Journal of the Korean Data Analysis Society (April 2020), 22(2), 565-574.
https://doi.org/10.37727/jkdas.2020.22.2.565

불균형 자료의 분류분석을 위한 가우스 혼합 군집모형을 이용한 샘플링 기법*

방성완[1], 김재오[2]

요 약

일반적으로 전통적인 분류분석 방법은 소수집단의 개체수가 다수집단의 개체수와 현격한 차이를 보이는 이항 범주형 불균형 자료(imbalanced data)의 분류분석에서 문제를 야기한다. 그것은 다수집단에 편향된 분류함수를 추정하므로서 대부분의 자료를 다수집단으로 분류하여 소수집단의 분류 정확도를 현저히 감소하는 현상이다. 이러한 문제를 효과적으로 해결하기 위하여 본 논문에서는 가우스 혼합 군집모형을 활용하여 불균형 자료의 분류분석을 위한 샘플링 기법을 제안한다. 이 방법은 소수집단에 대해 가우스 혼합분포를 추정하고 이를 기반으로 과대 추출하는 것이 핵심이다. 제안하는 방법을 SMOTE(synthetic minority over-sampling technique), ADASYN (adaptive synthetic sampling)과 같은 기존의 과대 추출 방법들과 다양한 상황 및 실제 예제에서 비교하여 그 우수성을 확인하였다. 특히, 불균형 자료 분석에서 중요하게 다루어지는 소수집단의 분류 정확도 측면에서 제안한 방법은 충분히 좋은 성능을 보였다. 본 연구에서는 이진 분류기로서 서포트 벡터 머신을 분류방법으로 사용하였으며, 전체 정확도, 민감도, 특이도 및 기하평균으로 성능을 평가하였다.

주요용어 : 가우스 혼합 군집모형, 과대추출, 불균형 자료, 샘플링 기법.

1. 서론

범주형 자료의 분류(classification)는 개체들의 소속집단을 예측하는 분류함수(classifier)를 추정하는 데이터 마이닝의 과정을 의미한다. 현실 세계의 많은 문제들이 범주형 자료를 분류하고자 하는 것을 고려할 때 이는 다양한 분야에서 활용되고 있다(Kwon, Kwon, 2018; Choi, 2019). 예를 들어 금융기관에서 대출심사를 하며 고객을 '안전', '불안전'으로 분류한다던지, 병원에서 특정 질병에 대해 '양성'(positive), '음성'(negative)으로 분류하는 것 등이 해당될 수 있다. 이러한 이항 범주형 자료 분석의 경우 많은 현실 문제에서 다수 및 소수집단(class)의 개체 수가 현격한 차이를 보이는 불균형 자료(imbalanced data)를 다루게 된다. 이상금융거래탐지(fraud detection), 희귀 질병의 검진(medical diagnosis of rare diseases), 이동전화사업에 있어 이탈탐지(churn) 등은 대표적인 불균형 자료에 대한 현실 문제의 사례이며 이러한 자료의 형태에 관한 분석요구는 지속적으로 증대되고 있으며, 기계학습(machine learning)분야에서 'learning from imbalanced data'라는 용어로 활발하게 연구되고 있다(Tang et al., 2009; Wang et al., 2017; Singh, Dhall, 2018). 다수집단(majority class)을 잘못 분류하는 경우보다 소수집단(minority class)을 잘못 분류하는 것은 불균형 자료 분석에 있어 더 높

*본 논문은 육군사관학교 화랑대연구소의 2020년도(군학-1) 연구활동비 지원을 받아 연구되었음.
[1]01805 서울시 노원구 화랑로 574, 육군사관학교 수학과 교수. E-mail : wan1365@gmail.com
[2](교신저자) 32800 충남 계룡시 신도안면 계룡대로663 사서함 501-8, 육군본부 빅데이터분석센터 분석장교.
E-mail : c14180@gmail.com
[접수 2020년 3월 20일; 수정 2020년 4월 17일; 게재확정 2020년 4월 20일]

