14.3%
	1	90:10	**70.6%**	8.8	60.2%	83.2%	**3.5%**	1.4	1.4%	5.4%
	3	50:50	**10.6%**	1.9	7.3%	13.6%	**36.7%**	7.4	28.1%	53.5%
	3	65:35	**14.9%**	2.1	12.0%	18.8%	**22.9%**	3.6	15.2%	27.2%
	3	75:25	**16.0%**	1.9	13.1%	19.4%	**20.7%**	1.9	16.7%	23.2%
	3	80:20	**18.8%**	3.3	13.6%	22.0%	**18.3%**	2.0	15.2%	21.2%
	3	90:10	**31.2%**	4.9	23.6%	40.3%	**12.9%**	1.2	11.3%	14.8%
	5	50:50	**4.7%**	1.8	2.6%	7.3%	**56.5%**	8.0	46.3%	65.6%
	5	65:35	**12.9%**	2.3	8.4%	16.2%	**29.8%**	6.1	22.5%	42.5%
	5	75:25	**14.4%**	1.0	13.1%	16.2%	**23.6%**	1.9	20.1%	27.0%
	5	80:20	**15.2%**	1.8	12.0%	17.8%	**22.5%**	2.1	19.4%	26.3%
	5	90:10	**23.7%**	3.6	20.4%	31.4%	**15.8%**	0.8	14.8%	17.3%

변화시켜 나온 결과이다.

<표 7>의 결과에서 오분류 비용이 1인 경우 또한 over sampling만 수행한 결과와 동일하다. 다음의 <표 8>은 SMOTE 방법을 응용한 over sampling을 이용하여 나온 결과이다.

<표 8>을 보면 오분류 비용이 커질수록 소수범주 오분류율이 상대적으로 낮은 것을 알 수 있다. 그러나 이 때 심사업무 축소율 지표는 한 경우에 대하여 30%인 심사업무량 감소 기준을 초과하였다. 다음의 <표 9>는 이상치를 제외하고 under sampling을 한 결과이다.

<표 6>부터 <표 9>에서 나타난 70가지 경우를 보면, 심사업무 축소율 30%를 만족시키지 못하는 경우는 4개의 경우에서 나타났다. 이들을 제외한 나머지 경우들에 대하여 심사의 오분류율도 최소화하고, 동시에 심사 업무량도 최대한 줄일 수 있도록 (식 1)의 값이 적은 순서로 나열한 것이 <표 10>이다.

다음의 <표 10>에서 보면 USWO 방법에 80:20의 sampling 비율 그리고 오분류 비용이 5인 경우에 가장 좋은 모델이 생성되는 것으로 나타났다. 또한 전체적으로 USWO 모델이 상위권에 있어, 이상치를 제거한 후 under sampling을 하는 방법이 전반적으로 효율적인 것을 알 수 있다. 따라서 현재 실험한 70가지 방법 중에서는 다음의 <표 10>의 1순위 방법을 적용하는 것이 가장 효율적이라고 할 수 있다. 이 모델을 실제 업무에 적용한 경우 테스트 데이터 기준으로 볼 때, 전체 4,441건 중 901건만을 사전 검사하여, 전체 191건의 삭감 건 중에서 160건 정도를 발견할

〈표 7〉 ROS X 오분류비용 방법의 실험 결과(10회 반복 수행)

