Information Systems Review
Vol. 24, No. 4 November 2022 http://dx.doi.org/10.14329/isr.2022.24.4.023

비대칭 마진 SVM 최적화 모델을 이용한 기업부실 예측모형의 범주 불균형 문제 해결

Optimization of Uneven Margin SVM to Solve Class Imbalance in Bankruptcy Prediction

조 성 임 (Sung Yim Jo) 부산대학교 경영대학 박사과정
김 명 종 (Myoung Jong Kim) 부산대학교 경영대학 교수, 교신저자

요 약

Support Vector Machine(SVM)은 기업부실 예측문제 등 다양한 분야에서 성공적으로 활용되어 왔으나 범주 불균형 문제가 존재하는 경우 다수 범주의 경계영역은 확장되는 반면, 소수 범주의 경계영역은 축소되고 분류 경계선이 소수 범주로 편향되어 분류 성과에 부정적인 영향을 미치는 것으로 보고되고 있다. 본 연구는 범주 불균형 문제에 대한 대칭 마진 SVM(EMSVM)의 한계점을 개선하기 위하여 비대칭 마진 SVM(UMSVM)과 임계점 이동 기법을 결합한 최적화 비대칭 마진 SVM인 OPT-UMSVM을 제안한다. OPT-UMSVM은 소수 범주 방향으로 치우진 분류 경계선을 다수 범주로 재이동함으로써 소수 범주의 민감도를 개선하고 최적화된 분류 성과를 산출함으로써 SVM의 일반화 능력을 향상시키는 장점을 가진다. OPT-UMSVM의 성과 개선 효과를 검증하기 위하여 불균형 비율이 상이한 5개의 표본군을 구성하여 10-fold 교차타당성 검증을 수행한 결과는 다음과 같다. 첫째, 범주 불균형이 미미한 표본에서 UMSVM은 EMSVM의 성과 개선 효과가 미약한 반면, 범주 불균형이 심화된 표본에서 UMSVM은 EMSVM의 성과개선에 크게 공헌하고 있다. 둘째, OPT-UMSVM은 EMSVM 및 기존의 UMSVM과 비교하여 범주 균형 및 범주 불균형 표본 모두에서 보다 우수한 성과를 가지고 있으며, 특히 범주 불균형이 심화된 표본에서 유의적인 성과 차이를 보였다.

키워드 : 비대칭 마진 SVM, SVM, 범주 불균형, 부도예측, OPT-UMSVM

Ⅰ. 서 론

기업부실은 주주, 채권자, 종업원 및 거래처 등 기업의 직접적인 이해관계자들에게 상당한 비용을 부담시킬 뿐만 아니라, 1997년 대우, 한보 등과 같은 대기업의 부실은 금융기관 부실과 연계되어 국가경제 전체의 위기로 파급되어 막대한 사회적 손실을 초래한다. 이러한 의미에서 기업의 부실 가능성을 사전에 예측하고 부실의 파급효과를 최소화하기 위한 정교한 기업부실 예측모형의 개발은 재무 및 회계학 분야의 오랜 연구주제가 되어

† 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(IITP-2022-0-01201).

왔다(Altman, 1968; Barboza *et al.*, 2017; Beaver, 1966). 기업부실 예측모형 개발에 있어 다변량 판별분석 및 로지스틱 분석과 같은 통계분석을 비롯하여 최근에는 의사결정트리, 인공신경망 및 서포트벡터머신(Support Vector Machine, SVM) 등과 같은 다양한 머신러닝 기법을 적용하여 기업부실 예측모형의 정확도를 개선하기 위한 노력들이 지속되고 있다(Chen, 2011; Horak *et al.*, 2020; Olson *et al.*, 2012).

특히, SVM은 범주 간 서포트 벡터들(support vectors)의 최단 거리인 마진(margin)을 최대화하는 최대마진이론(maximum margin theory)에 기반하고 있기 때문에 분류 오류를 최소화하면서 범주 간 경계영역(decision boundary between classes)을 가장 선명하게 분리시키는 분류 경계선을 효율적으로 탐색할 수 있을 뿐만 아니라, 구조화된 정규화 모수(regularization parameter)를 적용하여 학습되지 않은 새로운 데이터에 대해서도 높은 분류 정확성을 보이는 일반화 능력이 탁월하여 분류 및 예측 문제에서 우수한 성과를 보여주었다(Burges, 1998; Cortes *et al.*, 1995; Shawe-Taylor, 1998; Vapnik, 1999).

