Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 3, pp. 232-241, June 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.3.232

<Original Research>

불균형 데이터 분류를 위한 Expectation-maximization 알고리즘과 경계 관측치를 이용한 SMOTE

이강혁[1] · 이강훈[2] · 고태훈[1†]

[1]가톨릭대학교 의과대학 의료정보학교실 / [2]한양대학교 공과대학 건설환경공학과

SMOTE Using Border-points and Expectation-maximization Algorithm for Imbalanced Data Classification

Kang Hyuck Lee[1] · Kang Hoon Lee[2] · Taehoon Ko[1]

[1]Department of Medical Informatics, College of Medicine, The Catholic University of Korea,
[2]Department of Civil and Environmental Engineering, College of Engineering, Hanyang University

Class imbalance refers to a classification problem which occurs when the class distribution is skewed in the dataset. It is known that class imbalance degrades the predictive performance of classifiers. As a solution to this problem, we propose GBL-SMOTE, a modified method of SMOTE technique considering the clusters and their border-points. GBL-SMOTE solves the class imbalance problem in the learning stage by generating artificial observations at the cluster boundary of minority class observations. In this study, we create an artificial class imbalanced dataset to show the effect of GBL-SMOTE compared to original SMOTE and other SMOTE-variants. In addition, through experiments on the publicly available data, it was shown as an experiment that learning a support vector machine using GBL-SMOTE has better prediction performance than using other oversampling techniques.

Keywords: SMOTE, Borderline-SMOTE, Cluster, Gaussian Mixture Model, GBL-SMOTE

1. 서 론

불균형 데이터는 한 범주에 속한 데이터의 관측치 수와 다른 범주에 속한 데이터의 관측치 수의 균형이 맞지 않을 때 발생하며, 현실 세계에 있는 많은 분류 문제들에서 데이터 불균형 현상이 발생한다. 대표적으로 감지(Fawcett *et al*., 1997), 의료 진단(Zahirinia *et al*., 2015), 생체 인식(Triguero *et al*., 2015), 공장 생산 결함, 보험 청구 및 표적 마케팅 등이 있다.

대부분의 분류 알고리즘은 각 범주에 속한 데이터의 관측치 수가 비슷할 경우를 가정하여 분류를 진행한다. 불균형 데이터를 이용하여 분류 알고리즘을 사용할 경우, 분류기의 학습 시간이 지체되거나 전혀 작동되지 않은 경우가 발생할 수 있으며, 분류기를 통한 분석 결과에 대해서도 신뢰할 수 없다는 문제점을 가질 수 있다. 또한, 분류 결과가 다수 범주에 편향된 결정을 내리는 경향이 있으며, 극단적인 경우 분류기는 모든 개체를 다수 범주로 분류한다(Weiss *et al*., 2003).

데이터의 불균형을 해결하기 위해 여러 접근법이 제안되었다(He *et al*., 2009; Lopez *et al*., 2013). 샘플링 수준(Chawla *et al*., 2009; Batista *et al*., 2004), 알고리즘 수준(Zadronzy *et al*., 2001), cost-sensitive learning(Zardronzy *et al*., 2003; Sun *et al*., 2007) 접근 방법이 있으며, 그 중 샘플링 기반의 기법들은 많이 사용되며 다수 범주의 데이터의 관측치 수를 줄이거나, 소수 범주의 데이터의 관측치 수를 증가시켜 데이터의 불균형을 조절한다.

† 연락저자 : 고태훈, 우편번호 서울시 서초구 반포대로 222 가톨릭대학교 성의교정, Tel : 02-2258-7947, Fax : 0508-906-5445,
E-mail : thko@catholic.ac.kr
2020년 10월 21일 접수; 2020년 12월 8일 수정본 접수; 2020년 12월 10일 게재 확정.

고전적으로 SMOTE(Synthetic Minority Oversampling Technique)는 임의의 소수 관측치를 선택한 후, 가장 가까운 이웃 중 하나를 연결하여 가중치를 통해 인공 관측치를 무작위로 생성한다(Chawla *et al*., 2002). 소수 범주의 인공 관측치들의 생성으로 인해 데이터의 불균형을 조절하여 분류기의 성능을 향상시켰다. 하지만, 인공 관측치를 생성하기 위해 선택된 소수 범주의 관측치들의 수에 의해 분류기의 성능이 정해지는 단점이 있으며, 새롭게 생성된 인공 관측치들의 개체 수에 의한 과적합 현상과 이상치의 발생 현상을 해결하지는 못했다.

SMOTE 기법으로 해결되지 않은 과적합, 이상치 등의 문제를 보완하기 위해 Haibo는 ADASYN(Adaptive Synthetic Sampling)을 제안하였다(Haibo *et al*., 2008). ADASYN 기법은 각각의 소수 범주의 관측치에 따라 생성시킬 인공 관측치의 수를 결정함으로써 SMOTE 기법에 비해 조금 더 체계적으로 데이터를 생성시켜 SMOTE 기법의 단점을 보완하였다. Han은 SMOTE 기법에서 경계영역(Borderline)에 해당하는 위치에서 SMOTE 기법을 기반으로 인공 데이터를 생성하는 Borderline-SMOTE을 제안하였다(Han *et al*., 2005).

Zhang은 SMOTE 기법에 가우시안 혼합 모델(Gaussian Mixture Model) 군집화 기법을 함께 사용하여 SMOTE 기법의 단점을 보완하는 G-SMOTE 기법을 제시하였다(Zhang *et al*., 2018). 소수 범주의 데이터에 대해 가우시안 혼합 모델을 이용하여 군집을 형성한 후, 각각의 군집에서 SMOTE 기법을 이용하여 인공 관측치들을 생성하였다. 하지만, SMOTE 기법을 기반으로 하여 인공 관측치들을 생성하기에 각 군집에서 인공 관측치 생성 시 특정 부분에서 과적합 현상과 이상치로 판단되는 관측치에서 인공 관측치들이 생성되는 문제가 발생하였다.