sampling 방법	오분류 비용	sampling 비율	소수범주오분류율				심사업무축소율			
			평균	표준 편차	최소값	최대값	평균	표준 편차	최소값	최대값
ROS	1	50:50	**56.7%**	1.3	56.0%	60.2%	**6.7%**	0.2	6.3%	4.9%
	1	65:35	**55.5%**	0.0	55.4%	55.5%	**6.9%**	0.0	6.9%	6.4%
	1	75:25	**57.1%**	3.1	53.9%	62.8%	**6.6%**	0.5	5.7%	7.1%
	1	80:20	**58.6%**	0.0	58.6%	58.7%	**6.4%**	0.0	6.4%	6.9%
	1	90:10	**65.7%**	3.2	61.3%	71.2%	**4.3%**	0.4	3.5%	6.9%
	5	50:50	**27.7%**	0.8	27.2%	29.8%	**13.0%**	0.2	12.5%	13.1%
	5	65:35	**33.0%**	0.0	33.0%	33.1%	**11.4%**	0.0	11.4%	11.4%
	5	75:25	**41.1%**	2.6	37.7%	46.1%	**10.2%**	0.7	9.1%	11.0%
	5	80:20	**51.8%**	0.0	51.7%	51.8%	**8.3%**	0.0	8.3%	8.4%
	5	90:10	**51.9%**	6.0	42.4%	61.3%	**8.1%**	1.0	6.8%	9.7%
	10	50:50	**20.4%**	0.0	20.4%	20.4%	**18.7%**	0.2	18.2%	19.1%
	10	65:35	**24.6%**	0.1	24.6%	24.7%	**15.0%**	0.0	15.0%	15.0%
	10	75:25	**26.5%**	1.4	24.6%	28.8%	**13.4%**	0.8	12.3%	14.3%
	10	80:20	**31.9%**	0.0	31.9%	31.9%	**12.5%**	0.0	12.4%	12.5%
	10	90:10	**44.3%**	6.7	34.0%	57.1%	**9.5%**	1.1	7.6%	11.1%
	20	50:50	**15.9%**	0.2	15.7%	16.2%	**23.8%**	0.4	22.8%	24.0%
	20	65:35	**14.7%**	0.0	14.7%	14.7%	**22.3%**	0.0	22.3%	22.3%
	20	75:25	**18.9%**	1.9	16.8%	21.5%	**19.7%**	1.6	17.3%	21.6%
	20	80:20	**25.7%**	0.0	25.7%	25.7%	**15.2%**	0.0	15.2%	15.2%
	20	90:10	**33.2%**	5.4	25.7%	40.3%	**12.5%**	1.5	10.3%	15.2%

〈표 8〉 SMOTE X 오분류비용 방법의 실험 결과(10회 반복 수행)

sampling 방법	오분류 비용	sampling 비율	소수범주오분류율				심사업무축소율			
			평균	표준 편차	최소값	최대값	평균	표준 편차	최소값	최대값
SMOTE	1	50:50	**34.8%**	1.5	32.5%	37.7%	**14.8%**	0.2	14.4%	15.2%
	1	65:35	**32.7%**	0.7	31.9%	34.0%	**14.4%**	0.2	14.1%	14.8%
	1	75:25	**32.5%**	2.0	27.7%	34.0%	**13.7%**	0.5	13.1%	14.6%
	1	80:20	**34.8%**	1.8	33.0%	38.2%	**13.1%**	0.2	12.6%	13.4%
	1	90:10	**42.5%**	1.3	40.3%	44.0%	**11.5%**	0.4	10.9%	12.1%
	5	50:50	**23.0%**	1.1	21.5%	24.6%	**17.9%**	0.5	17.4%	18.9%
	5	65:35	**25.6%**	1.5	23.0%	27.7%	**17.1%**	0.4	16.4%	17.6%
	5	75:25	**27.9%**	2.0	25.1%	30.4%	**16.9%**	0.7	16.1%	18.0%
	5	80:20	**27.8%**	1.9	25.7%	30.9%	**16.5%**	0.4	15.8%	17.0%
	5	90:10	**30.7%**	1.9	27.7%	33.5%	**14.9%**	0.5	14.1%	15.7%
	10	50:50	**18.4%**	2.9	13.1%	20.4%	**21.0%**	1.2	19.7%	23.1%
	10	65:35	**19.7%**	1.3	17.8%	22.0%	**19.1%**	0.8	18.1%	21.1%
	10	75:25	**22.2%**	1.6	18.8%	24.6%	**18.2%**	0.7	17.2%	19.1%
	10	80:20	**23.6%**	2.8	17.8%	28.8%	**18.1%**	1.0	16.7%	20.0%
	10	90:10	**28.7%**	2.1	26.2%	32.5%	**15.9%**	0.4	15.4%	16.4%
	20	50:50	**12.3%**	0.7	11.5%	14.1%	**30.2%**	1.0	28.7%	32.3%
	20	65:35	**14.1%**	2.0	11.5%	18.3%	**24.8%**	0.9	23.3%	26.0%
	20	75:25	**16.1%**	2.6	13.1%	20.4%	**22.1%**	1.1	19.8%	23.5%
	20	80:20	**17.3%**	2.5	13.1%	20.4%	**21.0%**	1.3	19.1%	22.8%
	20	90:10	**24.3%**	2.9	20.4%	29.3%	**17.9%**	1.0	17.0%	20.1%