이와 같은 분류 모형의 정확성 개선을 위한 노력에도 불구하고, 범주 불균형 문제(class imbalance problem)는 SVM과 같은 분류 및 예측모형의 성과 개선에 부정적인 영향을 미치는 대표적인 데이터 품질 문제로 인식되고 있다. 범주 불균형은 특정 범주의 표본이 다른 범주에 비하여 현저하게 편향되어 있는 데이터 분포의 왜곡 문제로 정의된다. 범주 불균형 문제는 부실예측(Kim *et al.*, 2015; Zhou, 2013), 의료 진단(Mazurowski *et al.*, 2008), 이상거래 인식(Dal Pozzolo *et al.*, 2014), 이미지 인식(Geng *et al.*, 2016), 문서 필터링(Li *et al.*, 2009) 등 다양한 현실 문제에서 빈번하게 관찰되고 있다. 범주 불균형 문제가 존재하는 경우, SVM은 다수 범주(majority class, negative)와 소수 범주(minority class, positive) 간 데이터 불균형 비율(Imbalance Ratio, IR)이 높아질수록 다수 범주의 경계영역이 확장되어 소수 범주의 경계영역을 침투함으로써 소수 범주 경계영역이 축소된다(Feng *et al.*, 2018; Kang and Cho, 2006; Li and Shawe-Taylor, 2003; Veropoulos *et al.*, 1999; Wu and Chang, 2005). 또한 IR이 높아질수록 분류 경계선 역시 소수 범주 방향으로 편향되어 다수 범주의 정확도(특이도, Specificity)는 완만하게 증가하지만, 소수 범주의 정확성(민감도, Sensitivity)는 급격하게 감소한다. 특히 IR이 극도로 편향되는 경우 소수 범주의 관측치는 모두 다수 범주로 오분류되어 소수 범주에 대한 분류 모형의 기능을 상실하게 된다.

이와 같이 범주 분균형 문제가 SVM 분류 모형의 성과에 미치는 부정적인 영향으로 인하여 많은 선행연구에서는 다양한 기법들을 활용하여 범주 불균형 문제를 해결하기 위한 노력들이 지속되어 왔다. 대표적으로 데이터 샘플링 기법(data sampling techniques)과 알고리즘 수정 기법(algorithm modification techniques)을 SVM과 결합한 하이브리드 모형이 활용되고 있다. 데이터 샘플링 기법은 소수 범주의 표본을 기준으로 다수 범주의 표본을 제거하는 undersampling 기법과 다수 범주의 표본을 기준으로 소수 범주의 표본을 증가시키는 oversampling 기법을 이용하여 불균형 데이터를 균형 데이터로 재구성하여 학습하는 것으로, 기업부실 예측문제와 같은 범주 불균형 문제에 대하여 모형의 예측 성과를 유의적으로 개선하는 것으로 보고되고 있다(Kim and Ahn, 2015; Sundarkumar and Ravi, 2015; Veganzones and Severin, 2018). 그러나 데이터 샘플링 기법은 데이터 전처리 과정이 인위적이며, 데이터의 첨삭에 따라 데이터의 특성이 변화하여 데이터의 고유한 성질이 변형될 수 있다는 단점이 있다(Galar *et al.*, 2012; He and Garcia, 2009).

알고리즘 수정기법은 범주 불균형 문제의 해결을 위하여 SVM 학습 알고리즘의 일부 모듈을 수정하는 방법으로 Cost-sensitive SVM과 비대칭 마진 SVM(Uneven Margin SVM, UMSVM)이 대표적으로 활용되고 있다. Cost-sensitive SVM은 다수 범

주와 비교하여 소수 범주 표본에 높은 가중치를 부여하여 소수 범주 표본의 손실계수를 증가시켜 소수 범주의 민감도를 개선하는 기법으로 범주 불균형 문제의 성과개선에 성공적으로 적용될 수 있으나, 진실한 손실계수를 측정하기 어렵다는 단점이 있다(Cao *et al.*, 2013; Kim *et al.* 2015). UMSVM은 전통적인 대칭 마진 SVM(Even Margin SVM, EMSVM)이 범주 간 동일 마진을 유지함으로써 범주 불균형 문제에 효과적으로 대처할 수 없다는 한계점을 개선하기 위한 알고리즘 수정 기법이다. UMSVM은 범주 불균형으로 인하여 축소된 소수 범주의 경계영역을 확대하여 소수 범주 방향으로 편향된 분류 경계선을 다수 범주 방향으로 재이동시켜 소수 범주의 민감도를 개선하여 소수 범주에 대한 학습을 강화할 수 있으며, 과적합 문제에 강건하다는 장점이 있다. 이러한 장점으로 인하여 UMSVM은 문서 필터링(Kuspriyanto *et al.*, 2010; Li and Shawe-Taylor, 2003; Li *et al.*, 2005; Li *et al.*, 2009; Ni *et al.*, 2010), 이미지 인식(Geng *et al.*, 2016) 등 자연과학 및 공학 등 다양한 문제에 적용되어 왔지만, 현재까지 기업부실 예측문제 등 비즈니스 분야의 범주 불균형 문제에 적용된 사례는 거의 보고되고 있지 않다. 특히, 현재까지 제안된 UMSVM은 가장 중요한 하이퍼 파라미터(hyper-parameter)로서 다수 범주와 소수 범주의 비대칭 마진 비율을 나타내는 마진 파라미터(margin parameter)를 시행착오에 의한 반복적인 작업을 통하여 탐색하였다(Li and Shawe-Taylor, 2003). 이에 따라 학습과정에 시간이 소요될 뿐만 아니라, 시행착오에 의해 탐색된 마진 파라미터도 학습 성과의 최적화를 보장하지 못한다는 단점이 있다. 이러한 문제점에 대하여 선행 연구에서는 UMSVM과 최적화 알고리즘의의 결합을 통하여 UMSVM이 수정될 필요가 있음을 권고하고 있다(Li and Shawe-Taylor, 2003).