은 비용 또는 위험을 수반한다. 이러한 이유로 인해 소수집단의 분류 정확도를 향상하는 것은 불균형 자료분석의 핵심이다. 그럼에도 불구하고 대부분의 전통적인 분류분석 방법은 모형의 전체 정확도를 향상시키기 위하여 분류함수가 다수집단을 편향적으로 추정하도록 유도하여 소수집단의 분류 정확도를 감소시키는 문제가 있다. 불균형 자료에 관한 분류분석에 있어 소수집단을 더 정확하게 분류하기 위한 대표적인 방법에는 소수집단을 더 추출하는 과대추출(over-sampling) 방법과 다수집단을 적게 추출하는 과소추출(under-sampling) 방법이 있다. 이것은 인위적으로 소수집단과 다수집단의 개체수를 균형되도록 조정하는 샘플링 기법으로 볼 수 있다. 또한 소수집단을 잘못 분류했을 때 가중치를 활용하여 비용을 증가시키는 오분류 비용을 차등하여 적용하는 방법이 있다. Veropoulos et al.(1999), Lin et al.(2002), Akabani et al.(2004), Ling, Sheng(2007), Tang et al.(2009), Anand et al.(2010), 그리고 Datta, Das(2015)는 가중치를 이용하여 오분류 비용을 차등 적용하는 방법에 대하여 연구하였으며, Kubat, Matwin(1997), Japkowicz (2000), Chawla et al.(2002), Tang et al.(2009)은 소수집단의 개체수를 증가하기 위한 과대추출 또는 다수집단의 개체수를 감소하는 과소추출을 시행하여 집단간 개체수의 균형을 맞추는 샘플링 방법을 연구하였다. 특히 Chawla et al.(2002)이 제안한 SMOTE(synthetic minority over-sampling technique)는 불균형 자료의 분류분석을 위해 제안된 과대추출 기법중 하나이다.

SMOTE는 일반적인 임의 과대추출(random over-sampling)에서 야기되는 과적합 문제를 극복하기 위하여 보간법(interpolation)을 활용하여 소수집단의 개체 간 새로운 개체를 생성하는 방법이다. SMOTE의 가상 개체 생성의 아이디어는 여러 다양한 방법으로 확장되어 연구되었다. Han et al.(2005)은 Borderline‐SMOTE로, He et al.(2008)은 ADASYN(adaptive synthetic sampling)으로, Bunkhumpornpat et al.(2009)은 Safe-level-SMOTE로, Bunkhumpornpat et al.(2012)은 DB(density-based) SMOTE로 기존의 SMOTE 알고리즘을 개선하였다. 그러나 SMOTE는 소수집단의 개체와 개체 사이의 직선상에서만 새로운 개체를 생성함으로 분류분석에서 소수집단의 영역을 확장하는 데는 제한적이다. 이러한 제한점을 보완하기 위하여 본 논문에서는 가우스 혼합 군집모형을 이용하여 소수집단의 혼합분포를 추정하고 이를 이용하여 새로운 개체를 과대 추출하는 샘플링 기법인 GMC-SMOTE(Gaussian mixture clustering-SMOTE)를 제안하였다.

본 논문의 구성은 2절에서 불균형 자료의 분류분석을 위한 다양한 방법론과 분류함수의 성능을 평가하는 척도에 대하여 간략히 소개하고, 3절에서 가우스 혼합 군집모형을 이용하여 소수집단의 가상 개체를 임의 생성하는 GMC-SMOTE를 제안하였다. 4절에서는 임의 생성 자료와 실제자료를 활용한 모의실험을 통해 기존의 과대추출 방법과 제안한 GMC-SMOTE의 성능을 비교하였고, 마지막으로 5절에서는 결론 및 향후 연구를 제시하였다.

2. 불균형 자료의 분류분석을 위한 다양한 방법론

불균형 자료의 분류분석을 위하여 전통적인 분류분석 기법을 그대로 적용하게 되면 다수집단의 많은 개체수로 인하여 편향된 추정을 유도하게 됨에 따라 소수집단을 정확하게 분류하지 못하는 현상이 발생할 수 있다(Owen, 2007; Oommen et al., 2011; Park, Bang, 2015). 본 절에서는 이항 범주형 불균형 자료의 분류분석에서 분류함수의 균형된 추정을 위하여 많이 이용되는 방법론과 추정된 분류함수의 성능을 평가하는 척도에 대하여 간략히 소개하기로 한다.

2.1. 과대추출(over-sampling) 기법

과대추출 방법 중 가장 간단한 방법은 임의 과대추출 방법(random over-sampling)으로 난수를 발

생하여 개체를 임의로 반복해서 추출하는 방법이다. 불균형 자료에 관한 분류분석에서 임의 과대추출 방법은 소수집단을 다수집단의 개체수만큼 임의로 추출하는 방법이다. 이를 통하여 소수집단을 더 정확하게 분류할 수 있다(Chawla et al., 2003; Jeong et al., 2008). 그러나 임의 과대추출 방법은 소수집단의 개체에 관하여 반복 추출함으로서 새로운 자료를 생성하지 못하게 되며 이는 과적합의 문제를 야기할 수 있다. SMOTE(Chawla et al., 2002)는 소수집단의 각각의 개체에 대하여 근접한 k개의 개체를 선택하고, 그 개체들 사이에서 새로운 가상(synthetic)의 개체들을 임으로 생성하는 방법으로, 임의 과대추출 방법의 과적합 문제를 해결하기 위해 개발되었다. SMOTE에서 소수집단에 대해 개체를 생성하는 과정은 다음과 같다. 먼저, 임의로 선택된 소수집단의 일부 개체에 대해 최근접한 k개 개체와의 각각의 거리를 측정한다. 다음은 무작위 난수(0과 1 사이)를 발생하여 각각의 개체와의 거리와 곱하고, 선택된 개체에 더해 새로운 개체를 생성하게 된다. 따라서 새롭게 생성된 개체는 두 개체 사이의 임의의 지점이 되고, 이렇게 생성된 새로운 개체는 과대적합을 피하고 균형된 분류함수를 추정하는 장점이 있다.