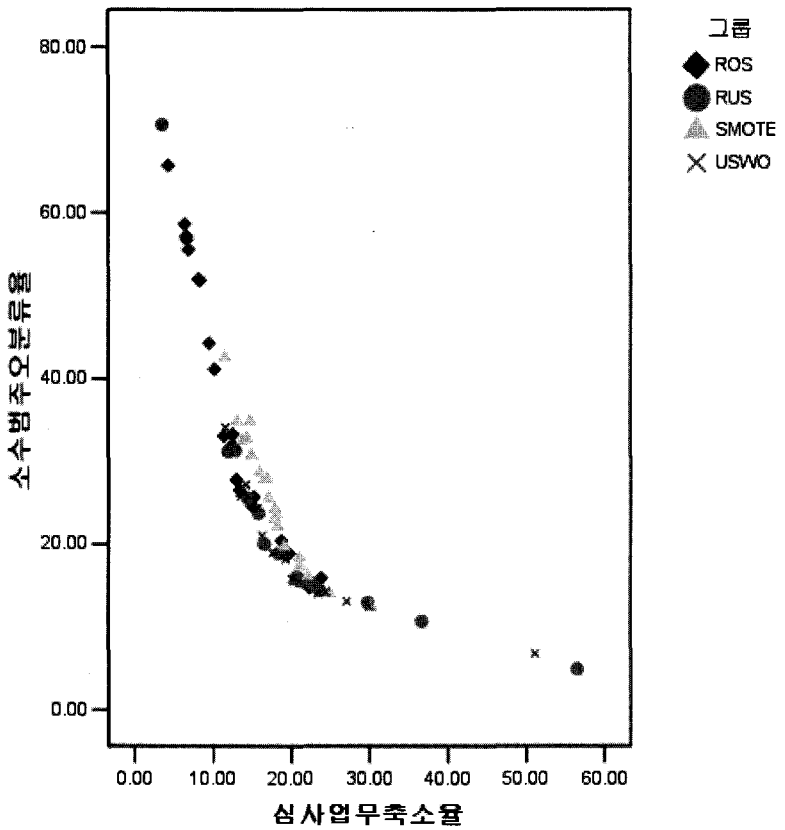
〈표 9〉 USWO X 오분류비용 방법의 실험 결과(10회 반복 수행)