본 연구에서는 기업부실 예측문제에 내재된 범주 불균형 문제에 대하여 UMSVM을 적용함으로써 UMSVM이 비즈니스 분야의 범주 불균형 문제를 효과적으로 해결할 수 있는지 규명하고자 한다. 특히, 본 연구에서는 임계점 이동 기법(threshold moving or post scaling method)을 적용하여 마진 파라미터를 최적화할 수 있는 최적화 UMSVM인 OPT-UMSVM을 제안한다.

본 연구의 실험을 위하여 2015년~2018년의 4개년 동안 500개의 부도 기업과 7,500개의 기업-연도별 정상 기업 표본을 수집하여 IR에 따라 5개의 하위 표본을 구성하였으며, 설명변수로서 7개의 재무비율을 선정하였다. 본 연구에서는 EMSVM에 대한 UMSVM과 OPT-UMSVM의 성과 개선 효과를 측정하기 위하여 각 표본별로 10회의 교차타당성(cross validation) 분석을 수행하였다.

본 연구의 주요 분석 결과는 다음과 같다. 첫째, 범주 불균형이 미약한 표본에서는 EMSVM에 대한 UMSVM의 성과 개선 효과가 크게 나타나지 않은 반면, 범주 불균형이 심화된 표본에서 UMSVM은 EMSVM의 성과 개선에 크게 공헌하고 있음을 확인하였다. 특히, UMSVM의 성과 개선효과는 범주 불균형 비율이 심화됨에 따라 크게 증가하는 것으로 나타나고 있는데 이는 마진의 비대칭성이 증가할수록 다수 범주의 특이도는 완만하게 감소하지만, 소수 범주의 민감도는 급격하게 증가하는 비대칭적인 상충효과에 기인하고 있음을 확인하였다. 둘째, 본 연구에서 제안한 OPT-UMSVM의 성과를 EMSVM 및 기존 UMSVM(Li and Shawe-Taylor, 2003)의 성과와 비교한 결과, OPT-UMSVM은 범주 불균형이 미약한 표본에서는 EMSVM과 UMSVM에 대한 OPT-UMSVM의 성과 개선 효과가 유의적이지 못하지만, 범주 불균형이 심화된 표본에서는 EMSVM과 UMSVM의 성과 개선에 유의적으로 공헌하고 있음을 발견하였다.

본 연구의 공헌점은 다음과 같다. 첫째, 범주 균형 및 불균형 등 다양한 데이터 분포에 대하여 UMSVM의 적용효과를 실증함으로써 비즈니스 영역에서의 UMSVM의 성과 개선 효과를 분석하였다. 특히, 비즈니스 분야의 대표적인 범주 불균형 문제인 기업부실 예측문제를 대상으로 UMSVM이

성공적으로 적용될 수 있음을 실증함으로써 비즈니스 분야에서 UMSVM의 도입 타당성에 대한 실증적 자료를 제공하고 있다. 둘째, UMSVM과 최적화 기법을 결합한 OPT-UMSVM을 활용하여 학습시간을 단축하고 최적화된 분류 성과를 산출할 수 있는 학습 알고리즘을 제안함으로써 UMSVM의 최적화에 대한 이론적 기반과 실증적 증거를 제시하고 있다.

본 연구는 다음과 같이 구성된다. 제Ⅱ장에서는 범주 불균형 문제가 SVM 분류 모형에 미치는 영향과 범주 불균형 문제를 해결하기 위한 선행연구를 고찰한다. 제Ⅲ장에서는 EMSVM과 UMSVM의 알고리즘 및 UMSVM의 최적화 알고리즘에 대하여 기술하고자 한다. 제Ⅳ장에서는 표본의 수집 및 변수의 선정과정 등 연구 설계에 대하여 설명한다. 제Ⅴ장에서는 국내 기업을 대상으로 실증한 연구결과를 제시하고 마지막으로 제Ⅵ장에서는 본 연구의 결론을 요약하고 향후 연구방향에 대하여 소개하고자 한다.

