의사결정나무 분석에서 불균형 자료의 분석 연구 : 종합병원의 건강보험료 청구 심사 사례

An Study on Decision Tree Analysis with Imbalanced Data Set : A Case of Health Insurance Bill Audit in General Hospital

허 준*, 김종우**

* SPSS Korea (주)데이터솔루션 (hoh@spss.co.kr)
** 한양대학교 경영대학 경영학부 (kjw@hanyang.ac.kr)

Abstract

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 국내의 대형 종합병원의 사례를 소개하고자 한다. 데이터 마이닝을 적용함에 있어, 주요한 문제점 중의 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 모형을 만들어 내기 위하여, 데이터 불균형 문제의 기본 해법인 Sampling과 오분류 비용의 다양하고 혼합적인 적용을 통하여, 적합한 조건을 가지는 의사결정 나무 모형을 도출하였다.

1. 서론

데이터 마이닝의 지도 학습 기법(Supervised Learning)을 수행하는 과정 중에는 예기치 못한 문제점들이 도출되곤 하는데 그 중 대표적인 것 하나가 바로 목표변수(Target Variable)의 범주 불균형 문제, 다시 말하면 데이터 불균형 문제이다. 불균형 데이터(imbalance data)의 문제란 예를 들어, 목표 변수가 이탈/정상과 같이 이분형으로 되어져 있는 경우 이들의 범주에 속하는 데이터 수의 비율이 현격히 차이가 나는 경우를 의미한다.(오장민, 장병탁, 2001) 일반적으로, 목표변수의 빈도 분포는 동등한 비율로 존재하는 것이 지도 학습 기법을 통해서 패턴을 인식하는데, 가장 좋은 분포가 되어질 것이다. 왜냐하면, 대표적인 지도학습 기법을 적용할 때, 어느 한 목표변수의 범주가 비정상적으로 큰 경우, 지도학습 모형은 전체적인 오분류를 작게 하기 위해서, 다수의 목표변수의 범주로 분류를 많이 하게 되고, 이 경우 소수의 목표변수 범주는 다수의 목표변수 범주로 취급이 되어지기 때문이다.(Weiss and Provost, 2001) 따라서 데이터 마이닝을 이용하여, 올바른 패턴 인식 모형을 개발하기 위해서는 목표변수의 분포가 5:5는 아니더라도, 최소한 패턴을 인식할 수 있는 수준의 비율은 유지하여야 한다. 그러나 실제 상황에서는 이렇지 못한 경우가 많이 있다. 대표적으로 사기 적발(Fraud Detection)과 같은 문제의 경우에 있어서가 가장 보편적이라고 할 수 있다.(Fawcett and Provost, 1997) 예를 들어서, 신용카드 회사에서 일정기간 전체 카드 사용건수 중에서 부정하게 다른 사람의 카드를 도용하여 사용하는 사례는 거의 미비한 수준이고, 거의 모든 사용 건은 정상적인 사용 건이라고 할 수 있다. 그 비율이 심지어는 99.999 : 0.001의 수준으로 비교분석을 할 수 있는 부정사용 건이 거의 발생하지 않을 수 있다. 이런 경우 부정사용 건을 적발하기 위한 지도학습 모형을 만들고자 할 때, 일반적인 데이터 마이닝 방법을 적용하면, 거의 패턴이 나타나지 않게 된다. 이런 신용카드의 부정사용 건 사례 외에도, 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(문자)를 해당 문자 그룹 범주에 정확하게 분류하는 사례(Lewis and Ringuette, 1994) 등 다양한 상황에서 불균형 데이터 문제는 발생하게 된다. 본 논문에서는 국내의 한 종합 병원의 건강보험 심사청구 사례에서 나타난 불균형 데이터 문제 해결 과정을 살펴보기로 한다.

1.1 논문의 배경-건강보험심사평가

국내의 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 적용분을 제외하고, 나머지 차액을 병원에 지불하게 된다. 병원은 전체 의료비 중 보험료 해당 금액을 건강보험 관리공단에 청구를 하여, 그 금액을 받게 되는데, 금액을 받기 전 의료기관(병원)은 의료행위에 대한 내역을 일정 양식 및 자료 파일화하여, 건강보험심사평가원에서 과다 진료 및 부당 청구 심사를 받은 후 심사가 완료된 의료/투약 행위 건에 대하여, 최종적으로 보험금을 받게 된다. 이런 심사 청구 과정은 불필요한 검사나 진료, 투약 행위 등 부당 의료행위로 인한 환자의 부담을 가중시키는 것을 사전 차단하여, 올바른 의료 서비스가 정착하는데 목적이 있다.(건강보험심사평가원, 2006) 이 과정에서 특히 대형 종합병원은 건강보험심사 평가원에 월 1회 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 청

구를 진행하고, 가급적 청구된 보험료가 삭감되지 않도록 하고 있다.(최길림, 1995) 대한병원협회에 따르면 전국적으로 이런 보험 심사 청구에만 종사하는 인력은 약 4만여명으로 알려져 있다.(이수연 외, 2004) 비단, 의료기관뿐만 아니라 건강보험심사평가원에서도, 전국 병원에서 올라오는 의료보험 청구의 심사업무를 위해 의료지식이 있는 간호학 출신 전공자들 1000여명이 업무에 참여하고 있다.(장익암, 2000)

