응용통계연구 (2011)
24(3), 477–484
DOI: 10.5351/KJAS.2011.24.3.477

목표 범주가 희귀한 자료의 과대표본추출에 대한 연구

김은나[1] · 이성건[2] · 최종후[3]

[1]비씨카드, [2]성신여자대학교 통계학과, [3]고려대학교 정보통계학과

(2011년 3월 접수, 2011년 6월 채택)

요 약

반응/미반응 목표변수를 갖는 모집단에서 관심 목표범주의 빈도가 극히 작을 경우, 즉 희귀할(rare) 경우, 모형 구축을 위한 데이터마트를 형성할 때 반응/미반응 범주 구성비는 구축된 모형의 성능에 영향을 준다. 본 연구는 이러한 점에 착안하여 반응/미반응 범주 구성비와 모형성능의 관련성을 모형평가 통계량에 기반하여 판단한다. 이로써 데이터마트 형성에 이상적인 반응/미반응 범주 구성비를 탐지하려는데 본 연구의 목적을 두고 있다. 또한 일반적으로 목표범주의 빈도가 희귀할 경우, 분할 표본추출에 의하여 희귀사건(rare event)을 과대표본추출(oversampling)하는 것이 일반적이며, 이로부터 기인하는 사후확률에 대한 편향을 조정하게 된다. 본 연구에서는 사후확률 조정방법으로 오프셋(offset) 방법과 가중치 방법(sampling weights)을 적용하고 이를 비교하였다.

주요용어: 과대표본추출, 사후확률 조정, 희귀사건, 오프셋 방법, 가중치 방법.

1. 서론

통계학 및 데이터마이닝 분야에서 많이 수행하는 DM(direct mail)에 응답할 가능성이 높은 고객의 예측, 신용카드 이용 고객의 사기 탐지(fraud detection), 우/불량 고객 식별을 위한 신용평가 모형개발, 보험회사의 이탈고객 분석, 개인 휴대통신의 이탈고객 분석, 제조공정의 불량제품 탐지, 환자의 질병진단, 도산기업의 예측 등과 같은 자료에서는 일반적으로 목표 집단(목표변수의 관심 범주)의 빈도가 상대적으로 희귀하다.

본 논문의 연구 목적은 반응/미반응 목표변수를 갖는 모집단에서 목표범주의 빈도가 희귀한 경우 데이터마트의 반응/미반응 범주 구성비에 의존하는 구축모형의 성능을 비교 실험함으로써 바람직한 구축모형을 도출하기 위한 시사점을 도출하고자 한다. 예를 들어, 신용불량 고객들의 특성을 파악하기 위해 자료를 수집하면 대개 정상적인 고객은 95% 이상을 차지하고 불량인 고객은 5% 미만일 때가 많다. 이러한 현상은 도산기업이나 이탈고객의 유형을 발견하기 위해 자료를 수집해 보아도 마찬가지이다. 이와 같은 경우에는 효과적인 모형을 구축하기 위하여 목표변수의 범주 간 수적 형평성을 맞추는 것이 바람직하다 (장남식 등, 1999).

목표변수의 범주 간 수적 형평성을 맞추기 위하여 표본을 추출할 때 '표본크기(sample size)가 얼마나 되어야 하는가?'는 모형을 개발하고자 하는 경우에 나타나는 일반적인 질문 중 하나이지만, 불행하게도 정확한 해답은 없다. 표본크기는 '목표 집단의 반응율이 어느 정도로 예측되는가?'(목표 집단의 반응률이 낮을수록 더 많은 데이터가 필요하다), '모형개발을 위해 얼마나 많은 변수들을 사용할 계획인

[3]교신저자: (339- 700) 충남 연기군 조치원읍 서창리 208, 고려대학교 정보통계학과, 교수.
E-mail: jhchoi@korea.ac.kr