sampling 방법	오분류 비용	sampling 비율	소수범주오분류율				심사업무축소율			
			평균	표준 편차	최소값	최대값	평균	표준 편차	최소값	최대값
USWO	1	50:50	**15.7%**	1.3	13.1%	17.3%	**20.1%**	1.3	17.6%	22.0%
	1	65:35	**21.1%**	3.4	17.3%	26.2%	**16.3%**	2.2	13.7%	19.1%
	1	75:25	**25.8%**	3.0	21.5%	29.8%	**13.6%**	0.8	12.3%	14.6%
	1	80:20	**26.2%**	2.8	20.9%	29.8%	**13.4%**	0.7	12.6%	14.9%
	1	90:10	**34.0%**	3.2	29.3%	40.8%	**11.6%**	0.5	10.4%	12.5%
	3	50:50	**12.5%**	2.1	9.9%	15.2%	**29.9%**	6.2	20.9%	39.2%
	3	65:35	**14.0%**	0.9	12.0%	15.2%	**23.4%**	1.5	21.0%	25.7%
	3	75:25	**18.2%**	5.1	13.6%	29.3%	**19.3%**	2.9	12.3%	22.6%
	3	80:20	**18.9%**	1.7	16.8%	22.0%	**17.6%**	1.6	14.2%	19.5%
	3	90:10	**27.2%**	3.3	22.0%	34.0%	**14.2%**	1.2	11.3%	15.2%
	5	50:50	**6.7%**	3.8	2.6%	13.6%	**51.1%**	11.2	35.4%	67.9%
	5	65:35	**13.0%**	1.5	9.4%	15.7%	**27.1%**	3.3	22.2%	33.1%
	5	75:25	**14.1%**	1.1	12.6%	16.2%	**24.3%**	2.2	20.0%	26.5%
	5	80:20	**15.5%**	2.4	12.6%	20.4%	**20.3%**	1.6	17.5%	22.7%
	5	90:10	**24.4%**	3.0	20.4%	31.4%	**15.7%**	1.0	13.6%	17.0%

수 있다는 것을 보여주고 있다.

아래의 <그림 1>은 4가지 방법에 따라 소수범주 오분류율을 Y축으로 하고, 심사업무 축소율을 X축으로 하는 산점도이다. 좌표 (0,0)에 가까울수록 본 논문에서 정한 기준에 적합한 방법이라고 할 수 있다. <그림 1>을 보면, USWO 방법과 RUS 방법들이 좌표 (0,0)에 가까이 있으며, 특히 SMOTE 방법의 경우 산포 정도가 다른 방법과 비교하면, 매우 좁은 것을 알 수 있다. 다음은 방법별로 가장 안정성이 높은 방법을 찾기 위하여, 각 방법에 따라 2개의 지표에 차이가 있는지 1-way ANOVA를 이용하여 검정을 하였다. 1-way ANOVA 분석 후 사후분석은 Duncan 검정을 이용하였으며, 수행한 결과가 다음의 <표 11>과 같다.

<표 11>을 보면 먼저 4개의 방법에 따른 2개의 지표가 차이가 있는지에 대하여, 1-way ANOVA를 이용하여, 검정한 결과 유의수준 5%에서 유의한 것으로 나타났다. 자세히 살펴보면, 이상치를 제거한 후 under sampling을 한 경우가 소수범주 오분류율이 가장 좋은 것으로 나타났고, 다음으로 일반적인 under sampling과 SMOTE를 응용한 over sampling 방법이 좋은 것으로 나타났다. 이들 2개의 방법은 Duncan 사후 검정 결과 차이가 없는 것으로 나타났다. 이는 sampling을 임의적으로 하는 것보다는 sampling에 조정 및 변화를 주는 경우 더욱 좋은 결과를 도출한다는 선행연구 (Batista 등, 2004; Chawla 등, 2002; Laurikkala,



〈그림 1〉 소수범주오분류율과 심사업무 축소율의 산점도

〈표 10〉 모델의 선택 순위

순위	모델			소수범주 오분류율	심사업무 축소율	(식 1)
	방법	sampling 비율	오분류비용			
1	USWO	80:20	5	15.45%	20.28%	**17.86**
2	USWO	50:50	1	15.72%	20.10%	**17.91**
3	RUS	50:50	1	15.55%	20.95%	**18.25**
4	RUS	65:35	1	20.00%	16.54%	**18.27**
5	USWO	80:20	3	18.91%	17.64%	**18.28**
6	RUS	75:25	3	15.98%	20.68%	**18.33**
7	ROS	65:35	20	14.70%	22.30%	**18.50**
8	RUS	80:20	3	18.84%	18.33%	**18.59**
9	USWO	65:35	1	21.06%	16.26%	**18.66**
10	USWO	65:35	3	13.97%	23.39%	**18.68**
11	USWO	75:25	3	18.16%	19.27%	**18.72**
12	RUS	80:20	5	15.22%	22.52%	**18.87**
13	RUS	65:35	3	14.97%	22.85%	**18.91**
14	RUS	75:25	5	14.39%	23.58%	**18.99**
15	SMOTE	75:25	20	16.12%	22.06%	**19.09**
16	SMOTE	80:20	20	17.28%	20.99%	**19.14**
17	USWO	75:25	5	14.14%	24.34%	**19.24**
18	ROS	75:25	20	18.86%	16.69%	**19.27**
19	SMOTE	65:35	10	19.72%	19.07%	**19.40**
20	SMOTE	65:35	20	14.07%	24.84%	**19.45**
			...			
65	ROS	90:10	1	65.66%	4.31%	**34.99**
66	RUS	90:10	1	70.60%	3.5%	**37.05**