1.2 본 논문의 목적과 데이터 불균형의 문제

이런 심사 청구 업무에 있어서, 병원의 입장은 매우 민감할 수 있다. 이 심사 청구가 잘 이루어지면 좋겠지만, 만약 삭감이 발생하는 경우 2가지 입장에서 병원들은 손실을 입을 수 있다. 첫 번째는 수익의 손실이다. 진료나 투약, 각종 검사를 통해서 비용은 지출이 되었으나, 돈을 받을 길이 없기에 고스란히 그 비용의 부담은 병원으로 오게 되는 것이다. 두 번째는 이미지의 손실이다. 삭감 건이 많다는 것은 그 만큼 해당 병원이 과잉 진료 등 올바른 의료 서비스를 하지 못한다는 것을 의미하기 때문에, 병원에서는 무형적인 심각한 손실을 입을 수도 있다. 따라서 매우 많은 보험료 청구 건이 발생하는 대형 종합병원과 같은 보험료 청구 대상 병원에서는 어떤 경우에 건강보험심사평가원에서 보험료가 삭감이 되는지 사전에 파악하여 줄 수 있는 시스템이 있는 경우, 상당히 많은 업무적 자원도 감소시킬 수 있고, 직접적인 수익을 올릴 수도 있으며, 병원의 이미지 제고에 도움을 줄 수 있다.(유상진, 박문로, 2005) 본 사례의 대상이 되는 한 종합병원에서는 데이터 마이닝의 지도학습 기법 중 의사결정나무분석을 이용하여 사전에 보험료가 삭감이 되는 의료처방 규칙을 찾아내어, 실질적으로 과잉 진료의 경우에는 해당 의료진에게 경고 및 수정을 지시하고, 의학적으로 불가피한 상황의 경우에는 추가 자료를 청구서에 첨부하여, 최대한 청구된 보험 건이 삭감되는 것을 막고자 하는 목적을 가지고, 데이터 마이닝을 이용한 내부 프로젝트를 수행하였다. 본 과정을 데이터 마이닝으로 수행함에 있어, 가장 크게 대두된 문제가 바로 데이터의 불균형 문제인데, 대다수의 의료 처치행위는 과잉 의료행위 또는 부당 청구 건이 아니라 의학/약학적으로 어쩔 수 없는 행위인 것이 일반적이다. 이는 결과적으로 보험료가 삭감이 될 정도의 의료 행위 건은 정상 의료행위에 비해 매우 적다는 것이고, 그로 인하여, 실제로 의사결정나무 분석을 통해서, 삭감 규칙을 찾아내는 모형을 만들기에는 문제가 많이 발생한다는 것이다.

본 논문에서는 사례가 되는 국내의 한 대형 종합병원 신경외과 CT(Computerized Tomography: 전산화 단층촬영, 이하 CT) 건 중에서 건강보험심사평가원에서 삭감이 될 수 있는 의료 행위 건을 사전 탐지하는 의사결정나무 모형을 개발함에 있어, 데이터 불균형의 문제를 해결한 방법을 사례를 소개하고자 하고, 이를 통해서 다른 유사한 사례에서도 적용할 수 있도록 하고자 한다.

2. 관련연구

데이터의 목적변수가 한 쪽 범주에만 편중되어져 있는 데이터 불균형의 경우를 해결하고자 하는 연구는 많이 이루어졌다. 불균형 데이터를 극복하기 위해서, 기본적인 방법이 몇 가지가 있을 수 있는데, 가장 기본적인 방법으로 Sampling을 이용한 방법이 있으며, 다른 방법으로는 오분류 비용을 조정하거나 분류의 결정 기준(decision Thresholds)을 조정하는 방법이 대표적이라고 할 수 있다.

샘플링(Sampling) 방법에는, 다수 범주 집단에서 임의적 Sampling을 하여, 소수 범주와 균형(balance)을 이루도록 하는 “Under Sampling" 방법과 소수범주 집단을 반복적으로 복사를 하여 다수 범주 집단과 균형을 이루게 하는 "Over Sampling" 방법이 있다. Japkowicz(Japkowicz, 2000)는 가상 데이터를 생성하여, 이 2가지 Sampling 방법을 비교하여 데이터 불균형을 해소하는 연구를 수행하였다. 단순한 Sampling 방법이 아닌 다양한 방법을 결합하여 Sampling의 효과를 높이고자 하는 연구의 하나로 Chawla 등(2002)은 k-NN 기법을 이용한 소수 범주 집단의 데이터를 Over Sampling하는 방법으로 SMOTE(Synthetic Minority Over-Sampling Technique)를 제시하였다. SMOTE(Synthetic Minority Over-sampling Technique)는 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, 소수범주를 그대로 반복하는 것이 아니라, 소수 범주들과 가까운 주변의 소수 범주값 사이의 값을 Sampling하여, Over Sampling에서 나타날 수 있는 과적합(Overfitting) 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 맞추어 이를 군집분석 기반 하의 문제 해결 방법을 통해 데이터 불균형 문제 해결을 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 SVM과 C4.5를 이용하여, 불균형 데이터 하에서의 효율성 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 모형을 개발하는 것으로 이상치 및 문제가 되는 건을 제거 후 더욱 더 분류를 잘하기 위한 목적을 가지고 있다. 이와 비슷한 개념으로 Hart(1968)의 CNN(Condensed Nearest Neighbor Rule)이라는 것을 이용하여, 다수의 범주의 중심에서 떨어진 레코드를 제거한 후 불균형 문제를 해결하고자 하는 방법도 연구되었다. 국내에서도 강필성 등(2004)이 SVM 앙상블 기법을 적용한 Over Sampling을 통해서 데이터 불균형 문제를 해소하고자 하였다.

두 번째 방법은 오분류 비용을 조정하거나 각종 가중치를 줌으로 인해서, 데이터의 불균형을 해소하고자 하는 방법이다. 이는 원 데이터 구조를 그대로 유지하면서, 오분류에 가중치를 두어, 데이터의 불균형을 해소하고자 하는 방법이다. 오분류 비용의 조정은 의사결정나무 분석 기법에서만 사용이 가능한 것이라고 할 수 있고, 그 외 로지스틱 회귀분석 등의 기법에서는 목적변수에 가중치를 달리 주어 분류가 되는 기준을 변화시켜 불균형한 데이터의 문제를 해소하고자 하였다. 또한 Christianini and Shawe-Taylor (Christianini and Shawe-Taylor, 2000)도 오분류 비용의 조정에 따른 불균형 데이터의 성능 향상을 언급하였다. 또한 Huang 등(2004) 은 Biased Minmax Probability Machine이라는 방법을 고안하여, 오분류 기각역의 변화에 따른 불균형 데이터의 해결을 연구하였다. 또한 김지현과 정종빈(2004)이 각종 오분류 비용을 통한 복원 Sampling 방식이나 소수 범주에 가중치 적용을 통해서, 단순한 Sampling 균형보다 더 좋은 효율을 내었다고 보고하였다. 다양한 방법들에 대한 소개 및 개발 이외에도 여러 기법을 비교 연구하며, 상황에 맞는 최적의 모형을 찾고자 하는 연구들이 있었는데, Batista 등(2004)은 다양한 불균형한 데이터들에서 SMOTE를 비롯한 다양한 Sampling을 이용한 데이터 불균형 해소 방법들의 비교를 통해서 최적의 방법을 찾고자 하였다. 또한 Huang 등(2005)은 계좌 정보를 중심으로 한 은행 고객의 신용 위험도의 데이터에서 나타난 불균형의 문제에 대하여 다양한 데이터 마이닝 기법을 비교하여, 최적의 모형을 찾고자 하는 연구를 하였다. 그리고 Chawla(2004) 등은 현재까지 기계학습에서 발생하는 데이터 불균형 해결의 주요 연구 실적 및 방향을 정리하는 연구를 수행하기도 하였다.