표 **2.1.** 원시자료

캠페인월	Resp.(= Y)	Non Resp.(= N)	Total	Resp. Rate
200808	2,970	67,151	70,121	4.2%
200809	2,242	73,565	75,807	3.0%
200810	2,320	75,465	77,785	3.0%
200811	2,287	84,141	86,428	2.6%
200812	2,552	77,171	79,723	3.2%
200901	1,650	77,136	78,786	2.1%
Total	14,021	454,629	468,650	3.0%

가?'(많은 변수들을 고려할수록 더 많은 데이터가 필요하다)와 같은 여러 가지 요인들에 의존한다. 따라서 분석의 목적에 따라 적절하게 추출된 표본의 활용은 비용과 시간의 절약, 보다 효율적인 모형화 작업을 위해서 매우 중요하다 (강현철 등, 2006).

이에 본 연구에서는 반응/미반응 목표변수를 갖는 모집단에서 데이터마트 구성비에 따라 모형 성능 비교 실험을 함으로써 바람직한 모형을 구축하기 위한 데이터마트 구성에 대한 시사점을 도출하고자 한다. 모형으로는 의사결정나무모형, 로지스틱 회귀모형, 신경망모형이 활용되며, 모형 평가를 위한 통계량은 정확도(accuracy), 특이도(specificity), 민감도(sensitivity)를 사용하였다.

본 연구의 2절에서는 데이터마트 구성비에 따른 모형 성능을 비교실험하고, 3절에서는 과대표본추출에 대한 사후확률의 조정에 대해 논의하였다.

2. 데이터마트 구성비에 따른 모형 성능 비교

2.1. 분석 자료

본 연구를 위하여 사용된 자료는 A카드사에서 교차판매(cross-sell)를 위하여 TM(tele-marketing)을 수행한 2008년 8월부터 2009년 1월까지 6개월간 대상 고객자료이다. 자료에서 목표변수는 TM 대상 고객의 캠페인 반응/미반응의 이분형 변수(binary variable)이다. 본 연구에 사용된 원시자료의 크기는 표 2.1과 같고, 반응/미반응의 비율이 3.0%/97.0%로 TM에 반응한 고객의 비율이 매우 낮아 미반응의 비율이 반응의 약 32.4배를 차지하고 있다.

모형에 고려된 설명변수는 A카드사의 내부정보와 은행연합회(KFB), 한국신용평가(KIS), 한국신용정보(NICE), 한국개인신용(KCB) 등의 외부정보이며, 분석에 이용되는 변수는 총 198개이다.

2.2. 데이터마트 구성 및 모형 구축

반응/미반응 범주의 데이터마트 구성비에 따른 구축모형의 정확도, 민감도, 특이도를 비교해 보고자 반응 고객은 전체를 추출하고 미반응 고객은 반응 고객의 20배, 15배, 10배, 8배, 6배, 4배, 2배, 1배를 표본으로 추출하여 데이터마트를 형성하였다. 표본추출방법으로는 목표변수 각 범주에 대하여 단순임의추출법(simple random sampling)을 사용하였다.

분석 모형으로는 의사결정나무(decision tree), 로지스틱 회귀(logistic regression), 신경망(neural network)을 각각 적용하여 모형을 구축한다. 이때 분석용 자료(training data) 70%, 검증용 자료(Validation Data) 30%로 분할하였다.

의사결정나무는 CHAID 알고리즘, 유의수준 0.20를 적용하여 모형을 구축하였고, 로지스틱 회귀는 단계적 방법을 사용하여 모형을 구축하였다. 단계적 방법의 매 단계에서 유의수준 0.20하에서 유의한 변

표 **2.2.** 미반응 개체에 대한 표본추출

반응 : 미반응 비율	Resp.(= Y)	Non Resp.(= N)	Total	Resp. Rate
1 : 20	14,021	280,420	294,441	4.8%
1 : 15	14,021	210,315	224,336	6.3%
1 : 10	14,021	140,210	154,231	9.1%
1 : 8	14,021	112,168	126,189	11.1%
1 : 6	14,021	84,126	98,147	14.3%
1 : 4	14,021	56,084	70,105	20.0%
1 : 2	14,021	28,042	42,063	33.3%
1 : 1	14,021	14,021	28,042	50.0%