〈표 11〉 방법별 평균의 차이

방법 설명	소수범주 오분류율의 평균	심사업무 축소율의 평균	소수범주 오분류율 F-검정(p)	심사업무 축소율 F-검정(p)
RUS	**24.06%**	**17.59%**	64.307 (0.000)	56.817 (0.000)
ROS	**37.76%**	**12.22%**		
SMOTE	**25.73%**	**17.67%**		
USWO	**20.58%**	**18.18%**		

소수범주오분류율 Duncan 검정 결과	**USWO < RUS=SMOTE < ROS**
심사업무 축소율 Duncan 검정 결과	**ROS < RUS=SMOTE=USWO**

2001; Japkowicz, 2000) 결과와도 부합한다고 할 수 있다. 심사업무 축소율 지표에서는 단순한 over sampling이 가장 업무를 축소시켜주는 것으로 나타났으나, 나머지 3가지 방법 역시 30% 업무 축소에는 거의 다 적합한 것으로 나타났다. 본 논문의 또 다른 목적 중 하나는 위의 <표 6>~<표 9>에서 나온 70가지의 조건 중에서 개별적으로 가장 효율적인 모델을 찾는 것 이외에 과연 다양한 sampling 방법과 오분류 비용을 동시에 조정하는 방법이 일반적으로 단일한 1개의 방법(sampling 또는 오분류 비용의 조정)으로 수행한 것보다 더 효율적인지 확인을 해 보는 것이다. 즉, 일반적으로 불균형 데이터의 해결을 위해서는 sampling이나 오분류 비용 방법 중 1개만을 이용하는데, 이들 2가지 방법을 적절히 이용하여, 1개의 방법만을 사용하는 것보다 더 좋은 효과를 나타내는지 확인하고자 하는 것이다. 그 결과는 <표 12>와 같다.

먼저 데이터가 15~20개로 많지 않기 때문에 비모수 검정 방법 중 2개의 독립된 집단을 검정하는 Mann-Whitney 방법을 이용하였다. 다음의 <표 12>의 결과를 보면, 전부 유의수준 5%이내에서 sampling과 오분류 비용을 동시에 조정하는 방법이 소수범주 오분류율을 낮춘다고 나타났다. 또한 심사업무 축소율 측면에서 보면, 2가지 방법을 혼합하여 사용한 경우가 더욱 많은 심사건을 수행하는 것으로 나타났으나, 전부 본 사례 병원에서 제시하는 30%의 기준보다는 낮은 것으로 나타났다. 이 결과를 통해서, 단순히 sampling이나 오분류 비용만 조정하는 것보다는 이 2가지를 같이 병행하여 모델을 개발하는 것이 본 사례에서는 더욱 효과적이라는 것을 알 수 있다.