본 연구에서는 불균형 데이터 분류에 좀 더 효과적으로 사용할 수 있는 GBL-SMOTE라는 기법을 제안하고자 한다. 이 기법은 가우시안 혼합 모델을 통해 소수 범주의 데이터에 대해 군집화를 형성시킨 후, Borderline-SMOTE 기법을 이용하여 인공 데이터를 생성한다. 각각의 군집의 경계면에 인공 데이터가 생성됨으로써 SMOTE 기법을 이용하였을 때보다 과적합 현상이 완화되며 분류 경계면 주변의 샘플이 생성되는 효과가 있다. 이번 연구에서는 시뮬레이션을 통해 제안하는 기법인 GBL-SMOTE가 군집의 경계면에서 인공 데이터가 생성됨을 보이고자 한다. 그리고 실제 데이터를 이용하여 기존 SMOTE, G-SMOTE 기법의 단점을 GBL-SMOTE 기법을 통해 보완하여 불균형 데이터 셋에 대한 분류 모델 성능을 평가하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 선행 연구에 사용된 기법들의 알고리즘에 대해 알아보며, 제 3장에서는 제안 기법에 대해 자세히 설명한다. 제 4장에서는 예제 데이터를 통해 기존 연구와 제안 기법과의 성능 비교를 실시한다. 마지막으로 제 5장에서는 본 논문의 결론과 추후 실시할 수 있는 연구에 대해 알아본다.

2. 선행 연구 고찰

2.1 SMOTE

SMOTE 기법은 기존에 있는 데이터를 단순 복제하는 기법과는 다르게 소수 범주의 관측치들 중 K개의 가장 가까운 관측치들과의 선형 관계 사이에 가중치를 두어 인공 관측치를 생성하는 기법이다. SMOTE 기법의 알고리즘은 다음과 같다.

첫 번째로, SMOTE 기법은 오버샘플링을 실행할 임의의 소수 범주의 관측치를 선택한다. 이 때, 소수 범주의 데이터 수 이상으로 인공 데이터를 생성할 경우에는, 소수 범주의 모든 관측치들이 선택된다. 생성하는 인공 데이터의 수가 소수 범주의 데이터 수보다 작을 경우에는 소수 범주의 모든 관측치들 중 일부분이 임의로 선택된다.

두 번째로, 인공 데이터를 생성하기 위해 선택된 소수 범주의 관측치의 K개의 가장 가까운 이웃 관측치들을 선별한 후, 임의로 $m(\leq K)$개의 관측치를 선택한다. 선택된 m개의 관측치들과 첫 번째 단계에서 기준이 된 소수 범주의 관측치와 선형 관계를 형성한 후 가중치를 곱한다.

$$x_{syn} = x + (x_j - x) \times w. \qquad j = 1, 2, \cdots, m$$

이 때, x는 첫 번째 단계에서 선택된 소수 범주의 관측치, x_j는 두 번째 단계에서 x와 가장 가까운 이웃에 있는 소수 범주의 관측치들 중 한 관측치, w는 0과 1사이에서 발생시킨 난수로 가중치를 의미한다.

마지막으로, 첫 번째 단계에서 선택한 소수 범주의 관측치를 제외한 나머지 관측치들 중에서 다시 임의로 선택해서 나머지 단계를 반복한다.

2.2 Borderline-SMOTE

SMOTE 기법은 다수 범주의 관측치들과 소수 범주의 관측치들의 위치에 대한 고려없이 소수 범주의 관측치들을 이용하여 인공 관측치들을 생성한다. 이것은 다수 범주의 관측치들과 소수 범주의 관측치들이 서로 겹치는 문제를 발생시킬 수 있으며, 기계 학습 알고리즘을 이용한 분류기를 사용할 때 분류기의 성능을 저하시킬 수 있다. 이 때, 다수 범주의 데이터와 소수 범주의 데이터가 서로 겹쳐 있는 영역을 경계영역이라 정의하며, Borderline-SMOTE 기법은 이런 문제점을 해결하기 위해 경계영역에 있는 소수 계급의 관측치들에 대해 SMOTE 기법을 적용시켜 인공 관측치들을 생성시킨다(Han *et al*., 2005).

Borderline-SMOTE 기법은 소수 범주에 속하는 각각의 모든 관측치들에 대해 범주에 상관없이 N개의 가장 가까운 관측치들을 선택한다. 두 번째로, 구한 관측치들 중에서 다수 범주에 속하는 관측치의 수를 확인한다. 다수 범주의 관측치의 수를 $|N_{maj}|$라 할 때, $\frac{N}{2} \leq |N_{maj}| \leq N$의 경우 "Danger", $0 \leq |N_{maj}| \leq \frac{N}{2}$

의 경우 “Safe”, $|N_{maj}| = N$의 경우 “Noise” 집단으로 구분할 수 있다. 마지막으로, “Danger” 집단에 속하는 소수 범주의 관측치들에 대해 SMOTE 기법을 이용하여 인공 관측치들을 생성한다.

2.3 G-SMOTE

SMOTE 기법, Borderline-SMOTE 기법은 0과 1 사이의 난수를 생성하여 관측치들간의 선형 관계에 가중치를 곱하여 인공 관측치를 생성하였다. 그러나 인공 관측치를 생성하는 과정 중에 자주 선택되는 특정 소수 범주의 관측치와 가장 가까운 이웃 관측치 간에 둘 이상의 인공 관측치들이 생성되어 높은 확률로 과적합과 이상치, 다수 범주의 데이터와 소수 범주의 데이터가 겹치는 문제점이 발생한다. 따라서, 인공 관측치 생성에 있어서 다양성을 보장하기 위해 오버샘플링을 시행하기 전에 소수 계급의 관측치들을 가우시안 혼합 모델을 이용하여 군집화를 진행한다.

$$f(x|\mu,\Sigma) = \sum_{k=1}^{R} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp\left[(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right]$$

Expectation-maximization(EM) 알고리즘을 통해 R개의 가우시안 혼합 분포를 구성한다고 가정하였을 때, k번째 가우시안 혼합 모델의 가중치, 평균 벡터 그리고 공분산은 c_k, μ_k그리고 Σ_k라고 할 수 있다. 소수 범주의 데이터에 대해 R개의 정규분포로 군집을 구성한 후, 각각의 군집에 속한 소수 범주의 관측치들의 수를 통해 각각의 군집에서 새로 생성할 인공 관측치들의 비율을 설정하였다. 이 때, 인공 관측치들을 생성하는 방법은 SMOTE 기법을 이용한다.

3. GBL-SMOTE

본 논문에서 제안한 기법은 먼저 데이터에 대해 EM 알고리즘을 이용하여 소수 범주의 데이터에 대해 군집을 형성한다. 이후, 각각의 군집에 속한 소수 범주의 데이터와 군집을 형성하지 않은 다수 범주의 데이터를 통해 결정 영역의 관측치들을 찾는다. 마지막으로 확인된 소수 범주의 관측치들을 대상으로 SMOTE 기법을 통해 인공 관측치들을 생성한다. 아래는 GBL-SMOTE의 알고리즘이다.