Ⅱ. 범주 불균형 문제 및 해결 방안

2.1 범주 불균형 문제

범주 불균형 문제는 분류 모형에서 다수 범주와 소수 범주의 범주 간 경계영역을 왜곡시켜 SVM 등 분류 모형의 예측 성과를 저하시키는 것으로 보고되고 있다. Kang and Cho(2006)은 소수 범주와 다수 범주의 불균형 비율이 크지 않은 경우에는 EMSVM의 범주 간 경계영역은 유사하게 설정되지만, 범주 불균형이 심해질수록 다수 범주가 소수 범주의 경계영역을 침범하게 되어 소수 범주의 경계영역이 점차적으로 축소되는 것으로 보고하였다. Veropoulos *et al.*(1999)과 Li and Shawe-Taylor(2003)는 데이터의 범주 불균형 문제가 심각할수록 소수 범주의 영역이 축소되어 분류 경계선이 소수 범주 방향으로 편향되며, 결과적으로 EMSVM의 성과가 저하되는 것으로 보고하고 있다.

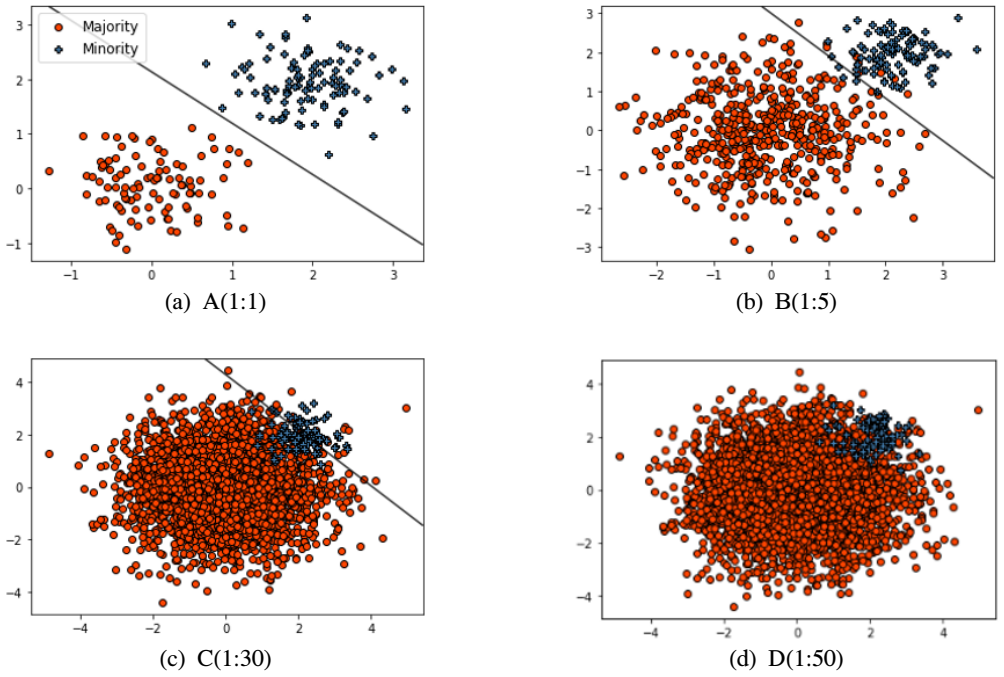
본 연구에서는 범주 불균형 문제가 EMSVM의 분류 성과에 미치는 영향을 분석하기 위하여 다양한 IR을 적용한 인공 데이터를 생성하였다. 인공 데이터에서 소수 범주의 표본 수는 100개이며 이를 기준으로 데이터 불균형 비율을 각각 1:1, 1:5, 1:30, 1:50로 하여 <표 1>과 같이 구성하였다.

〈표 1〉 인공 데이터의 표본 구성[1)]

DataSets(IR)	Minority	Majority	Total
A (1:1)	100	100	200
B (1:5)	100	500	600
C (1:30)	100	3,000	3,100
D (1:50)	100	5,000	5,100

<표 1>의 데이터 표본에 대하여 EMSVM을 이용하여 분류 경계선의 이동과정을 관찰한 결과는 <그림 1>에 제시되어 있다. <그림 1>에서 실선으로 나타난 부분이 범주 간의 분류 경계선이다. 범주간 균형인 데이터 A(1:1)의 경우(<그림 1>의 (a)), 두 범주 간 영역의 크기가 유사하게 나타나고 두 범주 사이의 경계가 선명함을 알 수 있다. 그러나 불균형이 심해진 B(1:5) 및 C(1:30)에서는 다수 범주의 경계영역이 소수 범주의 경계영역을 침투하여 다수 범주의 경계영역은 점점 확대되는 반면, 소수 범주의 영역이 점점 축소되고 있으며 분류 경계선 역시 소수 범주 방향으로 치우치는 것을 확인할 수 있다(<그림 1>의 (b) 및 (c)). 범주 불균형이 극단적으로 심화된 D(1:50)에서는 다수 범주가 소수 범주의 전 영역에 침투하여 데이터 영역 전체가 다수 범주의 영역으로 분류되고 소수 범주의 영역이 완전히 소멸되어 범주 간의 분류 경계선이 설정되지 않는 것으로 분석되었다(<그림 1>의 (d)).