3. 본 논문의 제안방법과 사례의 개요

본 논문에서 소개되는 사례는 국내의 대형 종합병원의 신경외과 CT 검사에 대한 보험료 청구 사례이다. 데이터 마이닝을 적용하기 위하여, 본 자료를 훈련용(Training) 자료와 테스트용(Testing) 자료로 분리한 데이터의 개요는 다음의 <표 1>과 같다. 본 자료에서 총 7개월 동안의 CT 촬영 결과 정보로서 이용하여 훈련용 데이터로는 5개월의 데이터를 이용하였으며, 2개월의 데이터를 테스트용 데이터로 이용을 하였다. 사례병원의 전체적인 보험료청구 건에 대한 실질 삭감 비율은 약 1.5%정도이고, 위의 <표 1>에서 보는 바와 같이 신경외과 CT의 경우에는 평균 3~4%로 전체와 비교하여, 매우 높은 분야라고 할 수 있다. 이는 CT 분야의 보험료 삭감과 비삭감의 데이터 불균형도 심각하지만, 다른 분야로 갈수록 더욱 더 삭감과 비삭감의 데이터 불균형이 심각한 상황이라는 것을 알 수 있다.

3.1 모형성능 평가기준과 사전정의

위의 <표 1>에서 보는 것과 같이 매우 심각한 불균형 데이터일 때, 개발된 모형의 성능 평가기준을 오분류율 또는 정확도율(= 1-오분류율)을 사용하는 것은 적절한 성능 평가 기준이 되지 못한다.(김지현, 정종빈, 2004) 위의 <표 1>에서 보는 바와 같이 별다른 모형을 만들 것도 없이 *"무조건 삭감이 되지 않는다."* 라는 규칙으로만 정해도 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 *소수범주의 오분류율*이고, 둘째는 전체에서 차지하는 *심사 확인 건수의 축소 비율*이다.

<표 1> 사례 병원에서 분석에 활용된 데이터

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

<표 2>를 통해서 보면 첫 번째 성능 평가기준은 소수범주의 오분류율 $\frac{FP}{(FP+TN)}$ 가 최소가 되는 것을 의미한다. 이는 개발한 모형에서 소수범주에서 대하여, 예측력이 좋아야 한다는 것을 의미한다. 다시 말해서, 다른 데이터 마이닝 예측 모형과 다른 점은 전체적으로 잘 맞추는 것이 중요한 것이 아니라, 특히 비삭감은 고정한 채 삭감 규칙의 정확도가 특히 더 중요하다는 것을 의미한다. 두 번째 평가지표는 <표 2>에서 전체 건수 중에서 예측된 삭감 건수를 의미하는 $\frac{(FN+TN)}{(TP+FN+FP+TN)}$ 이 최소가 되는 것이다. 즉, 예측 모형에서 삭감으로 예측한 것이 전체 건수와 비교하여 최소가 되는 것이다. 이는 병원 내부의 심사자들이 사전 청구 심사를 할 때 심사 건수를 줄여주어야 한다는 의미가 있다. 즉, 최소의 노력을 통해서, 최대의 효과를 발생시켜야 하는 경제적인 논리가 적용된 성능평가기준이라고 할 수 있다. 그러나 이 2개의 기준은 서로 상충관계(trade-off)를 가지고 있다. 바로 통계학 가설검정 결과의 제 1종 오류와 제 2종 오류인데, 어느 한 쪽을 줄이면, 다른 한 쪽의 오류가 증가하는 문제가 바로 그것이다. 따라서 일반적으로는 가설 검정에서는 제 1종 오류를 고정시키고, 제 2종 오류를 최소화시키게 된다.(이용구, 1992) 따라서 본 사례 병원에서는 사전에 다음과 같은 목표를 설정하였다.

① 기본적으로 심사 업무량을 30% 이하로 줄일 것. (제 1종 오류의 고정)

② 기본 심사 업무량이 30% 이상 줄어 있는 경우에서 최고의 정확도를 가질 것

<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			

즉, 두 번째 성능 평가기준인 기본 심사 업무량이 30% 이하로 줄어야 하고,(예를 들어 10,000건의 심사를 인력에 의지하여 수행하고 있는 경우, 이 건수가 3000건 이하로 심사가 줄어야 하는 것을 의미함.) 그 중에서 최상의 정확도를 가지는 모델을 개발하는 것이 목표이다. 심사 업무량의 목표량인 30% 이하 축소는 시스템 개발 및 도입비와 오차비용을 비용으로, 심사 업무량의 감소를 통한 절감 비용을 수입으로 계산하여 ROI가 1년 안에 흑자로 될 수 있는 지점을 잡아서 결정을 하였다.[1] 또한 이 2가지 지표의 상대적인 가중치 결정의 문제는 병원의 업무특성상 정하기가 어려워 동등한 가중을 주었다. 그러나 본 프로젝트 구성원의 업무적 상식에서, 인력을 감소시키는 것은 각종 인사/노무 업무와 관련된 문제를 야기할 수 있어, 동등한 가중치로 두었을 때 최고의 모형이 궁극적인 목적이지만, 2번째 평가 기준을 무시한 상태에서, 최고의 정확도만을 나타낼 수 있는 모형은 어떤 것인지도 추가적으로 살펴보기로 하였다.

3.2 본 사례에서 사용한 모형 방법론

관련 연구에서도 살펴보았듯이 의사결정 나무분석에서 불균형 데이터를 분석하는 방법은 몇 가지 방법이 있었다. 이를 크게 2가지로 살펴보면, 하나는 Sampling을 통해서, 데이터의 균형을 맞추어 주는 방법과 오분류 비용을 이용하여, 일정한 가중치를 주는 방법이라고 할 수 있다. 대표적인 2가지 방법 이외에도 2장에서 제시된 것과 같이 다양한 종류의 Sampling 방법들이 제시되었다. 본 사례에서는 현재 소개되어져 있는 주요한 불균형 해소 방법들을 조합하여 사용하고자 하였다. 즉, 주요한 방법들을 조합하기 위한 다양한 경우의 수를 생성하고, 이들을 동일한 테스트 데이터를 이용하여, 실험한 다음 가장 효율적인 방법을 선택하고자 하였다. 본 사례에서 사용한 방법의 종류는 다음과 같다.