Algorithm : GBL-SMOTE

Input
- $\boldsymbol{U}$: all samples
- $\boldsymbol{M}$: samples in minority class
- r : the number of nearest neighbors in $\boldsymbol{U}$ for each minority sample
- k : the number of nearest neighbors in $\boldsymbol{M}$ for each borderline sample

Output
- $\boldsymbol{S}$: Minority samples and corresponding synthetic samples

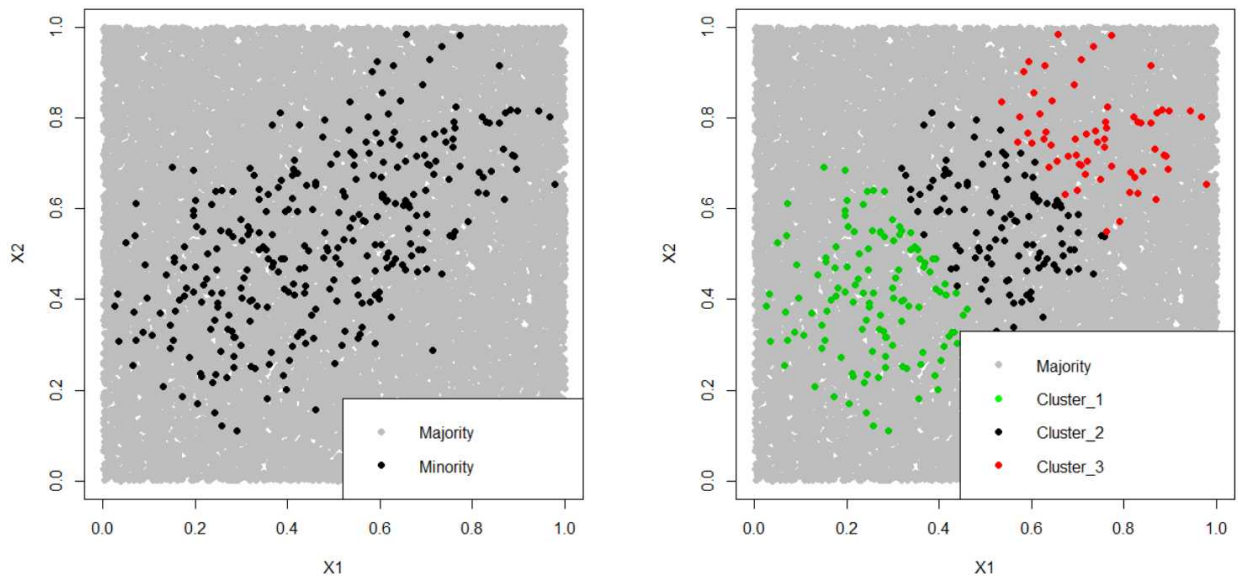
```
Create optimal Gaussian Mixture Model with the smallest BIC for M
for i ← 1 to c do // c is the number of clusters
    for all m in M do
        R ← r nearest neighbors of m in U
        n ← the number of samples in R and not in M
            if r/2 ≤ n ≤ r then
                add m to D_i // D_i is a set of borderline samples in i-th cluster
            end
    end
        for all d in D_i do
            L ← k nearest neighbors of d in D_i
            for all d in D_i do
                l ← a random sample from L
            d' ← d + p × (d − l) // p is a random number in (0,1)
                                 // d' is a synthetic sample
                add d' to S_i' // S_i' is a set of synthetic samples
            end
        end
    S = M ∪ S_i' // S is the union of minority samples and synthetic samples
end
return S
```

본 연구에서 제안하는 GBL-SMOTE 특징과 장점을 보기 위하여 인공 데이터에 대한 실험을 시행하였다. 우선, 1×1 크기의 공간에서 균일 분포를 이용하여 임의의 10,000개의 관측치들을 생성한다. 이 후, (0.5, 0.5), (0.7, 0.7), (0.2, 0.2)를 중심으로 하는 원에서 임의로 관측치들을 추출하여 소수 범주의 데이터를 생성한다. 이 때, 극단적인 불균형 정도를 주기 위해 300개의 관측치들만 추출한다. <Figure 1(a)>는 9,700개의 다수 범주의 데이터와 300개의 소수 범주의 데이터의 모습을 보여준다. <Figure 1(b)>는 300개의 소수 범주의 데이터에 대해 가우시안 혼합 모델을 생성한 것을 나타낸다. 군집을 형성하기 위해 EM 알고리즘을 사용하였고, 3개의 군집이 형성되었다.

<Figure 2(a)>는 <Figure 1(a)>의 소수 범주의 데이터에서 SMOTE 기법을 이용하여 인공 관측치들을 생성한 모습이며, <Figure 2(b)>는 <Figure 1(b)>에서 소수 범주의 데이터를 SMOTE 기법을 이용하여 인공 관측치들을 생성하는 G-SMOTE 기법을 보여준다. 이 때, 각각의 기법으로 생성된 인공 관측치들은 총 2,700개로 동일하다. G-SMOTE 기법의 경우 SMOTE 기법과 비교하였을 때, 각각의 군집에서 생성되는 인공 관측치들의 개수가 제한된다. 이번 실험에서는 각각의 군집에서 900개씩 생성하였다. <Figure 2(a)>와 <Figure 2(b)>를 비교하였을 때, 전체적으로 생성되는 인공 관측치들의 위치는 서로 유사함을 보인다. 하지만 (b)에 비해 (a)의 경우, 인공 관측치들이 특정 부분에서 많이 생성되는 현상이 나타나고, 이로 인해 실제 클래스 경계면과 다른 분류 경계가 학습될 가능성이 높다. 이로 인해 인공 관측치에 의한 과적합 현상이 나타날 것으로 예상된다.