2.2. 과소추출(under-sampling) 기법

불균형 자료의 분류분석에서 과소추출은 과대추출과 반대로 다수집단의 자료를 소수집단의 개체 수만큼 선택하여 선택된 개체만 분류함수의 추정에 사용하는 방법이다. 먼저 임의 과소추출 방법(random under-sampling)은 다수집단과 소수집단의 개체수 비율을 유사하게 맞춰 소수집단에 대해 더 정확하게 분류할 수 있다. 이러한 과소추출 방법은 전체 개체수가 많아지는 과대추출에 대비하여 모형의 적합속도가 빠른 것이 장점이다. 그러나 다수집단의 개체를 임의로 제거하면서 분류함수 추정에 중요한 영향을 미치는 개체가 제거될 수 있으며, 다수의 개체가 제거됨에 따라 훈련 개체 수가 부족하게 되어 전체적으로 정보 손실이 발생할 수 있는 단점을 내재한다(Drummond, Holte, 2003). K-평균 군집 과소추출 방법은 K-평균 군집모형(K-means clustering)을 기반으로 실시하는 과소추출 기법으로 다수집단의 개체를 이용하여 K개의 군집을 만들고, 각 군집의 평균점을 새로운 훈련개체로 사용한다(Zhang et al., 2010; Bang, Jhun, 2014). 이때 군집수 K를 조정하여 다수집단과 소수집단의 비율을 조정하여 분류의 결과가 다수집단으로 편향되지 않도록 하며, 군집의 평균점을 훈련 개체로 활용하기 때문에 임의 과소추출에 비해 중요 정보의 손실위험을 줄일 수 있다.

2.3. 오분류 비용의 차등적용 기법

불균형 자료의 분류분석에서 균형된 분류함수를 추정하기 위한 또 다른 방법은 모형 적합에 사용되는 손실함수(loss function)에서 집단별 오분류에 대한 비용을 차등적으로 적용하는 것이다(Veropoulos et al., 1999; Lin et al., 2002; Akabani et al., 2004; Ling, Sheng, 2007; Tang et al., 2009; Anand et al., 2010; Datta, Das, 2015). 따라서 집단별 분포에 의존하지 않고 소수집단의 개체에 대한 오분류 비용을 상대적으로 상향 적용하여 균형된 분류함수를 추정할 수 있다.

2.4. 성능 평가(performance measures)

분류분석에서 추정된 분류함수의 성능은 새로운 개체를 실제 집단으로 잘 분류할 수 있는지를 평가하는 분류 정확도로 나타낼 수 있으며, Table 1의 오차행렬(confusion matrix)을 이용하여 다양한 분류 정확도를 표현할 수 있다. Table 1의 오차행렬에서 소수집단의 개체를 정확하게 소수집단의 개체로 예측한 개체수를 나타내는 TP와 다수집단의 개체를 정확하게 다수집단의 개체로 예측

한 개체수인 TN이 있으며, 실제 소수집단의 개체를 다수집단으로 부정확하게 분류한 개체수를 의미하는 FN과 실제 다수집단의 개체임에도 불구하고 소수집단으로 분류한 개체수인 FP가 있다.

Table 1. Confusion matrix

	'+' prediction	'-' prediction
'+' class	TP (True '+')	FN (False '-')
'-' class	FP (False '+')	TN (True '-')

전체정확도(overall accuracy)는 일반적인 분류분석에서 가장 많이 사용되는 평가 척도로서

$$Overall\ accuracy = \frac{TP+TN}{TP+TN+FN+FP}$$

와 같이 정의되고, 모든 새로운 개체에 대하여 소속집단이 올바르게 분류된 개체의 비율을 나타낸다. 전체정확도는 집단별 개체수가 균형적인 자료의 분류분석에서 가장 일반적으로 사용되는 평가척도이지만, 집단 별 자료의 수가 현저하게 차이가 나는 불균형 자료의 경우에는 다수집단 자료의 영향으로 분류함수에 대한 성능 평가가 왜곡될 가능성이 높다. 특히 집단별 개체수가 현저히 차이가 나는 경우에는 모든 개체를 다수집단으로 분류한다고 하더라도 오분류되는 소수집단의 개체수가 상대적으로 매우 작기 때문에 전체정확도는 여전히 높게 평가된다. 따라서 전체정확도만으로는 추정된 분류함수를 평가하기에는 한계가 있으며, 이러한 경우에는 각 집단 별로 분류 정확도를 각각 평가하는 방법이 대안이 될 수 있다. 소수집단의 분류 정확도를 의미하는 민감도(sensitivity)와 다수집단의 분류 정확도를 의미하는 특이도(specificity)는 각각