1) 가우시안 정규분포로 인공 데이터를 생성하는 “scikit-learn” 라이브러리의 “make_bolds” 함수를 활용하였다.



(a) A(1:1) (b) B(1:5)

(c) C(1:30) (d) D(1:50)

〈그림 1〉 SVM을 이용한 범주 불균형 인공 데이터에 대한 분류 결과

범주 불균형 문제에서 파생되는 또 다른 문제는 분류 모형의 성과 측정치와 관련되어 있다. 현재 가장 보편적으로 사용되는 성과 측정치는 정확도(Arithmetic Accuracy) 이지만 정확도는 범주 불균형 문제에서 다수 범주의 특이도(Specificity)에 크게 의존하며, 소수 범주의 민감도(Sensitivity)를 고려하지 못하는 단점이 있다. 성과 측정치로서 정확도가 다수 범주의 특이도와 소수 범주의 민감도를 균형적으로 고려하지 못한다는 문제점을 해결하기 위하여 기하평균 정확도(Geometric Mean Accuracy) 및 AUC(Area Under ROC Curve)와 같은 대체적인 성과지표들이 제안되어 왔다(Du *et al.*, 2017; Kubat *et al.*, 1997). <표 2>는 이진 분류 모형의 정오분류표와 모형의 성과 측정에 사용되는 특이도, 민감도, 정확도 및 기하평균 정확도의 산출 방법에 대하여 간략히 설명하고 있다.

<표 3>에서는 <표 1>의 인공 데이터의 범주 불균형에 따른 EMSVM의 분류 성과를 제시하고 있

〈표 2〉 정오분류표와 다양한 성능 평가 지표

		Prediction	
		Positive	Negative
Real	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

특이도(Specificity, SPE): TN / (FP + TN)
민감도(Sensitivity, SEN): TP / (TP + FN)
정확도(Arithmetic Accuracy or accuracy, ACC): (TP + TN) / (TP + FN + FP + TN)
기하평균 정확도(Geometric Mean, GM) = $\sqrt{\text{민감도}\times\text{특이도}}$

다. 먼저, 특이도(SPE)는 IR의 변화에도 불구하고 큰 변화가 나타나지 않는 반면, 민감도(SEN)는 IR이 증가함에 따라 급격하게 감소하는 추이를 보이고 있다. 이에 따라 특이도에 크게 의존하는 정확도(ACC)는 완만하게 증가하는 반면, 특이도와 민감도를 동시에 고려하는 기하평균 정확도(GM)는 IR이 증가할수록 급격하게 감소하고 극단적인 불균형 범주 데이터(D(1:50))에서 GM은 0이 되는데 이는 분류 모형이 소수 범주 분류에 대한 기능을 상실하였음을 의미한다.

〈표 3〉 IR에 따른 SVM의 정확도 비교

DataSets(IR)	SPE	SEN	ACC	GM
A (1:1)	0.98	0.98	0.97	0.98
B (1:5)	0.99	0.92	0.98	0.95
C (1:30)	1.00	0.43	0.98	0.66
D (1:50)	1.00	0.00	0.98	0.00

주) 특이도(SPE), 민감도(SEN), 정확도(ACC) 및 기하평균 정확도(GM).

인공 데이터를 대상으로 한 실험에서 범주 불균형 문제가 EMSVM에 미치는 효과를 분석한 결과를 간략하게 요약하면, 첫째, 범주 불균형 문제가 심화될수록 다수 범주의 경계영역은 확장되는 반면 소수 범주의 경계영역은 급속하게 축소된다. 둘째, 범주 불균형 문제가 심화될수록 분류 모형의 분류 경계선은 소수 범주로 편향된다. 마지막으로 범주 불균형 문제가 심화될수록 분류 모형의 예측 성과는 급속하게 저하된다.

2.2 범주 불균형 문제의 해결 방안

범주 불균형 문제에 대하여 EMSVM의 성과개선을 위하여 보편적으로 활용되어 왔던 방법은 데이터 샘플링 기법과 알고리즘 수정 기법이다. 데이터 샘플링 기법은 학습 데이터의 범주간 분포 비율을 균형적으로 조정함으로써 학습 성과 저하의 문제를 효과적으로 해결하는 것으로 보고되고 있다(Kim and Ahn, 2015; Sundarkumar and Ravi, 2015; Veganzones and Severin, 2018). 그러나 데이터의 균형을 맞추기 위한 데이터 전처리 과정이 인위적이며, 그 결과로 인해 정보 손실이나 생성 데이터의 중복으로 인한 데이터의 품질 저하 문제가 제기되고 있다(Galar *et al.*, 2012; He and Garcia, 2009).