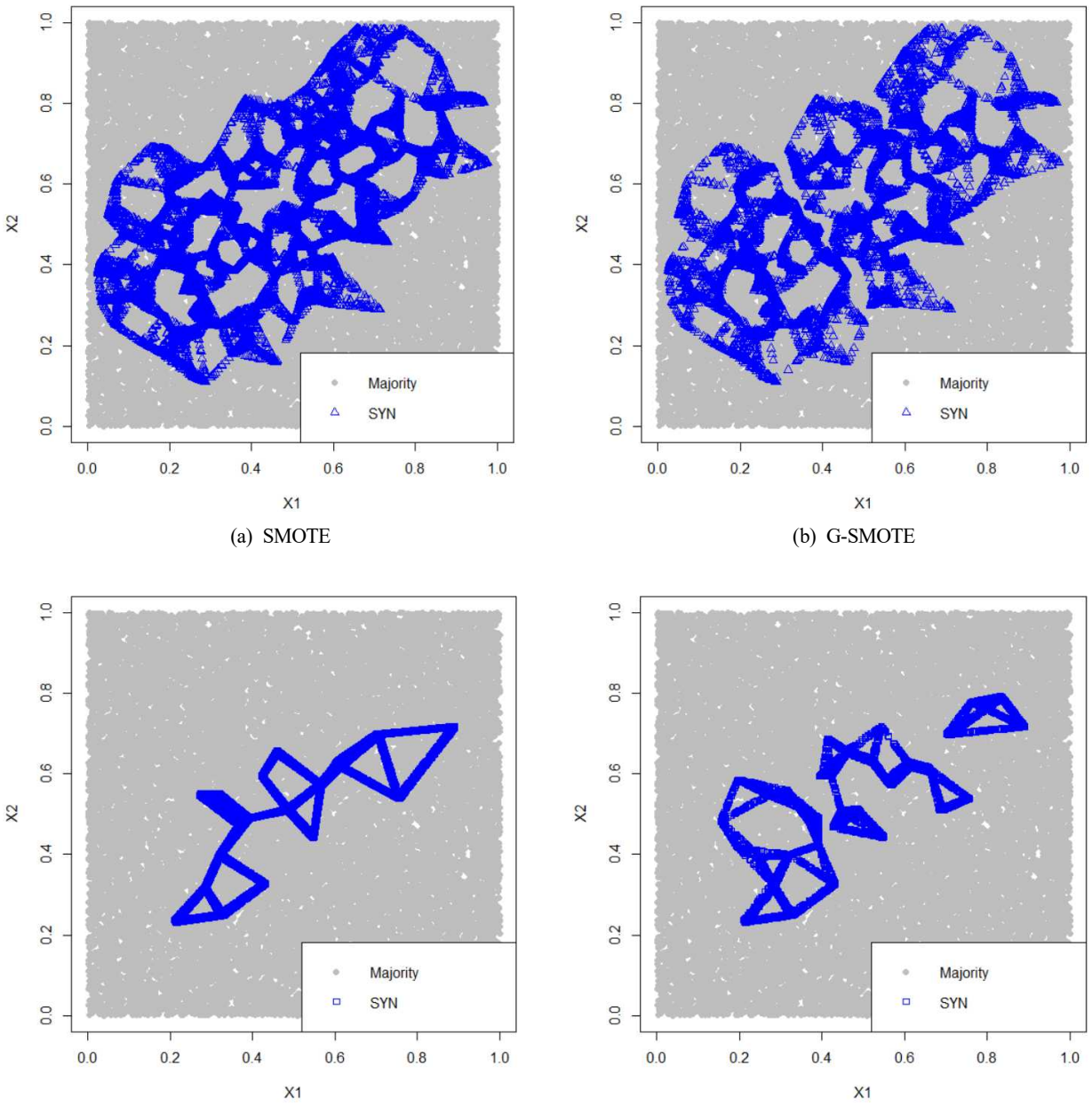
<Figure 2(c)>는 <Figure 1(a)>의 소수 범주의 데이터에서 Borderline-SMOTE 기법을 이용하여 인공 관측치들을 생성한 모습이며, <Figure 2(d)>는 <Figure 1(b)>에서 소수 범주의 데이터를 Borderline-SMOTE 기법을 이용하여 인공 관측치들을 생성하는 GBL-SMOTE 기법을 보여준다. <Figure 2(a)>, <Figure 2(b)>와 동일하게 생성한 인공 관측치들의 개수는 2,700개이다. 또한, 각각의 군집에서 생성되는 인공 관측치들의 개수 또한 900개로 동일하게 생성하였다. <Figure 2(c)>의 경우에 <Figure 2(a)>와 비교하였을 때, 소수 범주의 데이터에 대해서 경계영역에 속한 소수 범주의 관측치들에 제한하여 생성됨으로써, 생성된 인공 관측치들의 위치가 특정한 곳에 제한되는 모습을 확인할 수 있다. <Figure 2(d)>의 경우, 각각의 군집의 속한 소수 범주의 데이터에 대해 경계영역에 속한 소수 범주의 관측치들을 대상으로 인공 관측치들이 생성되었다. <Figure 1>의 경우 생성된 인공 관측치들의 위치가 서로 유사하였던 것과는 다르게 <Figure 2(c)>와 <Figure 2(d)>의 경우 생성되는 인공 관측치들의 위치가 서로 다른 모습을 보여준다. <Figure 1(a)>와 <Figure 2(c)>를 함께 생각하였을 때, <Figure 2(c)>의 생성된 인공 관측치들의 위치는 <Figure 1(a)>의 내부에서 생성되는 모습을 보여준다. <Figure 1(b)>와 <Figure 2(d)>를 함께 생각하였을 때, <Figure 2(d)>의 생성된 인공 관측치들의 위치는 <Figure 1(b)>에서 군집의 경계면에서 생성되어 군집의 경계면을 강화하는 모습을 보인다. 또한, <Figure 2(a)>와 <Figure 2(b)>에 비해 제한된 위치에서 인공 관측치들이 생성되었으며, 상대적으로 이상치의 위치에서 인공 관측치가 생성되지 않는 모습을 보여준다.

모의 실험을 통해 각각의 기법을 사용하였을 때, 어떤 위치에서 인공 관측치들이 생성되는지 확인하였다. SMOTE 기법과 G-SMOTE 기법의 경우 서로 유사한 위치에서 인공 관측치들이 생성되는 모습을 보였다. 또한, Borderline-SMOTE 기법의 경우 데이터의 내부에서 경계영역이 발견되어 인공 관측치들이 생성됨으로써 상대적으로 데이터의 내부가 두꺼워지는 모습을 보였다.



(a) Raw data (b) Clustering with Raw data

Figure 1. Raw data and Clustering with Raw data



(a) SMOTE (b) G-SMOTE

(c) Borderline-SMOTE (d) GBL-SMOTE

Figure 2. Comparison of SMOTE, G-SMOTE, Borderline-SMOTE and GBL-SMOTE

본 연구에서 제안한 GBL-SMOTE 기법의 경우 각각의 군집의 결정 영역에서 생성된 소수 범주의 인공 관측치들로 인해, 각각의 군집의 경계면을 강화시켜주는 모습을 보였다. 상대적으로 SMOTE 기법과 Borderline-SMOTE 기법, G-SMOTE 기법에 비해 GBL-SMOTE 기법은 과적합 현상과 이상치의 생성이 적게 발생하는 모습을 확인할 수 있었다. 따라서, GBL-SMOTE 기법의 경우 군집 경계면에 소수 범주의 관측치들에 대한 인공 관측치가 생성이 되면, 의사결정경계 부분에 다수 범주와 소수 범주의 불균형 정도가 완화되어 기계학습 분류기의 의사결정 경계가 뚜렷해지는 효과를 기대할 수 있다.