$$Sensitivity = \frac{TP}{TP+FN},\ Specificity = \frac{TN}{TN+FP}$$

로 정의되며, 민감도와 특이도 두 가지를 모두 고려할 수 있는 기하평균(g-mean)은

$$G-mean = \sqrt{Sensitivity \times Specificity}$$

와 같고, 불균형 자료의 분류분석에서 분류함수의 평가 척도로 많이 사용되고 있다(Kubat, Matwin, 1997; Wang et al., 2017; Singh, Dhall, 2018). 본 논문에서는 제안한 방법과 기존 방법의 성능을 비교하여 평가하기 위하여 전체 정확도(overall accuracy), 민감도(sensitivity), 특이도(specificity) 및 기하평균(G-mean)을 활용한다.

3. 가우스 혼합 군집모형을 이용한 과대추출 기법

일반적으로 불균형 자료의 분류분석에서는 소수집단의 분류 정확도가 더 중요하게 다루어진다. SMOTE는 소수집단의 가상 개체를 생성하여 균형된 분류함수를 추정하는 이점이 있으나, 기존의 소수집단 개체와 개체 사이의 직선상에서만 새로운 개체를 생성함으로 소수집단의 영역을 확장하는 데는 제한적이다. 따라서 본 논문에서는 가우스 혼합 군집모형(Gaussian mixed clustering)을 이용하여 소수집단을 K개로 군집화 하고, 추정된 각각의 군집모형으로부터 소수집단의 새로운 개체를 임의로 추출하는 GMC-SMOTE(Gaussian mixed clustering based SMOTE) 샘플링 기법을 제안하고자 한다.

가우스 혼합 군집모형은 혼합분포에 근거한 군집모형의 특수한 경우로 각각의 군집분포를 다변

량 정규분포로 가정한다. 즉, 표본 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$는 가우스 혼합분포 $f(\boldsymbol{x})$

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \phi_k(\boldsymbol{x} | \boldsymbol{\mu}_k, \Sigma_k), \text{(여기서 } \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1)$$

로부터 얻은 크기가 n인 확률표본(random sample)으로 k번째 군집의 확률분포 $\phi_k(\boldsymbol{x})(k=1,2,\cdots,K)$는 평균 벡터 $\boldsymbol{\mu}_k$와 공분산 행렬 Σ_k를 모수로 갖는 다변량 정규분포 $N(\boldsymbol{\mu}_k, \Sigma_k)$로 가정한다. 따라서 확률표본 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$으로부터 로그-가능도 함수는

$$l(\pi_k, \boldsymbol{\mu}_k, \Sigma_k | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{\mu}_k, \Sigma_k) \right)$$

와 같고 일반적으로 모수 π_k, $\boldsymbol{\mu}_k$, Σ_k는 EM 알고리즘(Dempster et al., 1977)을 통해 추정된다.

본 논문에서는 소수집단의 분포를 가우스 혼합분포로 가정하고 주어진 N^+개의 소수집단의 훈련자료 $\left\{\boldsymbol{x}_1^+, \boldsymbol{x}_2^+, \cdots, \boldsymbol{x}_{N_+}^+\right\}$로부터 K개의 군집으로 구성된 혼합분포를 추정한 후, 이로부터 상대적으로 부족한 소수집단의 개체를 임의 생성하는 GMC-SMOTE를 Algorithm 1과 같이 제안하고자 한다.

Algorithm 1. GMC-SMOTE algorithm

단계 1.	크기 $N(=N^+ + N^-)$의 훈련자료를 N^+개의 소수집단 자료 $\left\{\boldsymbol{x}_1^+, \boldsymbol{x}_2^+, \cdots, \boldsymbol{x}_{N_+}^+\right\}$와 N^-개의 다수집단 자료 $\left\{\boldsymbol{x}_1^-, \boldsymbol{x}_2^-, \cdots, \boldsymbol{x}_{N_-}^-\right\}$로 구분한다.
단계 2.	N^+개의 소수집단 자료 $\left\{\boldsymbol{x}_1^+, \boldsymbol{x}_2^+, \cdots, \boldsymbol{x}_{N_+}^+\right\}$에 근거하여 가우스 혼합 군집모형의 분포에 대한 모수의 추정값 $\hat{\pi}_k$, $\hat{\boldsymbol{\mu}}_k$, $\hat{\Sigma}_k(k=1,2,\cdots,K)$을 계산한다.
단계 3.	단계 2에서 추정한 혼합분포 $\hat{f}(\boldsymbol{x}) = \sum_{k=1}^{K} \hat{\pi}_k \phi_k(\boldsymbol{x} | \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k)$로부터 새로운 n^+개의 소수집단 자료를 생성하고, 이를 이용하여 $\widetilde{N}^+ = N^+ + n^+$개의 소수집단 자료를 형성한다.
단계 4.	N^-개의 다수집단 자료와 단계 3에서 형성한 $\widetilde{N}^+$개의 소수집단 자료를 이용하여 균형된 분류함수를 추정한다.