V. 결 론

본 논문에서는 데이터 마이닝의 지도학습 기법인 의사결정나무 추론을 이용하여, 병원의 상시 업무인 보험료 심사 업무를 줄이고, 효율을 높일 수 있는 방법을 제시하고 있다. 특히 그 중에서도, 데이터의 불균형으로 의사결정나무 추론을 제대로 활용 못하는 상황에서, sampling과 오분류 비용 조정과 같은 불균형 해소 기법의 조합을 통하여, 가장 효율적이고, 안정성 있는 모델을 찾아내는 방법을 제시하고자 하였다. 본 논문에서 제시한 불균형 해소 방법을 다른 병원

〈표 12〉 sampling과 오분류 비용조정을 복합한 방법과 단독으로만 사용한 방법의 비교

비교 구분	소수범주 오분류율의 평균	심사업무 축소율의 평균	소수범주 오분류율 검정결과 :p 값	심사업무 축소율 검정결과 :p값
RUS(오분류 비용=1) vs. (RUS X 오분류 비용 조정)	32.56% vs. 16.27%	13.47% vs. 25.94%	0.037	0.027
오분류 조정(sampling 없음) vs. (RUS X 오분류 비용 조정)	56.75% vs. 16.27%	6.92% vs. 25.94%	0.001	0.001
ROS(오분류 비용=1) vs. (ROS X 오분류 비용 조정)	58.72% vs. 30.77%	6.18% vs. 14.23%	0.001	0.001
오분류 조정(sampling 없음) vs. (ROS X 오분류 비용 조정)	56.75% vs. 30.77%	6.92% vs. 14.23%	0.005	0.002
SMOTE(오분류 비용=1) vs. (SMOTE X 오분류 비용 조정)	35.47% vs. 22.11%	13.49% vs. 19.42%	0.001	0.001
오분류 조정(sampling 없음) vs. (SMOTE X 오분류 비용 조정)	56.75% vs. 22.11%	6.92% vs. 19.42%	0.000	0.000
USWO(오분류 비용=1) vs. USWO X 오분류 비용 조정	24.57% vs. 16.44%	14.97% vs. 24.27%	0.037	0.020
오분류 조정(sampling 없음) vs. USWO X 오분류 비용 조정	56.75% vs. 16.44%	6.92% vs. 24.27%	0.001	0.001

이나, 신경외과 CT 이외에 다른 진료과에서도 활용할 수 있을 것으로 판단된다. 또한 의료 분야 이외에 신용카드 사기 적발 및 보험 사기 적발 등의 경우에서도 본 방법을 적용하여, 효과적인 모델을 생성할 수도 있을 것이다. 본 사례의 병원에서는 데이터 마이닝을 이용한 보험료 청구삭감 판정 시스템 개발 프로젝트를 본격적으로 추진하기에 앞서, 효과성 측정을 목적으로, 신경외과 CT건에 한하여 pilot 프로젝트로 수행하였고, 본 논문에서 제시한 것과 같은 성과를 얻었다. 이를 전체적으로 적용하기 위해서는 시스템화 시키는 것이 필요하여, 현재 이를 추진하고 있다. 추후 연구과제로는 본 논문에서 제시한 방법이외 다양한 데이터 불균형 해소 방법을 적용하여, 보험료 청구 심사 분석 업무에 적용하는 것과 병원의 다양한 진료과에서 적용하여, 해당 분야에 가장 효율적인 모델을 찾아내는 연구가 필요하다. 더 나아가 보험료 청구 삭감 판정 이외 여러 산업에서 문제가 되고 있는 각종 데이터 불균형 문제에 대한 효율적인 알고리즘 개발과 개발된 알고리즘의 비교 연구가 지속적으로 필요하다.

참 고 문 헌

강필성, 이형주, 조성준, "데이터 불균형 문제에서의 SVM 앙상블 기법의 적용", 한국정보과학회 가을 학술발표논문집, 제31권, 제2

호, 2005, pp. 706-708.

김지현, 정종빈, "계급 불균형 자료의 분류 훈련표본 구성방법에 따른 효과", 응용통계연구, 제17권, 제3호, 2004, pp. 445-457.

오장민, 장병탁, "불균형 데이터의 효과적 학습을 위한 커널 퍼셉트론 부스팅 기법", 한국정보과학회 춘계학술발표논문집(B), 2001, pp. 304-306.

유상진, 박문로, "데이터 마이닝 기법을 활용한 의료보험 진료비 청구 삭감분석 시스템 개발 및 구현에 관한 연구", Information Systems Review, Vol.7, No.1, 2005, pp. 275-295.