4. 실 험

4.1 데이터 설명

실험에 사용된 데이터는 “KEEL Data set repository”에 공개된 불균형 데이터들로 각각의 실험 데이터의 불균형 정도, 관측치 수, 설명 변수 그리고 군집 수는 <Table 1>을 통해 확인할 수 있다(Alcala-Fdez *et al*., 2011). 데이터는 다양한 경우에 대한 실험을 위해 불균형 정도는 1.87부터 129.44까지, 데이터 수는 214부터 4174까지의 데이터를 사용하였다. 불균형 정도는

다수 범주의 관측치 수를 소수 범주의 관측치 수로 나눈 값으로, 값이 클수록 클래스 불균형이 매우 크다는 것을 의미한다. 또한, 각각의 실험 데이터에 EM 알고리즘을 사용하여 가우시안 혼합 모델 학습하였으며, Bayesian information criterion(BIC)가 가장 작을 때의 군집 수를 최적의 군집 수로 정하였다. EM 알고리즘에 대한 자세한 수식 과정은 <Appendix>를 참고할 수 있다.

4.2 실험 설계 및 결과

본 논문에서 제안한 기법의 성능 평가를 위해 <Table 1>에서 제시한 데이터를 사용하여 분석을 진행하였다. 비교 분석에 앞서 소개한 SMOTE, Borderline-SMOTE, 가우시안 혼합 모델 군집화를 이용한 SMOTE 기법을 이용하여 비교 평가하였다. SMOTE의 k값의 경우 일반적으로 사용되는 k = 3을 사용 하였으며(Chawla *et al*., 2002), 인공 데이터의 경우 원본 데이터의 소수 계급의 데이터 수 만큼 생성하였다. 또한, 예제 데이터들에 대하여 seed를 다르게 설정하여 5-fold 교차검증을 10번 반복실험을 수행하여 총 50번의 군집화 과정을 계산하였다. 각 실험에서 BIC값이 최저일 때의 군집의 개수를 구하였으며, 이들의 평균값을 반올림하여 최종 군집의 수로 선정하였다. 매 실험마다 각 변수를 표준화(Standardization) 하여 평균 0, 분산 1로 스케일을 변경하였다. 실험 결과는 <Table 2>에 요약되어 있다.

분류기는 Support Vector Machine(SVM)을 사용하였으며, 샘플링 이후에 각 예제 데이터에 최적으로 설정된 초모수(Hyper-parameter)를 찾아 분류를 실행하였다. SVM의 커널함수는 Radial Basis Function(RBF) 커널을 사용하였으며, Gamma의 경우 1/(Number of features) 값을 사용하였다. 분류기의 성능 평가를 위한 척도로는 Reciver Operating Characteristic(ROC) 곡선의 Area Under Curve(AUC)를 사용하였다. ROC 곡선은 특이도(Specificity)에 따른 민감도(Sensitivity)의 변화를 나타낸 곡선으로, 밑면적인 AUC는 다수 계급과 소수 계급에 대한 정분류율을 동시에 고려할 수 있는 척도이므로 불균형 데이터의 분류 성능을 평가하기에 적합하다. <Table 3>은 오버샘플링 기법을 적용하지 않은 원본 데이터와 4개의 오버샘플링 기법을 적용한 후 기계학습 분류기에 학습시켰을 때의 결과를 나타내며, 각 데이터마다 50번의 실험을 통해 나온 결과의 평균값을 기입했다.

예제 데이터 1, 5, 6, 9, 11, 14, 17, 18의 결과를 확인하였을 때, 군집을 형성하지 않고 SMOTE, Borderline-SMOTE 기법을 사용한 것과 군집을 형성한 후 SMOTE, Borderline-SMOTE 기법을 사용한 것의 차이가 다른 예제 데이터 비해 상대적으로 명확하게 차이가 나는 것을 확인할 수 있다. 예제 데이터 5의 경우 SMOTE 기법과 GBL-SMOTE 기법의 AUC값의 차이가 0.04로 다른 데이터에 비해 조금 작은 차이를 보이지만, 예제 데이터 1의 경우에는 Borderline-SMOTE 기법과 GBL-SMOTE 기법의 AUC 값의 차이가 0.154로 큰 차이를 보이는 것을 확인할 수 있다.

예제 데이터 2, 6, 12, 13, 14의 경우에는 G-SMOTE 기법이 GBL-SMOTE 기법에 비해 AUC 값이 큰 것을 확인할 수 있다.

Table 1. Data Description

	Data	Imbalanced ratio	Observations	Independent Varaible
1	Abalone	16.40	731	7
2	Abalone19	129.44	4174	7
3	Ecoli-0-1_vs_2-3-5	9.17	244	7
4	Ecoli1	3.36	336	7
5	Ecoli4	15.8	336	7
6	Glass0	2.06	214	10
7	Pima	1.87	768	8
8	Segment	6.02	2308	18
9	Winequality-red-4	29.17	1599	11
10	Winequality-red-8_vs_6	35.44	656	11
11	Yeast-0-5-6-7-9_vs_4	9.35	514	8
12	Yeast-1_vs_7	14.30	459	7
13	Yeast-1-2-8-9_vs_7	30.57	947	8
14	Yeast-2_vs_4	9.08	514	8
15	Yeast-2_vs_8	23.10	482	8
16	Yeast4	28.10	1484	8
17	Yeast5	32.73	1484	8
18	Yeast6	41.40	1484	8

Table 2. Result of 50 GMMs for Each Data

	Data	Average of the number of clusters	Standard deviation of the number of clusters	Final determined number of clusters
1	Abalone	3.4	0.813	3
2	Abalone19	8.3	0.675	8
3	Ecoli-0-1_vs_2-3-5	3.34	3.061	3
4	Ecoli1	7.34	2.056	7
5	Ecoli4	6.94	2.271	7
6	Glass0	5.72	1.702	6
7	Pima	6.62	1.483	7
8	Segment	5.38	2.664	5
9	Winequality-red-4	8.3	0.909	8
10	Winequality-red-8_vs_6	6.08	1.676	6
11	Yeast-0-5-6-7-9_vs_4	4.46	2.573	4
12	Yeast-1_vs_7	5.02	2.637	5
13	Yeast-1-2-8-9_vs_7	4.18	0.691	4
14	Yeast-2_vs_4	8.4	1.723	8
15	Yeast-2_vs_8	3.98	1.392	4
16	Yeast4	4.36	0.757	4
17	Yeast5	4.52	0.762	4
18	Yeast6	4.6	0.707	5

예제 데이터 2의 경우에는 G-SMOTE 기법과 GBL-SMOTE 기법의 차이가 0.004로 상대적으로 근소한 차이를 보이지만, 예제 데이터 13의 경우에는 0.034 차이가 나는 것을 확인할 수 있다.