4. 모의실험

본 논문에서 제안한 GMC-SMOTE의 성능을 평가하기 위하여 불균형 자료의 분류분석에 대한 모의실험을 진행하였으며, 제안된 GMC-SMOTE의 성능 비교를 위해 과대 샘플링 기법으로 널리 사용되고 있는 SMOTE(Chawla et al., 2002)와 ADASYN(He et al., 2008)를 모의실험에 포함하였다. 모의실험 1에서는 가우스 혼합모형으로 임의 생성된 Checkboard 자료를 이용하였으며, 모의실험 2에서는 Table 2에서 보는 바와 같이 UCI machine learning repository(Dua, Graff, 2019)에서 제공되는 2개의 실제자료를 이용하였다. Table 2의 Pima 자료는 입력변수 중 결측치가 많은 insulin 변수를 제거한 후 결측치가 있는 개체를 제거하였다.

Table 2. The description of the data sets

The names of data set	Number of instances			Number of input variables	Percentage of minority class
	Total	Majority class	Minority class		
Checkboard	1000	900, 800, 700	100, 200, 300	2	10%, 20%, 30%
Pima	532	355	177	7	33.3%
Haberman	306	225	81	3	26.5%

각각의 불균형 자료에 대하여 과대 샘플링 기법을 적용한 후에는 support vector machine(Cortes, Vapnik, 1995; Vapnik, 1998)을 이용하여 분류분석을 시행하였다. SVM은 대표적인 기계학습의 지도학습 방법으로 이진형 자료분류에 있어 파라미터의 조율을 통하여 일반적으로 높은 정확도를 갖는다. 분류 정확도를 비교 평가하기 위한 지표로 Park, Bang(2015)에서 사용한 전체정확도(overall accuracy), 민감도(sensitivity), 특이도(specificity) 그리고 기하평균(G-mean)을 사용하였다. 일반적으로는 전체 정확도를 모형의 성능을 평가하기 위해 사용하지만, 불균형 자료 분석에 있어서는 소수집단에 관한 예측력을 평가할 수 있는 민감도와 기하평균을 사용하는 것이 더 타당하다. 소수집단의 개체를 과대추출하기 위하여 SMOTE, ADASYN, GMC-SMOTE를 적용하였으며, SMOTE와 ADASYN은 R에서 제공하는 “smotefamily” 패키지(Siriseriwan, 2016)를 이용하였다. 또한 GMC-SMOTE의 적용시 가우스 혼한 군집모형은 R에서 제공하는 “mclust” 패키지(Scrucca et al., 2016)를 이용하였으며, 이때 BIC(Bayesian information criteria)(Schwartz, 1978)을 이용하여 군집 수 K를 결정하였다.

4.1. **모의실험** 1 : Checkboard synthetic data

모의실험 1에서는 다수집단(majority class)과 소수집단(minority class)의 개체를 각각 가우스 혼합모형 $f_1(\boldsymbol{x})$과 $f_2(\boldsymbol{x})$

$$f_1(\boldsymbol{x}) = \frac{1}{6}\sum_{k=1}^{6} N_2\left(\boldsymbol{\mu}_{1(k)}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right) \text{ and } f_2(\boldsymbol{x}) = \frac{1}{6}\sum_{k=1}^{6} N_2\left(\boldsymbol{\mu}_{2(k)}, \begin{bmatrix} 2.5 & 0 \\ 0 & 2.5 \end{bmatrix}\right) \tag{4.1}$$

로부터 생성하였다. 이때 $\boldsymbol{\mu}_{1(1)} = (5,5)'$, $\boldsymbol{\mu}_{1(2)} = (15,5)'$, $\boldsymbol{\mu}_{1(3)} = (10,10)'$, $\boldsymbol{\mu}_{1(4)} = (20,10)'$, $\boldsymbol{\mu}_{1(5)} = (5,15)'$, $\boldsymbol{\mu}_{1(6)} = (15,15)'$와 $\boldsymbol{\mu}_{2(1)} = (10,5)'$, $\boldsymbol{\mu}_{2(2)} = (20,5)'$, $\boldsymbol{\mu}_{2(3)} = (5,10)'$, $\boldsymbol{\mu}_{2(4)} = (15,10)'$, $\boldsymbol{\mu}_{2(5)} = (10,15)'$, $\boldsymbol{\mu}_{2(6)} = (20,15)'$을 이용하여 Checkboard 자료를 생성하였다.