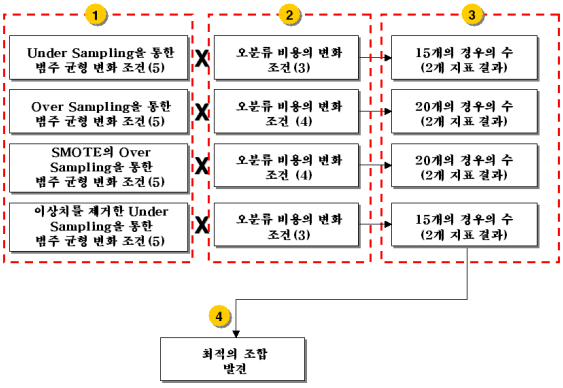
<표 3> 본 사례의 적용방법

방법	방법 설명	경우의 수
1	(Under Sampling을 통한 범주 균형 변화 조건 5가지) X (오분류 비용의 변화 조건 3가지)	15개의 경우의 수
2	(Over Sampling을 통한 범주 균형 변화 조건 5가지) X (오분류 비용의 변화 조건 4가지)	20개의 경우의 수
3	k-평균의 군집거리와 SMOTE 통한 범주균형 변화 조건 5가지) X (오분류 비용의 변화 조건 4가지)	20개의 경우의 수

1) 인건비 등이 공개될 수 있어, 해당 병원의 정보 보호 차원에서 정확한 수식을 공개하지 못함.

4	k-평균의 군집거리를 통해서, 다수 범주 중 이상치를 제외한 Under Sampling을 이용한 범주 균형 변화 조건 5가지) X (오분류 비용의 변화 조건 ·3가지)	15개의 경우의 수

<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 군집 분석을 이용하여, 다수 범주의 거리보다 소수 범주와 가장 유사한 거리를 가지는 건을 소수 범주로 하는 방법을 이용하여(허명회, 2005), Over Sampling을 하였다.[2] 또한 4번의 방법은 Laurikkala (2001)가 사용한 NCL방법과 같이, 다수 범주(본 사례에서는 비삭감 건)에서 소수범주와 유사한 성향을 가지는 이상한 건을 제외하고 Under Sampling하는 방법을 이용하였다. 제외의 방법으로는 K-평균 군집거리를 이용하여, 다수 범주의 군집 중심보다는 소수 범주의 군집 중심에 가까운 다수 범주 건들을 제외하는 방법을 이용하였다. 이렇게 다양한 방법의 조합을 이용하여, 가장 본 사례에 적합한 모델링 방법을 찾는 것이 본 논문의 목적이라고 할 수 있다. 본 사례에서 사용한 분석 프로세스는 <그림 1>과 같다.



<그림 1> 분석 프로세스

<그림 1>의 첫 단계에서 4가지 방법(Under Sampling, Over Sampling, SMOTE를 이용한 Over Sampling, 이상치를 이용한 Under Sampling)을 통한 범주 균형 조건이란 것은, 다수 범주와 소수 범주의 수를 50:50의 균형을 맞추는 조건뿐만 아니라, 66:34, 75:25, 80:20, 90:10 등 변화를 다양하게 하여 각각 5가지 경우를 고려하였다. 두 번째 단계인 오분류 비용의 변화 조건은 바로 의사결정나무 분석의 장점 중 하나인 오분류 비용을 조정하여, 오분류가 된 진료 건에 가중치를 주어서, 데이터 불균형 문제를 해결하고자 하는 것이다. <표 2>의 FN:FP 비가 1:1이 아니고 다른 경우가 바로 오분류 비용의 가중을 주는 것을 의미한다. 본 논문에서의 관심은 FP 즉, 삭감을 비삭감으로 예측하는 것을 최대한 줄이려는 것이므로 FP의 오분류 비용을 변화시켜서 최적의 조건을 찾고자 하였다. Under Sampling의 경우에는 아무런 가중을 주지 않는 1:1, 1:3 그리고 1:5 3개의 오분류 비용 조건을 정의하였고, Over Sampling의 경우에는 1:1, 1:5, 1:10, 1:20 등의 오분류 비용을 주었다. 조건을 이와 같이 설정한 것은 만약 비용의 비율이 1씩 올라가는 경우, 전 단계 경우의 수와 곱해져 기하급수적으로 실험 수가 증가하여, 최대 24배 정도의 차이를 보간(interpolation)으로 파악할 수 있도록 적정한 거리를 두어 파악하여 보았다.

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

세 번째 과정은 이렇게 해서 나온 70개(Under Sampling 30개, Over Sampling 40개)의 모델에서 각각 금번 논문의 성과 지표인 소수범주의 오분류율과 심사확인건수의 축소율을 산출하는 것이다. <그림 1>에서와 같이 총 70가지의 경우에서, 2개의 지표 값 모두 0에서 1사이의 값을 가지며, 두 지표 모두 0에 가까울수록 좋은 것이라고 할 수 있다.

네 번째 과정은 세 번째 과정에서 나온 지표들을 이용하여 산점도를 그려서 어떤 경우의 조합이 가장 좌표축 (0,0)에 가까운 지를 파악하고, 이를 최종 모형으로 선정하는 과정이다.

4. 사례문제의 해결 결과

본 사례의 데이터의 구성은 <표 4>와 같다.

<표 4> 데이터의 구성

활용 데이터 변수명
환자의 고객번호(기준변수) 내원일수, 투약일수, 진료일수, 총진료비, (보험)청구액, 본인 부담액, 가산율, 본인 부담율, 연령, 성별, 외래 촬영여부, 보험종류, 명세서당 CT종류, 총 CT 입원횟수, 상병코드, 조영제 사용여부, 청구수(이상 설명변수), 삭감여부(목표변수)

본 데이터는 사례 병원에서 2004년 1월부터 6월까지 6개월 간 신경외과 CT 촬영을 하고, 건강보험심사평가원의 심의 평가를 받은 자료이다. 본 자료를 이용하여 SPSS[3]사의 데이터 마이닝 솔루션인 클레멘타인 버전 10.0의 의사결정나무 분석 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 분석 알고리즘을 지원한다. 최초의 아무런 변환없이 4개의 의사결정나무 분석 알고리즘을 수행 시 <표 5>와 같이 C5.0이 가장 좋은 분류 성능을 보였기 때문에, C5.0을 본 사례에 적용하는 의사결정나무 분석 알고리즘으로 채택을 하였다

<표 5> 4개 의사결정나무 분석 알고리즘의 정확도 및 범주 교차 비교

구분			예측		정확도
			비삭감	삭감	
C5.0	실제 값	비삭감	9,869	29	96.11%
		삭감	372	40	
C&RT	실제 값	비삭감	9,898	0	96.00%
		삭감	412	0	
CHAID	실제 값	비삭감	9,898	0	96.00%
		삭감	412	0	
QUEST	실제 값	비삭감	9,898	0	96.00%
		삭감	412	0	

위의 <표 5>에서 보는 바와 같이 C5.0을 제외하고 다른 분석 기법들은 전혀 분류를 해주지 못하고 있다는 것을 알 수 있다. 이는 C5.0을 제외하고 다른 의사결정나무 분석 기법들은 모두 규칙을 비삭감이 될 것이라고 정의를 하였고, 이는 불균형 데이터의 가장 심각한 문제를 바로 보여주고 있기도 하다.