표 **2.3.** 분류기준값 0.10일 때 정확도, 민감도, 특이도

표본추출	Tree			Logistic Regression			Neural Network		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
원시자료	0.962	0.037	0.991	0.957	0.066	0.985	0.961	0.037	0.991
1 : 20	0.882	0.224	0.916	0.889	0.230	0.923	0.882	0.246	0.915
1 : 15	0.822	0.320	0.856	0.817	0.372	0.848	0.807	0.397	0.835
1 : 10	**0.689**	**0.528**	**0.705**	**0.683**	**0.586**	**0.693**	**0.717**	**0.506**	**0.738**
1 : 8	**0.638**	**0.612**	**0.641**	**0.600**	**0.684**	**0.589**	**0.623**	**0.661**	**0.618**
1 : 6	0.542	0.707	0.514	0.495	0.797	0.443	0.517	0.778	0.472
1 : 4	0.213	0.992	0.012	0.386	0.908	0.251	0.393	0.903	0.261
1 : 2	0.342	0.999	0.002	0.367	0.991	0.042	0.342	1.000	0.000
1 : 1	0.511	1.000	0.000	0.512	0.999	0.003	0.511	1.000	0.000

표 **2.4.** 분류기준값 0.20일 때 정확도, 민감도, 특이도

표본추출	Tree			Logistic Regression			Neural Network		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
원시자료	0.969	0.000	1.000	0.969	0.000	1.000	0.969	0.000	1.000
1 : 20	0.951	0.000	1.000	0.950	0.009	0.998	0.950	0.005	0.999
1 : 15	0.928	0.040	0.989	0.929	0.046	0.990	0.931	0.031	0.993
1 : 10	0.857	0.200	0.925	0.874	0.175	0.945	0.873	0.171	0.946
1 : 8	0.815	0.277	0.884	0.830	0.270	0.902	0.823	0.294	0.891
1 : 6	0.756	0.391	0.819	0.755	0.430	0.811	0.755	0.439	0.809
1 : 4	**0.661**	**0.562**	**0.686**	**0.639**	**0.635**	**0.639**	**0.650**	**0.621**	**0.658**
1 : 2	0.423	0.928	0.161	0.506	0.886	0.309	0.522	0.872	0.341
1 : 1	0.516	0.991	0.020	0.532	0.981	0.064	0.518	0.992	0.024

수는 모형에 진입하도록 하고, 유의수준 0.05하에서 유의하지 않은 변수는 제거되도록 지정하였다. 신경망은 MLP 모형(1개 은닉층, 3개 은닉노드)을 적용하였다.

2.3. 모형 성능 비교 실험

각 데이터마트의 검증용 자료 30%에 대한 오분류표를 기초로 분류기준값(cut off 또는 threshold)에 따라 정확도, 민감도, 특이도를 비교하였다. 분류 기준값으로는 0.1, 0.2, 0.3, 0.4 0.5를 고려하였다. 결과는 표 2.3∼표 2.7과 같다. 좋은 모형은 정확도 뿐만 아니라 민감도, 특이도 등도 함께 높게 나타나야 하지만 민감도와 특이도는 서로 반비례의 관계에 있기 때문에 어느 한쪽으로만 평가하기는 어렵다. 따라서 민감도, 특이도가 유사하면서 높은 값을 가지는 결과를 선택하는 것도 하나의 대안이 될 수 있다.