알고리즘 수정기법은 데이터에 대한 인위적인 조작없이 SVM의 일부 알고리즘을 수정하여 분류 경계선을 다수 범주 방향으로 재이동시켜 소수 범주의 경계영역을 확보하고 소수 범주의 민감도를 개선시키는 기법으로 Cost-sensitive SVM(Veropoulos *et al.*, 1999)과 UMSVM(Li and Shawe-Taylor, 2003)이 대표적으로 활용되고 있다. Cost-sensitive SVM은 분류 경계선을 벗어난 관측치(여유 변수, slack variable)에 대한 손실계수(penalty parameter)를 정의할 때, 일반적인 SVM이 범주별로 동일한 손실계수를 적용하는 것과는 달리 Cost-sensitive SVM은 소수 범주의 손실계수를 다수 범주의 손실계수 보다 큰 값을 부여함으로써 소수 범주의 데이터의 오분류율을 낮추도록 유도한다(Cao *et al.*, 2013; Tao *et al.*, 2019; Veropoulos *et al.*, 1999). Cost-sensitive SVM은 이해하기 쉽고 간편하게 소수 범주의 민감도를 높일 수 있어 자주 활용되고 있지만 여유 변수의 수에 따라 마진의 크기가 급변하므로 분류 경계선의 이동을 세밀하게 조정할 수 없으며, 현실 문제에서 손실계수의 상대적인 비율을 정확하게 측정하기 어렵다는 단점으로 인하여 과적합의 문제가 발생할 수 있다(Elrahman and Abraham, 2013; Wu and Chang, 2005; Yan *et al.*, 2017).

UMSVM은 소수 범주에 편향된 분류 경계선을 다수 범주로 재이동시키고자 EMSVM의 마진 내에서 각 범주와 분류 경계의 거리를 비대칭으로 적용하는 기법이다(Li and Shawe-Taylor, 2003). EMSVM은 범주 간 동일 마진을 유지하기 때문에 범주 불균형 문제에서 다수 범주 경계영역의 확장과 소수 범주 경계영역의 축소 문제에 대하여 효과적으로 대처하지 못하지만, UMSVM은 두 범주의 마진간 비대칭 비율의 조절을 통하여 분류 경계선을 직관적이면서도 쉽게 이동시키고, 이동거

리를 세밀하게 조절할 수 있다는 장점으로 인하여 문서 필터링 및 이미지 인식 등의 문제에 다양하게 적용되고 있다(Geng *et al.*, 2016; Kuspriyanto *et al.*, 2010; Li *et al.*, 2005; Li *et al.*, 2009; Ni *et al.*, 2010).

Ⅲ. UMSVM의 최적화 알고리즘

3.1 EMSVM과 UMSVM 알고리즘

n개의 관측치로 구성된 학습표본 $S=\{(x_i, y_i): i=1, \cdots, n\}$를 가정하자. 여기에서 $x_i \in R^d$는 d차원의 입력 변수 벡터(독립 변수)이고, $y_i \in (-1, +1)$는 이진 범주 레이블(종속 변수)이다. x를 소수 범주($y=+1$)의 입력 변수 벡터 x^+와 다수 범주($y=-1$)의 입력 변수 벡터 x^-로 구분하였을 때 EMSVM의 분류 경계선($f(x)$)과 소수 범주의 지지선($f(x^+)$) 및 다수 범주의 지지선($f(x^-)$)은 다음과 같이 식 (1)의 각각의 항목으로 정의된다.

$$\begin{aligned} f(x) &= < w, x > + b = 0 \\ f(x^+) &= < w, x^+ > + b = +1 \\ f(x^-) &= < w, x^- > + b = -1 \end{aligned} \quad (1)$$

EMSVM은 각 마진의 경계 지지선인 $f(x^+)$와 $f(x^-)$ 사이의 수직 거리(margin, $\frac{2}{|w|}$)를 최대화하는 것으로 이를 w의 내적을 나타내는 $< w, w >$의 최소화로 대체하고, 여유변수(ξ)와 여유변수에 적용되는 손실계수를 C라 할 때 EMSVM의 목적식과 제한식은 식 (2)와 같이 정의된다(Cortes *et al.*, 1995).