Table 3. Classification Performance of the Different Oversampling Techniques by AUC

Data \ Methods		RAW	SMOTE	BLSMOTE	G-SMOTE	GBL-SMOTE
1	Abalone	0.562	0.720	0.668	0.809	**0.822**
2	Abalone19	0.5	0.694	0.649	**0.718**	0.714
3	Ecoli-0-1_vs_2-3-5	0.711	0.854	0.870	0.870	**0.914**
4	Ecoli1	0.807	0.873	0.866	0.871	**0.886**
5	Ecoli4	0.787	0.886	0.871	0.915	**0.926**
6	Glass0	0.719	0.757	0.741	**0.833**	0.819
7	Pima	0.703	0.710	0.721	0.703	**0.741**
8	Segment	0.988	0.987	0.986	0.991	**0.993**
9	Winequality-red-4	0.5	0.630	0.645	0.670	**0.695**
10	Winequality-red-8_vs_6	0.5	0.652	0.654	0.661	**0.694**
11	Yeast-0-5-6-7-9_vs_4	0.587	0.734	0.775	0.787	**0.794**
12	Yeast-1_vs_7	0.558	0.642	0.666	**0.714**	0.695
13	Yeast-1-2-8-9_vs_7	0.5	0.643	**0.688**	0.674	0.683
14	Yeast-2_vs_4	0.774	0.810	0.798	**0.882**	0.857
15	Yeast-2_vs_8	0.648	0.617	0.693	0.700	**0.761**
16	Yeast4	0.5	0.741	0.754	0.751	**0.787**
17	Yeast5	0.734	0.912	0.908	0.889	**0.926**
18	Yeast6	0.571	0.853	0.851	0.871	**0.874**
Average of AUC		0.647	0.762	0.766	0.795	**0.810**

예제 데이터 8의 경우에는 인공 데이터를 생성하지 않고 분류를 진행하여 나온 AUC와 인공 데이터를 생성한 후 분류를 진행하여 나온 AUC의 차이가 매우 작으며, AUC도 가장 작은 값이 0.986, 가장 큰 값이 0.993으로 0.007의 상대적으로 작은 차이를 보인다.

또한, 예제 데이터 9, 10, 12, 13의 경우에는 원본 데이터의 AUC를 확인하였을 때, 다수 계급과 소수 계급의 관측치들을 구분할 수 있는 능력이 매우 작음을 확인할 수 있다. 4개의 예제 데이터의 경우 군집을 형성한 후 인공 데이터를 생성했을 경우, AUC 값이 상대적으로 많이 상승한 것을 확인할 수 있었다.

예제 데이터 15, 16, 17의 경우에는 본 연구에서 제안하는 GBL-SMOTE 기법을 이용하여 분류를 진행하였을 때, AUC가 가장 높은 값을 가지는 것을 확인할 수 있었다. 3가지 데이터의 경우, 군집의 경계면에서 인공 데이터를 생성함으로써 각각의 군집의 경계면을 강화시켜 분류기의 의사결정경계가 상대적으로 뚜렷해져 분류 평가지표인 AUC 값이 다른 기법들에 비해 가장 높게 나타났다.

종합적으로, 군집화 후 오버샘플링 기법을 적용시켜 기계학습 분류기에 학습을 진행하였을 때, 군집화를 이용하지 않은 경우보다 AUC가 높게 나타나는 것을 확인할 수 있었다. 또한, 18개의 예제 데이터 중 13개의 예제 데이터는 GBL-SMOTE를 이용하여 인공 데이터를 생성한 후 기계학습 분류기를 학습하였을 때, AUC가 가장 높게 나타나는 것을 확인할 수 있었다. 모의 실험을 통해 확인하였던, 군집의 경계면에서 인공 관측치들이 생성되어 다른 기법들에 비해 과적합 현상과 이상치에 둔감하게 반응하여 분류 결과값이 높게 나타나는 것을 예제 데이터를 이용한 실험을 통해 추가적으로 확인할 수 있었다.

5. 결 론

데이터 불균형으로 인한 분류 문제는 현실에서 수많은 분야에서 발생하며, 기계 학습 분야에서 오랫동안 연구되어왔다. 데이터 불균형을 해결하기 위해 샘플링 수준의 접근 방법으로 SMOTE 기법을 중심으로 다양한 오버샘플링 기법들이 제안되었지만, 과적합과 이상치의 존재 등 단점들이 존재하였다. SMOTE 기법의 경우 생성되는 인공 데이터의 개수가 특정 위치에서 임의로 정해짐에 따라 과적합 현상과 이상치로 판단되는 데이터에서 인공 데이터가 생성되는 단점이 있다. 또한, Borderline-SMOTE 기법의 경우 경계면에서 인공 데이터가 생성되지만, 이상치로 판단되는 데이터를 분별하지 못하는 단점이 있다. 가우시안 혼합 모델을 이용해 군집을 형성한 후, SMOTE 기법을 사용하여 인공 데이터를 생성하였을 경우에는 기존의 기법들이 가진 단점을 보완하는 모습을 보였지만, 이상치로 판단되는 관측치에서 인공 관측치들이 생성되는 단점을 해결하지는 못했다. 본 논문에서는 기존 기법들의 단점을 보완하기 위해 GBL-SMOTE 기법을 제안하였다. 18개의 예제 데이터를 통해 분석 결과를 비교하였을 때, 13개의 데이터의 경우 GBL-SMOTE 기법을 사용하였을 때, 더 좋은 결과가 나온 것을 확인할 수 있었다. 하지만 데이터의 관측치의 수와

변수의 개수가 많을수록 가우시안 혼합 모델을 이용한 군집화 형성에 계산량 증가에 따른 많은 시간이 소비되는 단점이 존재한다. 또한, SMOTE기법을 기반으로 하여, 오버샘플링을 하기 위해 선택하는 가장 가까운 이웃의 개수와 생성되는 인공 데이터의 수의 최적화 과정에 많은 시간이 소비되는 단점도 존재한다. 따라서, 추후에 진행될 수 있는 연구로는 가우시안 혼합 모델이 아닌 다른 군집화 기법을 사용한 것과 ADASYN과 같은 다른 SMOTE 기반의 오버샘플링 기법과 연계하는 것, 생성하는 인공 데이터 개수와 가장 가까운 이웃의 개수의 최적값을 찾는 방법에 대한 연구가 필요하다.