표 **2.5.** 분류기준값 0.30일 때 정확도, 민감도, 특이도

표본추출	Tree			Logistic Regression			Neural Network		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
원시자료	0.969	0.000	1.000	0.969	0.000	1.000	0.969	0.000	1.000
1 : 20	0.951	0.000	1.000	0.951	0.000	1.000	0.951	0.000	1.000
1 : 15	0.936	0.000	1.000	0.935	0.001	1.000	0.936	0.000	1.000
1 : 10	0.905	0.015	0.996	0.904	0.022	0.995	0.902	0.031	0.992
1 : 8	0.882	0.028	0.993	0.880	0.057	0.986	0.880	0.056	0.987
1 : 6	0.829	0.158	0.945	0.836	0.170	0.951	0.832	0.180	0.944
1 : 4	0.768	0.255	0.901	0.754	0.369	0.853	0.750	0.384	0.844
1 : 2	**0.610**	**0.673**	**0.577**	**.622**	**0.703**	**0.580**	**0.631**	**0.676**	**0.608**
1 : 1	0.553	0.935	0.155	0.583	0.917	0.235	0.568	0.925	0.195

표 **2.6.** 분류기준값 0.40일 때 정확도, 민감도, 특이도

표본추출	Tree			Logistic Regression			Neural Network		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
원시자료	0.969	0.000	1.000	0.969	0.000	1.000	0.969	0.000	1.000
1 : 20	0.951	0.000	1.000	0.951	0.000	1.000	0.951	0.000	1.000
1 : 15	0.936	0.000	1.000	0.936	0.000	1.000	0.936	0.000	1.000
1 : 10	0.907	0.000	1.000	0.906	0.002	0.999	0.906	0.003	0.999
1 : 8	0.886	0.000	1.000	0.885	0.006	0.999	0.886	0.000	1.000
1 : 6	0.889	0.036	0.994	0.852	0.039	0.992	0.858	0.040	1.000
1 : 4	0.787	0.128	0.957	0.791	0.157	0.956	0.789	0.179	0.947
1 : 2	0.675	0.386	0.825	0.680	0.490	0.779	0.681	0.501	0.775
1 : 1	**0.597**	**0.806**	**0.379**	**0.626**	**0.799**	**0.446**	**0.612**	**0.788**	**0.429**

표 **2.7.** 분류기준값 0.50일 때 정확도, 민감도, 특이도

표본추출	Tree			Logistic Regression			Neural Network		
	정확도	민감도	특이도	정확도	민감도	특이도	정확도	민감도	특이도
원시자료	0.969	0.000	1.000	0.969	0.000	1.000	0.969	0.000	1.000
1 : 20	0.951	0.000	1.000	0.951	0.000	1.000	0.951	0.000	1.000
1 : 15	0.936	0.000	1.000	0.936	0.000	1.000	0.936	0.000	1.000
1 : 10	0.907	0.000	1.000	0.906	0.000	1.000	0.906	0.000	1.000
1 : 8	0.886	0.000	1.000	0.886	0.000	1.000	0.886	0.000	1.000
1 : 6	0.853	0.000	1.000	0.853	0.003	0.999	0.853	0.000	1.000
1 : 4	0.795	0.000	1.000	0.796	0.039	0.991	0.795	0.034	0.992
1 : 2	0.682	0.223	0.920	0.689	0.272	0.906	0.690	0.307	0.889
1 : 1	**0.628**	**0.641**	**0.615**	**0.636**	**0.633**	**0.640**	**0.628**	**0.608**	**0.650**

∗ 정확도 = [정분류 개체수]/[전체 개체수]
∗ 민감도 = [진양(true positive)의 개체수]/[실제 양(positive)의 개체수]
∗ 특이도 = [진음(true negative)의 개체수/[실제 음(negative)의 개체수]

결과를 살펴보면, 각 모형별로 데이터마트 구성비가 원시자료 구성비에서 멀어질수록 정확도와 특이도는 낮아지고 민감도는 높아짐을 알 수 있다. 각 분류기준값에 대하여 0.1인 경우에는 1 : 8 또는 1 : 10, 0.2인 경우에는 1 : 4, 0.3인 경우는 1 : 2, 0.4와 0.5인 경우에는 1 : 1 비율이 특이도와 민감도가 유사한 지점으로 판단된다.