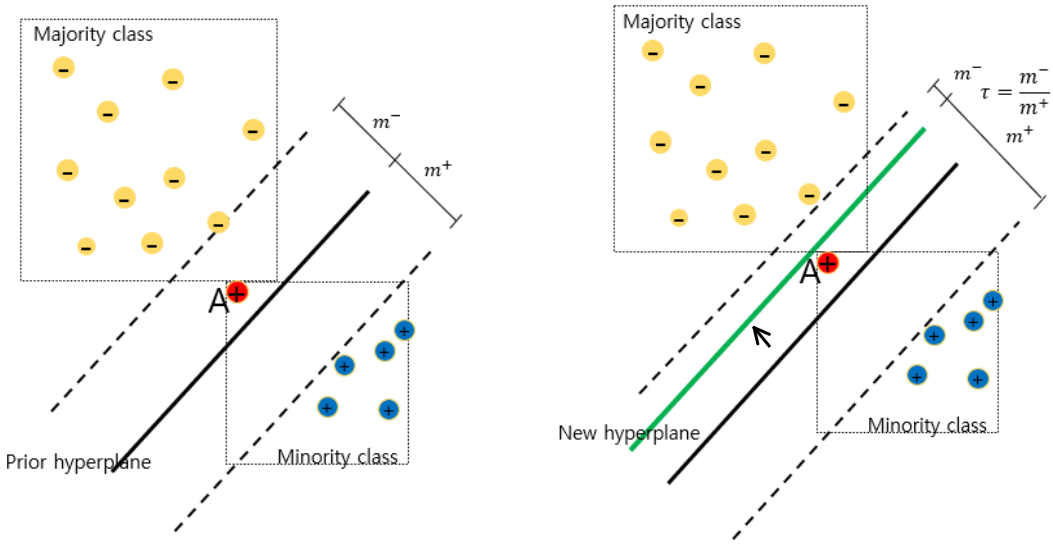
$$\begin{aligned} & function \quad \min < w, w > + C\textstyle\sum_{i=1}^{n} \xi & ① \\ & constraints \; < w, x^+ > + b \geq +1 - \xi \; if \; y_i = +1 & ② \\ & \qquad\qquad < w, x^- > + b \leq -1 + \xi \; if \; y_i = -1 & ③ \\ & \qquad\qquad \xi \geq 0 \qquad i = 1, \; ..., \; n \end{aligned} \quad (2)$$

범주 불균형 문제가 존재하는 경우 EMSVM은 다수 범주의 경계영역이 소수 범주의 경계영역을 침투하여 소수 범주의 경계영역이 축소하게 된다. 결과적으로 다수 범주의 경계영역이 확대됨에 따라 특이도는 다소 높아지지만 소수 범주의 경계영역은 급격하게 축소되어 민감도가 낮아지게 된다. UMSVM은 소수 범주의 마진을 다수 범주의 마진에 비하여 상대적으로 확장하여 민감도를 높이는 것을 목적으로 한다. 소수 범주의 마진(positive margin, m^+)과 다수 범주의 마진(negative margin, m^-)의 상대적인 비율인 $\frac{m^-}{m^+}$를 마진 파라미터(τ)라 할 때 UMSVM의 목적식과 제한식은 식 (3)과 정의된다.

$$\begin{aligned} & function \; \min < w, w > + C\textstyle\sum_{i=1}^{n} \xi & ① \\ & constraints \; < w, x^+ > + b \geq +1 - \xi \; if \; y_i = +1 & ② \\ & \qquad\qquad < w, x^- > + b \leq -\tau + \xi \; if \; y_i = -1 & ③ \\ & \qquad\qquad \xi \geq 0 \qquad i = 1, \; \cdots, \; n \\ & \qquad\qquad 0 \leq \tau \leq 1 \end{aligned} \quad (3)$$

식 (2)와 식 (3)을 비교하면 EMSVM과 UMSVM은 목적식과 소수 범주에 대한 제한식은 동일하게 유지되는 반면, 다수 범주에 적용되는 제한식이 변형된다. 식 (2)의 ③와 식 (3)의 ③을 비교하면 EMSVM의 다수 범주의 지지선은 $f(x^-)=-1$로 설정된 반면, UMSVM에서는 $f(x^-)=-\tau (0 \leq \tau \leq 1)$로 대체된다.

<그림 2>에는 EMSVM(a)과 UMSVM(b)의 분류결과를 간략하게 도시하였다. 여기에서 다수 범주의 관측치는 (-)이며 소수 범주의 관측치는 (+)로 도시되어 있다. 또한, 실선은 EMSVM과 UMSVM의 분류 경계선을 나타내며, 점선은 다수 범주와 소수 범주의 지지선을 의미한다. EMSVM에서 소수 범주와 다수 범주의 마진은 <그림 2>의 (a)와 같이 대칭적으로 동일하게 나타난다($m^+ = m^-$). <그림 2>의 (a)에서는 확장된 다수 범주 데이터의 영향으로 소수 범주 방향으로 편향된 EMSVM의 분류 경계선에 의하여 소수 범주의 관측치들이 다



(a) EMSVM (b) UMSVM

〈그림 2〉 UMSVM의 분류 경계선의 이동과 정확도 향상

수 범주로 오분류되는 가능성이 증가하게 된다. 예를 들어보면 관측치 A는 실제로는 소수 범주의 관측치이지만 EMSVM의 분류 경계선(prior hyperplane)을 기준으로 다수 범주로 오분류된 사례이다.