참고문헌

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011), Keel Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework, *Journal of Multiple-Valued Logic and Soft Computing*, **17**, 255-287.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004), A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, *ACM SIGKDD Explorations Newsletter*, **6**(1), 20-29.

Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. (2012), DBSMOTE: Density-based Synthetic Minority Over-sampling Technique, *Applied Intelligence*, **36**(3), 664-684.

Cordón, I., García, S., Fernández, A., and Herrera, F. (2018), Imbalance : Oversampling Algorithms for Imbalanced Classification in R., *Knowledge-Based Systems*, **161**, 329-341.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002), SMOTE : Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, **16**, 321-357.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004), Special Issue on Learning from Imbalanced Data Sets, *ACM SIGKDD Explorations Newsletter*, **6**(1), 1-6.

Chawla, N. V. (2009), Data Mining for Imbalanced Datasets : An Overview, *In Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA, 875-886.

Fawcett, T. and Provost, F. (1997), Adaptive Fraud Detection, *Data Mining and Knowledge Discovery*, **1**(3), 291-316.

Han, H., Wang, W. Y., and Mao, B. H. (2005), Borderline-SMOTE : A New Over-sampling Method in Imbalanced Data Sets Learning, *In International Conference on Intelligent Computing*, Springer, Berlin, Heidelberg, 878-887.

He, H. and Garcia, E. A. (2009), Learning From Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), 1263-1284.

Japkowicz, N. and Stephen, S. (2002), The Class Imbalance Problem : A Systematic Study, *Intelligent Data Analysis*, **6**(5), 429-449.

Lee, H., Kim, J., and Kim, S. (2017), Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions, *International Journal of Fuzzy Logic and Intelligent Systems*, **17**(4), 229-234.

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013), An Insight into Classification with Imbalanced Data : Empirical Results and Current Trends on Using Data Intrinsic Characteristics, *Information Sciences*, **250**, 113-141.

Pal, B. and Paul, M. K. (2017), A Gaussian Mixture based Boosted Classification Scheme for Imbalanced and Oversampled data, *In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 401-405.

Pan, T., Zhao, J., Wu, W., and Yang, J. (2020), Learning Imbalanced Datasets based on SMOTE and Gaussian Distribution, *Information Sciences*, **512**, 1214-1233.

Sanguanmak, Y. and Hanskunatai, A. (2016), DBSM : The Combination of DBSCAN and SMOTE for Imbalanced Data Classification, *In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 1-5.

Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007), Cost-sensitive Boosting for Classification of Imbalanced Data, *Pattern Recognition*, **40**(12), 3358-3378.

Triguero, I., del Río, S., López, V., Bacardit, J., Benítez, J. M., and Herrera, F. (2015), ROSEFW-RF : The Winner Algorithm for the ECBDL'14 Big Data Competition : An Extremely Imbalanced Big Data Bioinformatics Problem, *Knowledge-Based Systems*, **87**, 69-79.

Weiss, G. M. and Provost, F. (2003), Learning When Training Data are Costly : The Effect of Class Distribution on Tree Induction, *Journal of Artificial Intelligence Research*, **19**, 315-354.

Zadrozny, B. and Elkan, C. (2001), Learning and Making Decisions When Costs and Probabilities are Both Unknown, *In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 204-213.

Zadrozny, B., Langford, J., and Abe, N. (2003), Cost-Sensitive Learning by Cost-Proportionate Example Weighting, *In Third IEEE International Conference on Data Mining*,435-442.

Zahirnia, K., Teimouri, M., Rahmani, R., and Salaq, A. (2015), Diagnosis of Type 2 Diabetes Using Cost-Sensitive Learning, *In 2015 5th International Conference on Computer and Knowledge Engineering (ICCKE)*, 158-163.

Zhang, T. and Yang, X. (2018), G-SMOTE : A GMM-based Synthetic Minority Oversampling Technique for Imbalanced Learning, *arXiv preprint, arXiv:1810.10363*.

Zadrozny, B., Langford, J., and Abe, N. (2003), Cost-Sensitive Learning by Cost-proportionate Example Weighting, *In Third IEEE International Conference on Data Mining,* 435-442.

저자소개

이강혁 : 한양대학교 에리카캠퍼스 응용수학과에서 2015년 학사, 한양대학교 응용통계학과에서 2020년 석사학위를 취득하였으며 현재 가톨릭대학교 의과대학 의료정보학교실에서 연구원으로 재직 중이다. 연구분야는 머신러닝, 오버샘플링이다.

이강훈 : 성균관대학교 건설환경공학과에서 2017년 박사를 취득하고 현재 성균관대학교 건설환경연구소에서 선임 연구원으로 재직 중이다. 주 연구분야는 머신러닝을 이용한 폐기물 재활용, 수처리 공정이다.

고태훈 : 서울대학교 산업공학과에서 2008년 학사, 2017년 박사학위를 취득하였다. 서울대학교병원에서 연구교수를 역임하고, 현재 가톨릭대학교 의료정보학교실에서 연구조교수로 재직 중이다. 주 연구분야는 머신러닝 및 딥러닝 알고리즘 개선, 의료정보의 표준화, 개인정보보호이다.

< Appendix>

$r=1, 2, \cdots, R$ 에 대하여 $0<\pi_r<1$, $\sum_{r=1}^{R}\pi_r=1$ 이고, f_r 은 확률밀도함수라 할 때, 혼합분포(mixure distribution)는 식 (1)과 같은 확률밀도 함수를 갖는다.

$$\sum_{r=1}^{R}\pi_r f_r. \tag{1}$$

π_r 는 r번째 군집의 비율, f_r 는 r번째 군집의 확률밀도함수, R은 총 군집의 수이며, p차원 가우시안 혼합 모형은 식 (1)에서 $f_r(\circ)=\Phi(\circ;\mu_r,\Sigma_r)$ 인 경우에 해당한다. Φ는 p차원 정규분포의 확률밀도함수이며, 평균벡터가 μ_r 이며, 공분산행렬 Σ_r 을 갖는다.