모형적합을 위해 크기가 1,000인 훈련자료(training data)를 생성하였으며, 이때 소수집단의 비율을 10%(소수집단 100개, 다수집단 900개), 20%(소수집단 200개, 다수집단 800개), 30%(소수집단 200개, 다수집단 800개)로 하여 불균형의 정도를 달리하였다. 또한, SVM의 조율모수(tuning parameter)를 선택하기 위해 크기가 1,000(소수집단 500개, 다수집단 500개)인 검정자료(validation data)와 샘플링 기법들에 대한 SVM 분류함수의 정확도를 평가하기 위해 크기가 20,000(소수집단 10,000개, 다수집단 10,000개)인 평가자료(test data)를 각각 독립적으로 생성하였다. 이렇게 생성된 각각의 모의실험 자료에서 소수집단의 과대추출 비율을 100%로 하였으며, 독립적으로 100회 반복 시행하여 SVM, SMOTE, ADASYN과 본 연구에서 제안하는 GMC-SMOTE를 4가지 척도의 평균으로 성능을 평가하였다. Table 3에서 보듯이 다음과 같은 의미있는 해석이 가능하다. 먼저 소수집단의 분류 정확도를 알 수 있는 민감도의 경우 일반적인 SVM보다는 과대추출 기법을 기반으로 하는 방법이 월등히 좋은 성능을 나타내며 특히, 본 논문에서 제안하는 GMC-SMOTE는 가장 우수한 성능을 보였다. 둘째, 다수집단의 분류 정확도 기준이 될 수 있는 특이도는 일반적인 SVM이 가장 좋지만, GMC-SMOTE는 다른 과대추출 기법들과 큰 차이를 보이지 않았다. 마지막으로 전체 정확도와 민감도 및 특이도를 모두 고려하는 기하평균의 경우 GMC-SMOTE가 가장 우수한 성능을 보였으며, 모든 방법은 소수집단의 비율이 증가할수록 정확도가 향상되는 경향을 확인하였다.

4.2. **모의실험** 2 : Real data sets

본 논문에서 제안하는 GMC-SMOTE의 유용성을 확인하기 위하여 모의실험 2에서는 Table 2의

실제자료를 분석하였다. 각각의 자료에 대하여 모형의 적합 및 평가를 위해 자료를 2:1의 비율로 훈련자료(training data)와 평가자료(test data)로 임의로 나누었으며, SVM의 조율모수는 훈련자료를 이용한 10-겹 교차타당법(10-fold cross validation)으로 선택하였다. 샘플링 기법들의 성능 평가를 위해 전체정확도, 민감도, 특이도. 그리고 기하평균을 계산하였으며, 이러한 과정을 100번 독립적으로 반복하여 평균값을 계산하였다.

Table 4, 5는 각각 Table 2의 Pima와 Haberman 자료에 대한 결과이다. 먼저, 과대추출을 하지 않는 일반적인 SVM은 특이도는 높지만 소수집단의 분류 정확도인 민감도가 낮음을 알 수 있다. 둘째, 본 논문에서 제안하는 GMC-SMOTE는 기존의 SMOTE 및 ADASYN에 비해 민감도 뿐 아니라

Table 3. Simulation results for the Checkboard synthetic data set in Table 2

Percentage of minority class	Method	Test classification accuracy			
		Overall accuracy	Sensitivity	Specificity	G-mean
10%	Standard SVM	0.743 (0.002)	0.511 (0.005)	0.976 (<0.001)	0.705 (0.004)
	SMOTE	0.786 (0.002)	0.661 (0.004)	0.881 (0.001)	0.775 (0.002)
	ADASYN	0.792 (0.002)	0.696 (0.003)	0.887 (0.001)	0.786 (0.002)
	GMC-SMOTE	0.858 (<0.001)	0.848 (0.002)	0.867 (<0.001)	0.857 (0.001)
20%	Standard SVM	0.818 (0.004)	0.682 (0.003)	0.953 (<0.001)	0.807 (0.001)
	SMOTE	0.824 (0.002)	0.762 (0.004)	0.885 (0.001)	0.821 (0.002)
	ADASYN	0.811 (0.002)	0.766 (0.003)	0.855 (0.001)	0.810 (0.002)
	GMC-SMOTE	0.865 (0.001)	0.858 (0.002)	0.873 (0.003)	0.865 (0.001)
30%	Standard SVM	0.844 (0.001)	0.753 (0.002)	0.934 (<0.001)	0.839 (0.001)
	SMOTE	0.848 (0.002)	0.808 (0.002)	0.887 (0.002)	0.847 (0.002)
	ADASYN	0.825 (0.001)	0.817 (0.002)	0.833 (0.002)	0.825 (0.002)
	GMC-SMOTE	0.878 (<0.001)	0.861 (0.001)	0.875 (0.002)	0.878 (0.001)

Note) the numbers in parentheses are standard errors.