4.1 실험결과

먼저 훈련용 데이터를 통해서, 별도의 Sampling없이 오분류 비용만 조정하여 본 결과는 다음과 같이 나타났다.

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

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

<표 6> 오분류 조정만 한 경우의 분석결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감건수/전체건수)*100
			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[4)]	실제값	비삭감	3,939	311	**49.2%**	**9.2%**
		삭감	94	97		

위의 <표 6>을 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 (예측삭감건수/전체건수)*100의 지표 역시 오분류 비용에 어떤 값을 주어도 큰 변화가 없는 것으로 나타났다. 이는 단순히 오분류 비용만 조정을 하는 경우에는 어떠한 오분류 비용을 선택하여도, 일정한 결과를 보이는 것을 알 수 있다. 본 결과는 추후 본 논문에서 제시하고 방법들의 결과와 비교될 것이다.

다음의 <표 7>부터 <표 13>까지는 Sampling의 조건 변화와 오분류 비용 조건 변화에 따른 예측과 실제값의 교차표와 본 논문의 성능 평가 기준인 소수범주 오분류율과 심사확인 건수의 축소 비율을 나타낸 것이다. 소수범주 오분류율은 전체 삭감 건 191건 중 예측을 잘못한 비율로써, 모델의 정확성을 의미하고, 심사 확인 건수 축소비율은 전체 4,441건을 100%로 두었을 때, 심사 확인 건수 축소비율만큼만 진료 건을 심사 및 검토하면, *100-소수범주 오분류율(%)* 만큼의 정확한 삭감 건을 찾아낼 수 있다는 것이다. 또한 (예측삭감건수/전체건수)*100 이라는 지표는 병원에서 관련된 심사 건의 축소를 파악할 수 있는 지표로써, 낮으면 낮을수록 적은 수의 심사를 하는 것을 의미한다. 다음의 <표 7>부터 <표 10>까지는 첫 번째 방법의 결과이다.

<표 7> 오분류 비용 1일 때 다양한 Under Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감건수/전체건수)*100
			비삭감	삭감		
1	50:50	비삭감	3,625	625	**18.8%**	**17.6%**
		삭감	36	155		
1	65:35	비삭감	3,740	510	**24.1%**	**14.7%**
		삭감	46	145		
1	75:25	비삭감	3,908	342	**34.0%**	**10.5%**
		삭감	65	126		
1	80:20	비삭감	3,863	387	**29.8%**	**11.7%**
		삭감	57	134		
1	90:10	비삭감	4,158	92	**73.2%**	**3.2%**
		삭감	140	51		

<표 8> 오분류 비용 3일 때 다양한 Under Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감건수/전체건수)*100
			비삭감	삭감		
1	50:50	비삭감	3,625	625	**18.8%**	**17.6%**
		삭감	36	155		
1	65:35	비삭감	3,740	510	**24.1%**	**14.7%**
		삭감	46	145		
1	75:25	비삭감	3,908	342	**34.0%**	**10.5%**
		삭감	65	126		
1	80:20	비삭감	3,863	387	**29.8%**	**11.7%**
		삭감	57	134		
1	90:10	비삭감	4,158	92	**73.2%**	**3.2%**
		삭감	140	51		

<표 9> 오분류 비용 5일 때 다양한 Under Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감건수/전체건수)*100
			비삭감	삭감		
5	50:50	비삭감	1,801	2,449	**4.2%**	**59.3%**
		삭감	8	183		
5	65:35	비삭감	2,810	1,440	**10.0%**	**36.3%**
		삭감	21	170		
5	75:25	비삭감	3,414	836	**16.2%**	**22.4%**
		삭감	31	160		
5	80:20	비삭감	3,374	876	**13.6%**	**23.4%**
		삭감	26	165		
5	90:10	비삭감	3,704	546	**24.6%**	**15.5%**
		삭감	47	144		

위에서 <표 7>은 오분류 비용을 1로 한 것으로, 일반적으로 오분류 비용을 적용하지 않고, 단순히 Under Sampling만 한 것을 의미한다. 다음의 <표 10>부터 <표 13>까지는 2번째 방법의 결과이다.

<표 10> 오분류 비용 1일 때 다양한 Over Sampling 결과

FA	구분	예측	소수범주	(예측삭감

4) 24를 최대로 한 것은 삭감과 비삭감의 비율이 24배가 나타나기 때문이다.(허명회, 이용구, 2003)

오분류비용배수			비삭감	삭감	오분류율	건수/전체건수)*100
1	50:50	비삭감	4,030	220	56.0%	6.9%
		삭감	107	84		
1	65:35	비삭감	4,030	220	55.5%	6.9%
		삭감	106	85		
1	75:25	비삭감	4,043	207	57.1%	6.5%
		삭감	109	82		
1	80:20	비삭감	4,044	206	58.6%	6.4%
		삭감	112	79		
1	90:10	비삭감	4,118	132	63.4%	4.5%
		삭감	121	70		

<표 11> 오분류 비용 5일 때 다양한 Over Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감 건수/전체 건수)*100
			비삭감	삭감		
5	50:50	비삭감	3,808	442	27.7%	13.1%
		삭감	53	138		
5	66:34	비삭감	3,873	377	33.0%	11.4%
		삭감	63	128		
5	75:25	비삭감	3,879	371	37.7%	11.0%
		삭감	72	119		
5	80:20	비삭감	3,975	275	51.8%	8.3%
		삭감	99	92		
5	90:10	비삭감	3.956	294	52.3%	8.7%
		삭감	100	91		

<표 12> 오분류 비용 10일 때 다양한 Over Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감 건수/전체 건수)*100
			비삭감	삭감		
10	50:50	비삭감	3,556	694	20.4%	19.1%
		삭감	39	152		
10	66:34	비삭감	3,728	522	24.6%	15.0%
		삭감	47	144		
10	75:25	비삭감	3,795	455	26.7%	13.4%
		삭감	51	140		
10	80:20	비삭감	3,830	420	31.9%	12.4%
		삭감	61	130		
10	90:10	비삭감	3.927	323	44.0%	9.7%
		삭감	84	107		