이수연, 하호욱, 손태용, "의료기관과 심사기관의 심사업무인식도 비교연구", 병원경영학회지, 제9권 제3호, 2004, pp. 71-97.

장익암, "보험심사 간호사의 업무 스트레스와 대응방법 조사연구", 한양대학교 대학원 간호학과 석사학위 논문, 2000.

최길림, "의료보험입원진료비 청구누락방지를 위한 병원 자체심사에 관한 연구", 인제대학교 보건대학원 석사논문, 1995.

허명회, 이용구, 데이터마이닝 모델링과 사례, SPSS 아카데미, 2003.

허명회, "K-means Clustering을 활용한 분류예측", 제 10회 SPSS 사용자 사례 발표회, 2005.

Batista G., Pati, R.C., and Monard, M.C. "A Study of The Behavior of Several Methods for Balancing Machine Learning Training Data", *SIGKDD Exploring*, Vol.6, No.1, 2004, pp. 20-29.

Brieman, L., J.H. Friedman, R.A. Olshen and C. J. Stone, Classification and Regression Trees. Wadsworth, Belmont, 1984.

Chawla, N.V, Kevin W. Boywer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique", *Journal of Artificial Intelligence Research*, Vol.16, 2002, pp. 231-357.

Chawla, N. V., Nathalie Japkowicz, and Aleksander Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets", *SIGKDD Exploring*, Vol.6, No.1, 2004, pp. 1-6.

Cristianini, N., and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge: Cambridge University Press, 2000.

Fawcett, T. and F. Provost. "Combining Data Mining and Machine Learning for Effective User Profile", *In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, OR. AAAI. 1996, pp. 8-13.

Fawcett, T. and F. Provost, "Adaptive Fraud Detection", *Data Mining and Knowledge Discovery*, Vol.1, 1997, pp. 291-316.

Guo, H., and H. L. Viktor, "Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach", *SIGKDD Explorations*, Vol.6, No.1, 2004, pp. 30-39.

Hart, P.E., "The Condensed Nearest Neighbor Rule", *IEEE Transactions on Information Theory*, Vol.14, No.3, 1968, pp. 515-516.

Huang, Kaizhu, Haiqin Yang, Irwin King, and Michael R. Lyu, "Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine", *Proceedings of the '04' IEEE Computer society conference on computer vision and pattern recognition (CVPR'04)*, 2004, pp. 558-563.

Huang, Yueh-Min, Chun-Min Hung, and Hewijin Christine Jiau, "Evaluation of Neural Networks and Data Mining Methods on a Credit Assessment Task for Class Imbalance Problem", *accepted for publication in Nonlinear Analysis : Real World Applications*, 2005.

Japkowicz, Nathalie., "The Class Imbalance

Problem: Significance and Strategies", *In Proceedings of the 2000 International Conference on Artificial Intelligence*, 2000.

Jo, Taeho., and Nathalie Japkowicz, "Class Imbalances Versus Small Disjuncts", *SIGKDD Explorations*, Vol.6, No.1, 2004, pp. 40-49.

Kass, G. "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics*, Vol.29, No.2, 1980, pp. 119-127.

Laurikkala, J., "Improving Identification of Difficult Small Classes by Balancing Class Distribution", *Tech. Rep. A-2001-2*, University of Tampere, 2001.

Lewis, D. and Marc Ringuette, "A Comparison of Two Learning Algorithms for Text Categorization", *In Proceedings of SDAIR-94, 3rd Annual Symposium on DocumentAnalysis and Information Retrieval*, 1994, pp. 81-93.

Loh, W. and Y. Shin Forthcoming: Split Selection Methods for Classification Trees, Statistica Sinica, Taiwan, 1997.

Kubat, M., Robert C. Holte and Stan Matwin, "Machine Learning for The Detection of Oil Spills in Satellite Radar Images", *Machine Learning*, Vol.30, 1998, pp. 195-215.

Quinlan, R., C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, California, 1992.