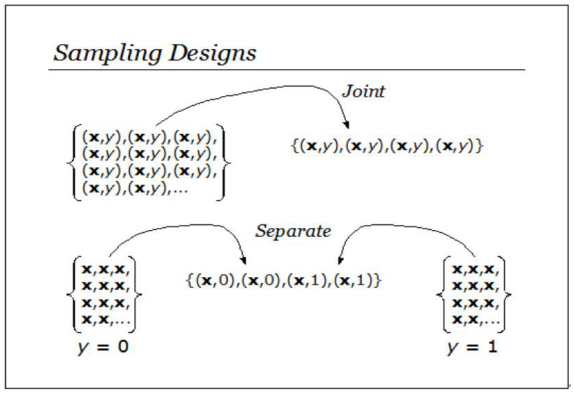

그림 **3.1.** 표본추출의 설계

3. 과대표본추출에 대한 사후확률의 조정

3.1. 과대표본추출(Oversampling) 및 효과

아래의 그림 3.1과 같이 표본추출을 설계할 때 결합 표본추출(joint sampling)은 설명변수-목표변수(input-target) 쌍이 결합 분포로부터 랜덤하게 표본추출 된다. 반면, 분할 표본추출(separate sampling)은 목표 집단 각 범주의 분포로부터 각각 독립적으로 표본추출 된다.

분할 표본추출은 관심을 갖는 사건이 희소할 경우, 희귀사건을 과대표본추출(oversampling)하는 것이 일반적이다 (Scott과 Wild, 1986). 희귀사건 데이터를 단순임의추출하면 그 사건의 수가 다른 집단에 비해서 상대적으로 더 적어지기 때문에 해당 사건을 파악하기가 더욱 어려워진다. 이 경우 분석용 데이터마트는 미반응 값으로 대부분 구성되기 때문에 반응과 미반응을 판별해 주는 모형 구축을 위해 필요한 데이터의 구성비가 한쪽으로 편향(bias)되는 문제가 발생하게 된다. 따라서 이런 상황에서 표본추출은 추가적으로 반응 표본을 충분히 가질 수 있도록 미반응과 비교해서 상대적으로 반응에게 더 많은 비중을 부여하는 표본추출과정이 필요하다.

이와 같이 분류집단들이 매우 불균등하게 나타날 때 상대적으로 사건의 빈도가 희귀한 집단의 사례들을 과대표본추출 한다. 사건의 빈도가 적어질수록 그 사건은 점점 더 관심이 증대되거나 중요해지는 경우가 있는데, 예를 들어 광고물에 반응하는 사람, 부정거래를 하는 사람, 채무 파산자 등이 이에 해당한다. 이러한 분류문제에서 매우 낮은 반응률에 직면하게 될 때 실제 전문가들은 대부분 상대적으로 효과적이면서 편리한 접근방법으로서 반응과 미반응의 비율을 균등하게 표본추출한다. 어떤 접근방법을 사용하든지 간에 구축 모형을 평가하고 사후확률을 추정/예측할 때는 과대표본추출에 따른 편향을 조정(adjustment)할 필요가 있다 (Galit 등, 2006).

분할 표본추출에서 모수의 최우추정량은 식 (3.1)과 같이 의사(擬似)모형(pseudo model)을 적합함으로써 결정된다 (Scott과 Wild, 1997).

$$\text{logit}\left(p_i^*\right) = \ln\left(\frac{\rho_1\pi_0}{\rho_0\pi_1}\right) + \beta_0 + \beta_1\chi_{1i} + \cdots + \beta_k\chi_{ki} \tag{3.1}$$

π_0와 π_1은 모집단에서 0과 1의 비율이며, ρ_0와 ρ_1은 표본에서 0과 1의 비율이다. 식 (3.1)에서 p_i^*은 사전확률 p_i에 대한 편향된 표본에 대응하는 사후확률의 추정값이다. 결국 과대표본추출의 효과는 상수 오프셋(offset) 만큼 로짓(logit)을 이동시키게 되는 것이다.