<표 13> 오분류 비용 20일 때 다양한 Over Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감 건수/전체 건수)*100
			비삭감	삭감		
20	50:50	비삭감	3,522	888	16.2%	22.8%
		삭감	31	160		
20	66:34	비삭감	3,424	826	14.7%	22.3%
		삭감	28	163		
20	75:25	비삭감	3,476	774	17.8%	21.0%
		삭감	34	157		
20	80:20	비삭감	3,718	532	25.7%	15.2%
		삭감	49	142		
20	90:10	비삭감	3.798	452	29.3%	13.2%
		삭감	56	135		

위의 <표 10>의 결과는 역시 일반적으로 Over Sampling 만 수행한 결과와 동일하다. <표 7>과 <표 10>의 결과 2 개만을 비교하여 볼 때 이 사례의 경우는 Under Sampling 이 좀 더 효율적으로 판단이 되어진다. 다음의 <표 14>부터 <표 17>까지는 SMOTE 방법을 응용한 Over Sampling을 이용하여 나온 결과이다

<표 14> 오분류 비용 1일 때 다양한 SMOTE 응용 Over Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감 건수/전체 건수)*100
			비삭감	삭감		
1	50:50	비삭감	3,716	534	33.5%	14.9%
		삭감	64	127		
1	65:35	비삭감	3,739	511	32.5%	14.4%
		삭감	62	129		
1	75:25	비삭감	3,785	465	34.0%	13.3%
		삭감	65	126		
1	80:20	비삭감	3,796	454	33.5%	13.1%
		삭감	64	127		
1	90:10	비삭감	3,867	383	41.9%	11.1%
		삭감	80	111		

<표 15> 오분류 비용 1일 때 다양한 SMOTE 응용 Over Sampling 결과

FA 오분류비용배수	구분		예측		소수범주 오분류율	(예측삭감 건수/전체 건수)*100
			비삭감	삭감		
5	50:50	비삭감	3,606	644	22.0%	17.9%
		삭감	42	149		
5	65:35	비삭감	3,636	614	23.0%	17.1%
		삭감	44	147		
5	75:	비삭감	3,640	610	25.1%	17.0%

	25	삭감	48	143		
5	80:20	비삭감	3,659	591	**27.7%**	**16.4%**
		삭감	53	138		
5	90:10	비삭감	3,759	491	**29.8%**	**14.1%**
		삭감	57	134		

<표 16> 오분류 비용 10일 때 다양한 SMOTE 응용 Over Sampling 결과

FA 오분류비용배수	구분		예측		**소수범주 오분류율**	**(예측삭감건수/전체건수)*100**
			비삭감	삭감		
10	50:50	비삭감	3,397	853	**13.6%**	**22.9%**
		삭감	26	165		
10	65:35	비삭감	3,558	692	**19.9%**	**19.0%**
		삭감	38	153		
10	75:25	비삭감	3,625	625	**22.5%**	**17.4%**
		삭감	43	148		
10	80:20	비삭감	3,634	616	**23.6%**	**17.2%**
		삭감	45	146		
10	90:10	비삭감	3,668	582	**26.7%**	**16.3%**
		삭감	51	140		

<표 17> 오분류 비용 20일 때 다양한 SMOTE 응용 Over Sampling 결과

FA 오분류비용배수	구분		예측		**소수범주 오분류율**	**(예측삭감건수/전체건수)*100**
			비삭감	삭감		
20	50:50	비삭감	3,084	1,166	**12.6%**	**30.0%**
		삭감	24	167		
20	65:35	비삭감	3,353	897	**13.6%**	**23.9%**
		삭감	26	165		
20	75:25	비삭감	3,433	817	**14.7%**	**22.1%**
		삭감	28	163		
20	80:20	비삭감	3,431	819	**14.7%**	**22.1%**
		삭감	28	163		
20	90:10	비삭감	3,630	620	**27.7%**	**17.1%**
		삭감	53	138		

위의 <표 14>부터 <표 17>까지를 보면 오분류 비용을 20으로 잡은 경우가 오분류율은 계속 떨어지는 것을 알 수 있으며, <표 17>에서 Over Sampling 비율이 50:50 인 경우가 소수 범주 오분류율이 가장 낮은 것을 알 수 있다. 그러나 이 경우 (예측삭감건수 / 전체건수)*100 지표는 30%로 심사업무량 감소 기준에 거의 도달을 한 것을 알 수 있다. 다음의 <표 18>부터 <표 20>까지는 이상치를 제외하고 Under Sampling을 한 결과이다.

<표 18> 오분류 비용 1일 때 이상치를 제외한 Under Sampling 결과

FA	구분	예측		**소수범주**	**(예측삭감**
오분류비용배수		비삭감	삭감	**오분류율**	**건수/전체건수)*100**
1	50:50	비삭감	3,516	734	**16.2%**	**20.1%**
		삭감	31	160		
1	65:35	비삭감	3,561	689	**17.3%**	**19.1%**
		삭감	33	158		
1	75:25	비삭감	3,829	412	**29.3%**	**12.3%**
		삭감	56	135		
1	80:20	비삭감	3,770	480	**22.0%**	**14.2%**
		삭감	42	149		
1	90:10	비삭감	3,874	376	**34.0%**	**11.3%**
		삭감	65	126		

<표 19> 오분류 비용 3일 때 이상치를 제외한 Under Sampling 결과

FA 오분류비용배수	구분		예측		**소수범주 오분류율**	**(예측삭감건수/전체건수)*100**
			비삭감	삭감		
3	50:50	비삭감	3,003	1,247	**11.5%**	**31.9%**
		삭감	22	169		
3	65:35	비삭감	3,400	850	**14.7%**	**22.8%**
		삭감	28	163		
3	75:25	비삭감	3,488	762	**29.3%**	**12.3%**
		삭감	32	159		
3	80:20	비삭감	3,661	589	**22.0%**	**14.2%**
		삭감	40	151		
3	90:10	비삭감	3,797	453	**34.0%**	**11.3%**
		삭감	58	133		

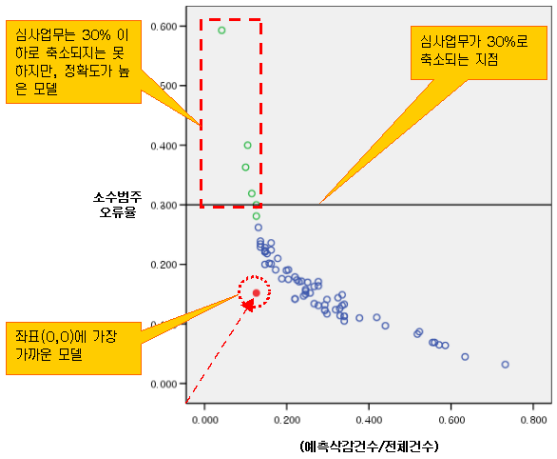
<표 20> 오분류 비용 5일 때 이상치를 제외한 Under Sampling 결과