Radivojac, P., Nitesh V. Chawla, A. Keith Dunker, and Zoran Obradovic, "Classification and Knowledge Discovery in Protein Databases", *Journal of Biomedical Informatics*, Vol.37, 2004, pp. 224-239.

Su, Chao-Ton, Long-Sheng Chen, and Yuehwern Yih, "Knowledge Acquisition through Information Granulation for Imbalanced Data", *Expert Systems with Applications*, Vol.29, 2005, pp. 1-11.

Weiss, G.M., and F. Provost, The Effect of Class Distribution on Classifier Learning. Technical Report, Department of Computer Science, Rutgers University, 2001.

http://www.nhic.or.kr, 건강보험관리공단 홈페이지.

http://www.hira.or.kr, 건강보험심사평가원 홈페이지.

Information Systems Review
Volume 9 Number 1
April 2007

Decision Tree Induction with Imbalanced Data Set: A Case of Health Insurance Bill Audit in a General Hospital

Joon Hur* · Jong Woo Kim**

Abstract

In medical industry, health insurance bill audit is unique and essential process in general hospitals. The health insurance bill audit process is very important because not only for hospital's profit but also hospital's reputation. Particularly, at the large general hospitals many related workers including analysts, nurses, and etc. have engaged in the health insurance bill audit process.

This paper introduces a case of health insurance bill audit for finding reducible health insurance bill cases using decision tree induction techniques at a large general hospital in Korea. When supervised learning methods had been tried to be applied, one of major problems was data imbalance problem in the health insurance bill audit data. In other words, there were many normal (passing) cases and relatively small number of reduction cases in a bill audit dataset. To resolve the problem, in this study, well-known methods for imbalanced data sets including over sampling of rare cases, under sampling of major cases, and adjusting the misclassification cost are combined in several ways to find appropriate decision trees that satisfy required conditions in health insurance bill audit situation.

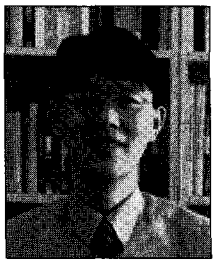
Keywords: Imbalanced Data Sets, Decision Tree Induction, Health Insurance Bill Audit, Over Sampling, Under Sampling, Misclassification Cost

* Consultant, Department of Consulting, SPSS Korea Data Solution Inc.
** Associate Professor, School of Business, Hanyang University, Seoul, Korea

◐ 저 자 소 개 ◑

허 준 (hoh@spss.co.kr)

중앙대학교 응용통계학과에서 경제학사, 중앙대학교 대학원 통계학과에서 경제학 석사를 취득하고, 한양대학교 경영학과 MIS 전공으로 박사과정을 수료하였다. 현대정보기술과 현대/기아 자동차 그룹의 오토에버 닷컴에서 근무하였으며, 현재 SPSS Korea 컨설팅팀에서 수석연구원 겸 컨설팅팀 팀장으로 재직 중이다. 주요 관심 분야는 데이터 마이닝 분야와 각종 수요예측 모델링, BSC를 이용한 성과평가, 통계적 분석 기반의 경영정보시스템 등이다.



김 종 우 (kjw@hanyang.ac.kr)

현재 한양대학교 경영학부 부교수로 재직 중이다. 서울대 수학과에서 이학사, 한국과학기술원 경영과학과에서 공학석사를 취득하고 한국과학기술원 산업경영학과에서 공학박사를 취득하였다. 한국과학기술원 경영정보연구센터 연수연구원, University of Illinois at Urbana-Champaign 방문연구원, 충남대학교 통계학과 부교수로 근무한 경력이 있다. 주요 관심분야는 경영정보시스템, 의사결정지원시스템, e-비즈니스, 추천시스템, 데이터 마이닝 응용, 지능정보시스템, B2B 비즈니스 프로세스 모델링 등이다.

논문접수일 : 2006년 04월 27일 **게재확정일 : 2007년 01월 15일**
1차 수정일 : 2006년 07월 01일