표 **3.1.** 사후확률 조정에 대한 정확도 비교

	Logistic Regression							
	1 : 1	1 : 2	1 : 4	1 : 6	1 : 8	1 : 10	1 : 15	1 : 20
Offset	0.685	0.686	0.687	0.685	0.681	0.683	0.680	0.679
Sampling Weights	0.566	0.567	0.570	0.568	0.563	0.567	0.565	0.566

3.2. Offset 방법 및 가중치 방법

의사모형은 오프셋(offset)을 모형에 반영함으로써 직접 적합할 수 있게 한다. 오프셋은 기본 모형이 적합된 후에 적용될 수 있다. 예측값으로 부터 오프셋을 빼는 것과 사후확률을 구하는 것은 식 (3.2)와 같다. 식 (3.2)에서 $\hat{p}_i^*$은 조정되지 않은 사후확률의 추정값이다.

$$\hat{p}_i = \frac{\hat{p}_i \rho_0 \pi_1}{(1-\hat{p}_i^*)\rho_1 \pi_0 + \hat{p}_i^* \rho_0 \pi_1}. \tag{3.2}$$

과대표본추출을 조정하는 또 다른 방법은 표본추출 가중치(sampling weights)를 반영하는 것이다. 표본추출 가중치는 표본이 실제 모집단을 잘 대표할 수 있도록 조정한다. 희귀사건이 과대표본추출 되었을 때 0은 표본에서 실제보다 적게 출현되지만 결과적으로 0의 경우는 1보다 분석표본으로 더 많이 추출되게 된다. 예측된 값은 선택 확률(각 범주에 대해서, (표본 내 사례수)/(모집단 내 사례수))에 반비례하는 가중치에 의해 조정된다. 표준화된 표본 가중치를 사용하는 것은 원래의 표본수를 합하기 때문에 편리하다.

$$\text{weight}_i = \begin{cases} \dfrac{\pi_1}{\rho_1}, & \text{만약 } y_i = 1, \\ \dfrac{\pi_0}{\rho_0}, & \text{만약 } y_i = 0, \end{cases}$$

$$\sum_{i=1}^{n} \text{weight}_i = n_0 \frac{\pi_0}{\rho_0} + n_1 \frac{\pi_1}{\rho_1} = n\pi_0 + n\pi_1 = n, \tag{3.3}$$

여기서 n_0, n_1은 각각 범주 0과 1의 빈도이다. 가중치(weight)는 $n\pi_0$와 $n\pi_1$의 표본에 대한 사례수를 조정하고, 조정된 표본의 비율은 모집단의 비율과 같다. 오프셋 방법과 가중치 방법에서 모수의 추정치는 정확하게 일치하지 않지만 대표본하에서는 통계적으로 일치한다 (Scott과 Wild, 1986).

3.3. 사례분석

과대표본추출에 대한 사후확률을 오프셋 방법과 가중치 방법으로 각각 조정하고, 로지스틱 회귀모형을 적용하여 데이터마트 구성비에 따른 정확도를 비교하였다. 이때 분리 기준값은 0.5이다. 사례분석을 위해 사용된 자료는 2절에서 사용된 자료와 동일하며, 2008년 8월, 9월, 10월, 11월 자료는 분석용 자료로 사용하였고, 2008년 12월, 2009년 1월 자료는 검증용 자료로 사용하였다. 표 3.1은 분석용 자료에 대한 결과이다.

표 3.1의 결과를 보면 오프셋 방법으로 조정된 사후확률의 정확도가 가중치 방법으로 조정된 사후확률의 정확도 보다 높음을 알 수 있다. 이와 같은 사례분석에서 볼 수 있듯이 과대표본추출의 문제에서 사후확률의 조정 절차가 필요하다고 하겠다. 일반적으로 선형모형에서는 오프셋 방법이 우월한 결과를 보이는 것으로 알려져 있으며, 반대로 비선형모형의 경우 가중치 분석이 수월성을 보인다 (Scott과 Wild, 1986).