Table 4. Simulation results for the Pima data set in Table 2

Percentage of over sampling	Method	Test classification accuracy			
		Overall accuracy	Sensitivity	Specificity	G-mean
0%	Standard SVM	0.776 (0.003)	0.551 (0.006)	0.890 (0.004)	0.700 (0.004)
100%	SMOTE	0.675 (0.004)	0.379 (0.017)	0.827 (0.007)	0.538 (0.012)
	ADASYN	0.692 (0.003)	0.519 (0.014)	0.774 (0.006)	0.624 (0.008)
	GMC-SMOTE	0.741 (0.003)	0.746 (0.008)	0.740 (0.005)	0.741 (0.003)
200%	SMOTE	0.684 (0.003)	0.497 (0.014)	0.783 (0.006)	0.613 (0.008)
	ADASYN	0.682 (0.004)	0.495 (0.010)	0.770 (0.001)	0.613 (0.006)
	GMC-SMOTE	0.725 (0.003)	0.821 (0.007)	0.679 (0.005)	0.746 (0.003)

Note) the numbers in parentheses are standard errors.

Table 5. Simulation results for the Haberman data set in Table 2

Percentage of over sampling	Method	Test classification accuracy			
		Overall accuracy	Sensitivity	Specificity	G-mean
0%	Standard SVM	0.723 (0.003)	0.155 (0.011)	0.932 (0.004)	0.331 (0.018)
100%	SMOTE	0.658 (0.004)	0.235 (0.011)	0.815 (0.007)	0.423 (0.009)
	ADASYN	0.618 (0.004)	0.321 (0.003)	0.726 (0.006)	0.471 (0.002)
	GMC-SMOTE	0.721 (0.004)	0.376 (0.010)	0.849 (0.006)	0.558 (0.007)
200%	SMOTE	0.631 (0.005)	0.324 (0.012)	0.745 (0.008)	0.479 (0.009)
	ADASYN	0.577 (0.005)	0.587 (0.005)	0.640 (0.010)	0.576 (0.009)
	GMC-SMOTE	0.646 (0.007)	0.531 (0.013)	0.693 (0.013)	0.594 (0.006)

Note) the numbers in parentheses are standard errors.

전체 정확도와 기하평균에서 전반적으로 우수함을 나타냄을 확인할 수 있다.

5. 결론

SMOTE는 불균형 자료에 대해 매우 유용한 샘플링 도구로 활용되지만, 보간법을 이용하여 소수집단의 개체와 개체 사이의 직선상에서만 새로운 개체를 생성함으로 분류분석에서 소수집단의 영역을 확장하는 데는 제한적이다. 이러한 문제를 극복하기 위하여 본 연구에서는 가우스 혼합 군집모형을 활용한 GMC-SMOTE를 제안하였다. GMC-SMOTE의 성능을 평가하기 위해 임의 생성자료 및 실제 자료를 분석하였다. 그 결과 소수집단의 비율에 관계없이 우수한 성능을 확인하였고 특히, 일반적으로 불균형 자료 분석에서 중요한 소수집단의 분류 정확도인 민감도 측면에서 기존의 방법에 비해 가장 우수한 성능을 보임을 확인하였다. 본 연구는 다양한 불균형 자료에 기반한 현실문제를 해결할 수 있는 대안을 제시하는데 의의가 크다고 할 수 있다. 일반 사용자들의 활용이 용이하도록 GMC-SMOTE의 프로그램을 제공하는 것은 추후 과제로 남긴다.

References

Akbani, R., Kwek, S., Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets, *In Proceedings of European Conference of Machine Learning,* 3201, 39-50.

Anand, A., Pugalenthi, G., Fogel, G. B., Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and undersampling, *Amino Acids*, 39(5), 1385-1391.

Bang, S., Jhun, M. (2014). Weighted support vector machine using k-means clustering, *Communications in Statistics-Simulation and Computation*, 43, 2307-2324.

Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C. (2009). Safe-level-SMOTE: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, *In Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 475-482.

Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C. (2012). DBSMOTE: density-based synthetic minority over-sampling technique, *Applied Intelligence,* 36, 664-684.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research,* 16, 321-357.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2003). C4.5 and imbalanced datasets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure, *Proceedings of the ICML*, 3.

Choi, H. (2019). Classification of bloodstream infection microbiome data using group information in taxonomy, *Journal of the Korean Data Analysis Society,* 21(2), 651-659. (in Korean).

Cortes, C., Vapnik, V. (1995). Support vector networks, *Machine Learning*, 20, 273-297.