EM 알고리즘(Expectation-Maximization)은 확률모형이 관측되지 않은 잠재변수(latent variable)에 의존하는 경우에 최대우도 추정량을 구할 때 사용된다. x는 관측된 변수들의 값이며, z는 관측되지 않은 잠재변수들의 값이라고 할 때, 모수 θ에 대하여 관측된 자료 x의 로그 우도는 $l(\theta;x)$ 이고 전체자료 (x,z)의 로그 우도는 $l(\theta;x,z)$ 로 나타낼 때, 다음은 일반적인 EM알고리즘이다.

1. 모수를 초기값 $\hat{\theta}(0)$ 으로 초기화 한다.

2. 기댓값 단계(Expectation step)
   j번째 단계에서 관측된 자료값과 전단계의 모수 추정값 $\hat{\theta}(j)$ 이 주어졌을 때, 다음과 같은 전체자료의 로그 우도의 기댓값을 구한다.
$$Q(\theta',\hat{\theta}(j)) = E\big(l(\theta';x,z)|x,\hat{\theta}(j)\big).$$

3. 최대화 단계(Maximization step)
$$\hat{\theta}(j+1)=\arg\max_{\theta} Q(\theta',\hat{\theta}(j)).$$

4. 단계 2와 3을 모수추정값이 수렴할 때까지 반복한다.

가우시안 혼합 모형의 EM 알고리즘에 대하여, $x=(x_1, x_2, \cdots, x_n)$ 은 R개의 성분으로 이루어진 p차원 가우스 혼합분포를 따른다. 또한, $z=(z_1,z_2,\cdots,z_n)$ 는 각 관측값이 어느 정규분포로부터 온 것인지를 나타내는 잠재변수라고 할 경우, 식 (2)와 같이 표현할 수 있다.

$$X_i|Z_i=r\sim N(\mu_r,\Sigma_r). \tag{2}$$

$P(Z_i=r)=\tau_r\in(0,1)$ 이고, $\sum_{r=1}^{R}\tau_r=1$ 이며, EM 알고리즘을 통해 추정할 모수 $\theta=(\tau, \mu_1, \mu_2, \cdots, \mu_R, \Sigma_1, \Sigma_2, \cdots, \Sigma_R)$ 이며 전체 데이터의 우도함수는 식 (3)과 같다.

$$L(\theta;x,z)=\prod_{i=1}^{n}\tau_{zi}\Phi(x_i,\mu_{zi},\Sigma_{zi}). \tag{3}$$

로그 우도함수는 식 (4)와 같다.

$$\sum_{i=1}^{n}\sum_{r=1}^{R} I(z_i=r)[\log\tau_r-\frac{1}{2}\log|\Sigma_r|-\frac{1}{2}(x_i-\mu_r)^T\Sigma_r^{-1}(x_i-\mu_r)-\frac{p}{2}\log(2\pi)]. \tag{4}$$

현재의 추정값 $\hat{\theta}(j)$가 주어졌을 때, Z_i의 조건부 분포는 베이즈 정리에 의해 식 (5)와 같이 나타낼 수 있다.

$$\widehat{\gamma_{r,i}}(j)=P(Z_i=r|X_i=x_i;\hat{\theta}(j))=\frac{\hat{\tau_r}(j)\Phi(x_i;\widehat{\mu_r}(j),\widehat{\Sigma_r}(j))}{\sum_{r=1}^{R}\hat{\tau_r}(j)\Phi(x_i;\widehat{\mu_r}(j),\widehat{\Sigma_r}(j))}. \tag{5}$$

다음으로, 기댓값 단계는 식 (6)와 같이 된다.

$$Q(\theta|\theta(j)) = \sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{\gamma_{r,i}}(j)[\log\tau_r - \frac{1}{2}\log|\Sigma_r| - \frac{1}{2}(x_i-\mu_r)^T\Sigma_r^{-1}(x_i-\mu_r) - \frac{p}{2}\log(2\pi)]. \tag{6}$$

최대화 단계에서, 먼저 τ의 경우 $\tau_1+\tau_2+\cdots+\tau_R=1$과 $0 \le \tau_r \le 1$, $r=1,2,\cdots,R$의 제약조건에서 식 (7)를 통해 구할 수 있다.

$$\hat{\tau}(j+1) = \arg\max_{\tau}(\sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{\gamma_{r,i}}(j)\log\tau_r). \tag{7}$$

라그랑즈 승수법에 의해 $\hat{\tau}(j+1)$은 식 (8)과 식 (9)의 해가 된다.

$$\frac{\partial}{\partial\tau_r}(Q-\lambda\sum_{s\neq r}\tau_s) = 0,\ j=1,2,\cdots,R \tag{8}$$

$$\frac{\partial}{\partial\lambda}(Q-\lambda\sum_{r=1}^{R}\tau_r - 1) = 1-\sum_{r=1}^{R}\tau_r = 0. \tag{9}$$

따라서, 식 (10)로 주어진다.

$$\hat{\tau_r}(j+1) = \frac{\sum_{i=1}^{n}\widehat{\gamma_{r,i}}(j)}{\sum_{i=1}^{n}\sum_{r=1}^{R}\widehat{\gamma_{r,i}}(j)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\gamma_{r,i}}(j). \tag{10}$$

다음으로, (μ_r, Σ_r)의 추정값은 식 (11)로 표현할 수 있다.

$$\left(\widehat{\mu_r}(j+1), \widehat{\Sigma_r}(j+1)\right) = \arg\max_{\mu_r,\Sigma_r} Q(\theta|\theta(j)). \tag{11}$$

μ_r과 Σ_r에 대한 편도함수로부터 식 (12)와 식 (13)을 얻는다.

$$\widehat{\mu_r}(j+1) = \frac{\sum_{i=1}^{n}\widehat{\gamma_{r,i}}(j)x_i}{\sum_{i=1}^{n}\widehat{\gamma_{r,i}}(j)}. \tag{12}$$

$$\widehat{\Sigma_r}(j+1) = \frac{\sum_{i=1}^{n}\widehat{\gamma_{r,i}}(j)(x_i-\widehat{\mu_r}(j+1))(x_i-\widehat{\mu_r}(j+1))^T}{\sum_{i=1}^{n}\widehat{\gamma_{r,i}}(j)}. \tag{13}$$

따라서 적절한 초기치를 설정한 후, 식 (12)와 식 (13)을 반복하여 추정값을 얻을 수 있다. 최적의 군집의 개수는 군집수에 따른 Bayesian information criterion(BIC)를 계산하여, 가장 작은 BIC를 갖는 군집수로 선정한다.