FA 오분류비용배수	구분		예측		**소수범주 오분류율**	**(예측삭감건수/전체건수)*100**
			비삭감	삭감		
5	50:50	비삭감	2,648	1,607	**10.5%**	**40.0%**
		삭감	20	171		
5	65:35	비삭감	3,251	999	**13.1%**	**26.2%**
		삭감	25	166		
5	75:25	비삭감	3,526	724	**14.7%**	**20.0%**
		삭감	28	163		
5	80:20	비삭감	3,626	624	**20.4%**	**17.5%**
		삭감	39	152		
5	90:10	비삭감	3,692	558	**24.6%**	**15.8%**
		삭감	47	144		

위의 14개의 표에서 최초 병원 담당자들이 제시한 기본 가정인 심사 업무 건이 30%이하로 줄어야 한다는 가정에

위배되는 것은 <표 9>의 오분류 비용이 5이고, Under Sampling을 통해서 비삭감과 삭감의 비율이 50:50과 66:34인 경우로, 각각 심사확인건수의 축소비율이 59.3%와 36.3%로 나타났다. 또한 <표 19>에서 비삭감과 삭감의 비율이 50:50인 경우, 그리고 <표 20>에서, 비삭감과 삭감의 비율이 50:50인 경우가 위배되었다.[5] 70개의 모형 중 이들 모형은 삭제됨을 전제하고, 전체적인 비교를 통해서 최적 모형을 찾아낸 것이 <그림 2>와 같다.

<그림 2>에서 보면 좌표축 (0,0)에 가장 가까운 것은 소수 범주 오류율이 0.126이고, 심사확인건수 축소율이 0.152인 모형이다. 해당 모형은 <표 8>에 있으며, Under Sampling은 비삭감 : 삭감 = 66 : 34이고, 오분류 비용을 3을 주었을 때 생성된 모형이다. 또한 이 때의 Under Sampling은 단순히 임의로 추출한 경우이다. 따라서 이 모형이 소수 범주 오류와 심사 건수의 축소라는 2개의 축을 동등하게 가중치를 두었을 때 가장 효율적으로 선택할 수 있는 모형이라고 할 수 있다. 만약 심사 업무의 축소에 대한 가정이 없다면, <표 9>에서 단순한 Under Sampling에 Sampling 조건이 비삭감 : 삭감 = 50 : 50 이고, 오분류 비용을 5를 주었을 때 심사 업무는 59.3% 정도만 줄일 수 있지만, 오분류율을 4.2%로 획기적으로 줄일 수 있게 되는 것으로 나타났다. 또한 <표 19>에서 이상치를 제거한 Under Sampling의 비삭감 : 삭감 = 50 : 50이고, 오분류 비용을 3으로 주었을 때는 업무량은 31.9%로 기준에서 약간만 넘어가지만, 오분류율은 11.5%로, 좌표(0,0)과 가장 가까운 모형보다 정확도가 더 나은 모형을 만들 수 있을 것이다. 앞에서도 언급했듯이, 업무량보다는 삭감 건의 발견이 중요한 목적인 경우에 이러한 모형을 활용하는 것도 하나의 선택 방법이 될 수 있다. 본 사례에서 또 하나 살펴봐야 할 것은 안정성이다. 좌표(0,0)과 가장 가까운 모델의 경우는 전체적인 다른 모형 추세와 비교하여 이상치처럼 돌출이 되어져 있어, 생각에 따라서는 Sampling에 의하여 우연히 발생한 특별한 모형일 수 있다. 따라서 다른 사례에 적용을 할 때, 안정적이기 위해서는, 방법별로 가장 안정적인 모형을 찾을 필요가 있다. 이를 위해 방법별로 각각 2개 지표 평균을 계산한 것이 다음 <표 21>과 같다.



<그림 2> 최적 모형의 선정

5) Under Sampling의 오분류 비용을 더 늘려서 실험을 하지 못한 건 본 가정 때문이었다. 오분류 비용을 더 늘리면 심사확인 건수는 더욱 증가하기 때문이다.

<표 21> 방법별 평균의 차이

방법	방법 설명	소수범주 오분류율의 평균	(예측삭감건수/전체건수)*100의 평균	평균비교 (1-way ANOVA) : F값 (p-값)
1	(Under Sampling을 통한 범주 균형 변화 조건 5가지) X (오분류 비용의 변화 조건 3가지)	**0.225**	**0.208**	**소수범주 오분류율** : 6.203(0.001), **(예측삭감건수/전체건수)** : 3.652(0.017)
2	(Over Sampling을 통한 범주 균형 변화 조건 5가지) X (오분류 비용의 변화 조건 4가지)	**0.303**	**0.145**	
3	k-평균의 군집거리와 SMOTE 통한 범주 균형 변화 조건 5가지) X (오분류 비용의 변화 조건 4가지)	**0.211**	**0.194**	
4	k-평균의 군집거리를 통해서, 다수 범주 중 이상치를 제외한 Under Sampling을 이용한 범주 균형 변화 조건 5가지) X (오분류 비용의 변화 조건 3가지)	**0.209**	**0.193**	

<표 21>을 살펴보면, 이상치를 제거한 후 Under Sampling을 한 경우가 소수범주 오분류율이 가장 좋은 것으로 나타났고, 다음으로 SMOTE를 응용한 Over Sampling 방법이 좋은 것으로 나타났다. 그리고 이는 단순한 임의의 Under Sampling이나 Over Sampling 보다도 좋은 것으로 나타났다. 이는 Sampling을 임의적으로 하는 것보다는 Sampling의 조정 및 변화를 주는 경우 더욱 더 최적의 결과를 도출한다는 선행연구(Batista 등, 2004; Chawla 등, 2002; Laurikkala, 2001; Japkowicz, 2000)의 결론에 부합한다고 할 수 있다. 심사 건수의 축소 지표에서는 단순한 Over Sampling이 가장 업무를 평균적으로 축소시켜주는 것으로 나타났으나, 전체적으로 기준인 30% 업무 축소에는 모든 방법이 전부 적합한 것으로 나타났다. 이를 일원배치 분산분석(One-way ANOVA)을 이용하여, 검정한 결과 유의수준 95%하에서 유의한 것으로 나타났다. 또한 위의 여러 표에서 나온 70개의 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 한 것과 차이가 있는지 확인을 해 볼 필요가 있다. 위의 <표 6>은 Sampling 없이 오분류 비용만을 조정한 것이고, <표 7>은 임의의 Under Sampling만 수행한 것, 그리고 <표 10>은 단순한 Over Sampling을 수행한 결과이다. 또한 <표 14>는 SMOTE를 응용한 변형된 Over Sampling만 이용을 한 것이고, <표 18>은 이상치를 제거한 Under Sampling 방법만 이용을 한 것이다. 이들을 제외한 나머지 분석 결과표들은 전부 Sampling 방법과 오분류 비용의 조정을 동시에 한 것이라고 할 수 있다. 이들 간의 평균 차이를 비교한 것이 <표 22>이다.