Datta, S., Das, S. (2015). Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs, *Neural Networks,* 70, 39-52.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological),* 39(1), 1-22.

Drummond, C., Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling?, *Proceedings of the ICML*, 3.

Dua, D., Graff, C. (2019). *UCI machine learning repository,* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Han, H., Wang, W. Y., Mao, B. H. (2005). Borderline－SMOTE: a new over－sampling method in imbalanced data sets learning, *Lecture Notes in Computer Science*, 3644, 878-887.

He, H., Bai, Y., Garcia, E. A., Li, S. (2008). ADASYN: adaptive synthetic sampling approach for imbalanced learning, *In proceedings of the 2008 IEEE International Joint Conference Neural Networks,* 1322-1328.

Japkowicz, N. (2000). The class imbalance problem; significance and strategies, *In Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning,* 1, 111-117

Jeong, H., Kang, C., Kim, K. (2008). The effect of oversampling method for imbalanced data, *Journal of the Korean Data Analysis Society,* 10, 2089-2098. (in Korean).

Kubat, M., Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection, *In Proceedings of the Fourteenth International Conference on Machine Learning,* 179-186.

Kwon, S., Kim, A. (2018). A comparative study of classification methods using data with label noise, *Journal of the Korean Data Analysis Society,* 20(6), 2853-2864. (in Korean).

Lin, Y., Lee, Y., Wahba, G. (2002). Support vector machines for classification in nonstandard situations, *Machine Learning,* 46, 191-202.

Ling, C. X., Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem, *Encyclopedia of Machine Learning,* 2011, 231-235.

Oommen, T., Baise, L. G., Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression, *Mathematical Geosciences,* 43(1), 99-120.

Owen, A. B. (2007). Infinitely imbalanced logistic regression, *The Journal of Machine Learning Research,* 8, 761-773.

Park, J., Bang, S. (2015). Logistic regression with sampling techniques for the classification of imbalanced data, *Journal of the Korean Data Analysis Society,* 17(4), 1877-1888. (in Korean).

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, 6(2), 461-464.

Scrucca, L., Fop, M., Murphy, T. B., Raftery, A. E. (2016). Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models, *The R Journal,* 8(1), 289-317.

Singh, N. D., Dhall, A. (2018). Clustering and learning from imbalanced data, *arXiv preprint arXiv:1811.00972.*

Siriseriwan, W. (2016). Smotefamily: a collection of oversampling techniques for class imbalance problem based on smote, *R Package version 1.3.1.* URL http://cran. r-project. org/package=smotefamily.

Tang, Y., Zhang, Y. Q., Chawla, N. V., Krasser, S. (2008). SVMs modeling for highly imbalanced classification, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* 39, 281-288.

Wang, Q., Luo, Z., Huang, J., Feng, Y., Liu, Z. (2017). A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM, *Computational Intelligence and Neuroscience,* 2017, 1-11

Zhang, Y. P., Zhang, L. N., Wang, Y. C. (2010). Cluster-based majority under-sampling approaches for class imbalance learning, *In Information and Financial Engineering (ICIFE) 2010 2nd IEEE International Conference*, 400-404.

Sampling Method Using Gaussian Mixture Clustering for Classification Analysis of Imbalanced Data*

Sungwan Bang[1], *Jaeoh Kim*[2]

Abstract

When analyzing imbalanced data with different class sizes, the classification accuracy in minority class (sensitivity) may drop significantly because traditional classifiers are biased toward the majority class so that they classifies almost all observations to majority class. The purpose of this study is to propose a sampling method for classifying imbalanced data using the Gaussian mixture clustering model. We compared our proposed method with the existing over-sampling methods such as SMOTE (synthetic minority over-sampling technique) and ADASYN (adaptive synthetic sampling), and confirmed the excellence of the proposed method in various situations. In particular, the proposed method outperformed all other methods in terms of the classification accuracy of the minority class, which are generally important in the analysis of imbalanced data. In this study, a support vector machine method is adopted as a classification method and each method is evaluated by overall accuracy, sensitivity, specificity, and geometric mean.

Keywords : Gaussian mixture clustering, imbalanced data, over-sampling, sampling techniques.

---

*This work was supported by 2020 research fund of Korea Military Academy.

[1]Professor, Department of Mathematics, Korea Military Academy, 574 Hwarang-Ro, Nowon-Gu, Seoul 01805, Korea. E-mail : wan1365@gmail.com

[2](Corresponding Author) Major, Center for Army Analysis and Simulation, HQs ROKA, 663, Gyeryongdae-ro, Sindoan-myeon, Gyeryong-si, Chungcheongnam-do, Korea. E-mail : c14180@gmail.com

[Received 20 March 2020; Revised 17 April 2020; Accepted 20 April 2020]