4. 결론 및 토의

본 연구에서는 반응/미반응 목표변수를 갖는 모집단에서 모형 구축을 위한 데이터마트를 형성할 때 데이터마트의 반응/미반응 구성비는 구축된 모형의 성능에 영향을 준다는 점에 착안하여 모형 성능 비교 실험을 하였다. 그 결과 분류기준값을 기준으로 데이터마트 구성비가 원시자료에서 1 : 1 표본추출에 가깝게 갈수록 정확도와 특이도는 떨어지고 민감도는 높아지는 양상을 보이고 있었다. 또한, 분류기준값 0.10일 때 1 : 8 표본추출, 0.20일 때 1 : 4 표본추출, 0.30일 때 1 : 2 표본추출, 0.40일 때 1 : 1.5 표본추출, 0.50일 때 1 : 1 표본추출이 추천되어졌다.

분류기준값을 선정하기 위해서 일반적으로 목표변수의 범주가 두 집단일 경우에는 분류기준값을 0.50으로 삼는 것이 보통이기는 하나 이는 절대적인 기준이 될 수는 없고 분석 자료의 성격에 많이 의존한다고 할 수 있다. 예컨대, 우/불량 고객 식별을 위한 신용평가 모형개발에서 우량고객을 '0', 불량고객을 '1'로 정의할 때 보수적인 관리방법을 채택하는 곳에서는 분류기준값을 '0'에 가까운 수로 정하여 우량고객을 상대적으로 강하게 제한할 수 있다 (이태림 등, 2004). 따라서 최적의 분류기준값 선정은 위해서는 모형개발의 목적과 관리방법에 의해 개발자의 주관적인 판단으로 이루어져야 하겠다. 또한 과대표본추출을 할 경우에는 추정된 사후확률에 대한 조정이 필요할 것이다.

참고문헌

강현철, 한상태, 최종후, 이성건, 김은석, 엄익현, 김미경 (2006). <고객관계관리(CRM)를 위한 데이터마이닝 방법론>, 자유아카데미.

이태림, 구자용, 박헌진, 이긍희, 최대우 (2004). <데이터마이닝>, 한국방송통신대학교출판부.

장남식, 홍성완, 장재호 (1999). <데이터 마이닝>, 대청미디어.

Galit, S., Nitin, R. P. and Peter, C. B. (2006). *Data Mining for Business Intelligence*, John Wiley & Sons, New York.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic regression models under case-control or choice based sampling, *Journal of the Royal Statistical Society B*, **48**, 170–182.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood, *Biometrika*, **84**, 57–71.

A Study on the Adjustment of Posterior Probability for Oversampling when the Target is Rare

U. N. Kim[1] · S. K. Lee[2] · J. H. Choi[3]

[1]BC Card; [2]Department of Statistics, Sungshin Women's University
[3]Department of Information & Statistics, Korea University

(Received March 2011; accepted June 2011)

Abstract

When an event of target variable is rare, a widespread strategy is to build a model on the sample that disproportionally over-represents the events, that is over-sampled. Using the data over-sampled from the original data set, the predicted values would be biased; however, it can be easily corrected to represent the population. In this study, we investigate into the relationship between the proportion of rare event on a data-mart and the model performance using real world data of a Korean credit card company. Also, we use the methods for adjusting of posterior probability for over-sampled data of the offset method and the weighted method. Finally, we compare the performance of the methods using real data sets.

Keywords: Over-sampling, adjusting of posterior probability, rare event offset method, weighted method.

but

[3]Corresponding author: Professor, Department of Information & Statistics, Korea University, Jochiwon-eup, Yeongi-gun, Chungnam 339-700, Korea. E-mail: jhchoi@korea.ac.kr