<표 22> Sampling과 오분류 비용조정을 복합한 방법과 단독으로만 사용한 방법의 비교

비교 구분	소수범주	(예측삭감건수/	소수범주	(예측삭감건수/

	오분류율의 평균	전체건수)*100의 평균	오분류율 검정결과 :p 값	전체건수)*100 검정결과 :p값
임의의 Under Sampling vs. (임의의 Under Sampling + 오분류 비용 조정)	0.3598 vs. 0.1576	0.1154 vs. 0.2548	0.014	0.010
단순한 오분류 조정 vs. (임의의 Under Sampling + 오분류 비용 조정)	0.5675 vs. 0.1576	0.0692 vs. 0.2548	0.001	0.001
임의의 Over Sampling vs. (임의의 Over Sampling + 오분류 비용 조정)	0.5812 vs. 0.3025	0.0624 vs. 0.1444	0.001	0.001
단순한 오분류 조정 vs (임의의 Over Sampling + 오분류 비용 조정)	0.5675 vs. 0.3025	0.0692 vs. 0.1444	0.004	0.002
SMOTE를 응용한 Over Sampling vs. (SMOTE를 응용한 Over Sampling + 오분류 비용 조정)	0.3508 vs. 0.2115	0.1336 vs. 0.1937	0.001	0.002
단순한 오분류 조정 vs. (SMOTE를 응용한 Over Sampling + 오분류 비용 조정)	0.5675 vs. 0.2115	0.0692 vs. 0.1937	0.000	0.000
이상치를 제외한 Under Sampling vs. (이상치를 제외한 Under Sampling + 오분류 비용 조정)	0.2360 vs. 0.1960	0.1520 vs. 0.2120	0.243	0.219
단순한 오분류 조정 vs. (이상치를 제외한 Under Sampling + 오분류 비용 조정)	0.5675 vs. 0.1960	0.0692 vs. 0.2120	0.001	0.001

먼저 데이터가 15~20개로 많지 않기 때문에 비모수 검정 방법 중 2개의 독립적인 집단을 비교하는 Mann-Whitney 방법을 이용하여 검정을 하였다. 위의 <표 22>의 결과를 보면, 이상치를 제외한 Under Sampling과 이상치를 제외한 Under Sampling에 오분류 비용을 조정한 것을 제외하고는 유의수준 95%이내에서 Sampling과 오분류 비용을 동시에 조정하는 방법이 소수범주의 오분류율을 낮춘다고 나타났다. 심사건수의 축소라는 지표의 측면에서 보면, 2가지 방법을 혼합하여 사용한 경우가 더욱 많은 심사 건을 수행하게 되는데, 전부 본 사례 병원에서 제시하는 30%의 기준에는 부합하는 것으로 나타났다. 이 결과를 통해서, 단순히 Sampling이나 오분류 비용만 조정하는 방법보다는 이 2가지를 같이 병행하여 모델을 개발하는 것이, 본 논문의 기준에서는 더욱 효과적인 것을 알 수 있다.

5. 결론

본 연구에서는 데이터 마이닝의 지도 학습 기법인 의사결정나무 분석을 이용하여, 병원의 상시 업무인 보험료 심사 업무를 줄이고, 효율을 높일 수 있는 방법을 제시하고 있다. 특히 그 중에서도, 데이터의 불균형으로 의사결정나무 분석을 제대로 활용 못하는 부득이한 상황에서, Sampling과 오분류 비용 등의 다양한 불균형 해소 기법의 조합을 통하여, 가장 효율적이고, 안정성 있는 모형을 찾아내는 방법을 제시하고자 하였다. 본 논문에서 제시한 불균형 해소 방법을 다른 종합/개인 병원이나, 신경외과의 CT 이외에 다른 진료과 분야에도 활용할 수 있을 것으로 판단된다. 또한 의료 분야 이외에도 불법 카드 사기 및 보험 사기의 경우에도 본 방법을 적용하여, 보다 나은 모형을 생성할 수도 있을 것이다. 본 사례의 병원에서는 데이터 마이닝을 이용한 보험료 청구삭감 판정 시스템을 본 프로젝트로 이행하기 전 효과성 측정을 하기 위하여, 신경외과 CT건에 한하여 Pilot 프로젝트로 수행하였고, 본 논문에서 제시한 것과 같은 성과를 얻었다. 이를 전체적으로 적용하기 위해서는 시스템화 시키는 것이 필요하고, 현재 이를 추진하고 있다. 향후 본 논문에서 제시한 방법이외에도 다양한 데이터 불균형 해소 방법을 적용하여, 보험료 청구 심사 분석 업무에 적용을 하여 보는 연구와 다양한 병원 그리고 진료과에서 적용을 하여, 효율적인 모형을 찾아내는 추가적인 연구가 필요하다. 또한, 청구 삭감의 판정 이외에 여러 산업에서 문제가 되고 있는 각종 데이터 불균형 문제에 대한 효율적인 알고리즘 개발과 개발된 알고리즘의 비교 연구가 필요하다.

참고문헌

강필성, 이형주, 조성준, “데이터 불균형 문제에서의 SVM 앙상블 기법의 적용”, *한국정보과학회 가을 학술발표논문집*, 제31권, 2호, 2005, pp.706-708.

김지현, 정종빈, “계급 불균형 자료의 분류 : 훈련표본 구성방법에 따른 효과”, *응용통계연구*, 제17권, 3호, 2004, pp445-457.

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 3호, 2004, pp.71-97.

이용구, 통계학 원론, 율곡출판사, 1992.

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*, 6(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*, 16, 2002, pp.231-357.

Chawla, N. V., Nathalie Japkowicz, and Aleksander Kolcz, "Editorial : Special issue onlearning

from imbalanced data sets." *SIGKDD Exploring*, 6(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*, 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*, 6(1), 2004, pp.30-39.

Hart, P.E., "The Condensed Nearest Neighbor Rule", *IEEE, Transactions on Information Theory IT-14*, 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*, 6(1), 2004, pp.40-49.

Kass, G. "An exploratory technique for investigating large quantities of categorical data." *Applied Statistics*, Vol.29, 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*, 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* 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*, 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

http://www.hira.or.kr