UMSVM은 <그림 2>의 (b)와 같이 분류 경계선을 결정하는 마진 파라미터인 τ를 이용하여 분류 경계선을 다수 범주 방향으로 이동시켜 새로운 분류 경계선(New hyperplane)을 설정하게 된다. τ의 범위는 $0 \leq \tau \leq 1$에서 설정되는데 $\tau = 1$인 경우 UMSVM와 EMSVM의 분류 경계선은 동일하게 유지되며, $\tau = 0$인 경우 분류 경계선은 다수 범주의 지지선까지 이동하게 된다. $0 < \tau < 1$인 경우 분류 경계선은 EVSVM의 분류 경계선과 다수 범주의 지지선 사이에서 위치하게 되며 소수 범주의 마진에 비하여 다수 범주의 마진이 축소되는 비대칭적 마진이 형성된다($m^+ > m^-$). 이와 같이 분류 경계선을 다수 범주 방향으로 이동함으로써 EMSVM에서 오분류된 관측치 A는 소수 범주로 정분류되며, 결과적으로 범주 불균형에서 과소 평가된 소수 범주의 민감도는 분류 경계선이 다수 범주 방향으로 이동함에 따라 소수 범주의 경계 영역이 확장되어 소수 범주의 민감도가 개선되는 효과를 가지게 된다.

UMSVM에서 소수 범주의 민감도 개선효과는 SVM에서 마진이 커질수록 검증 데이터의 일반화 능력이 높아지는 마진의 오차허용능력(error-tolerance ability)에 기초하고 있다(Li and Shawe-Taylor, 2003; Li *et al.*, 2005). Shawe-Taylor(1998)는 마진이 커질수록 SVM의 오분류율이 낮아져 모형의 정확도가 개선될 수 있음을 증명하였으며, 이 이론에 따라 범주 불균형 문제에 대하여 UMSVM은 소수 범주의 마진을 확장함으로써 소수 범주의 오차허용능력을 증가시켜 소수 범주의 오분류율을 낮추고 결과적으로 민감도를 높이는 역할을 하는 것으로 설명될 수 있다.

EMSVM의 분류 경계선에서 UMSVM의 분류 경계선으로 평행 이동할 때 두 분류 경계선은 식 (4) 및 식 (5)와 같은 관계성을 가지게 된다(Li and Shawe-Taylor, 2003; Li *et al.*, 2008).

$$\omega_2^* = \omega_1^* \tag{4}$$

$$b_2^* = b_1^* + \frac{1-\tau}{1+\tau} \tag{5}$$

먼저, 식 (4)에서 EMSVM의 weight(ω_1^*)는 UMSVM의 weight(ω_2^*)와 동일한 값을 가지게 되는데 이는 EMSVM과 UMSVM의 분류 경계선의 기울기가 동일하게 유지됨을 의미한다. EMSVM과 UMSVM의 분류 경계선 차이는 bias의 차이에 기인하여 나타나는데 식 (5)를 보면 UMSVM의 bias(b_2^*)는 $b_2^* = b_1^* + \frac{1-\tau}{1+\tau}$로서 EMSVM의 분류 경계선이 $\frac{1-\tau}{1+\tau}$만큼 평행 이동하여 UMSVM의 분류 경계선으로 설정됨을 알 수 있다.

3.2 UMSVM의 최적화 알고리즘

분류 경계선의 이동을 통하여 소수 범주의 민감도를 개선하는 UMSVM의 일반화 능력을 최대화하기 위해서는 EMSVM에서 설정된 분류 경계선을 UMSVM에서 이동시키는 거리를 결정하는 τ가 최적화되어야 한다. Li and Shawe-Taylor(2003)는 0과 1 사이의 임의의 실수 값 중에서 τ을 선택하여 최적의 결과를 도출하는 Random search 기법을 적용한 결과, 적정 τ은 소수 범주의 데이터의 숫자가 많을수록 1에 근접하는 경향을 보여 적정 τ가 표본의 불균형 비율과 관계되어 있음을 분석하였으나, τ와 데이터 불균형 비율 사이의 일관된 수학적 함수 관계를 발견하지 못하였다. Li *et al.*(2005)은 퍼셉트론(Perceptron) 알고리즘을 UMSVM과 합성하여 적정 τ을 찾고자 하였으나 이 기법은 퍼셉트론의 특성으로 선형 분리가 가능한 표본에 한하여만 적용할 수 있다는 한계점을 가지고 있다.

본 연구에서는 최적의 임계점을 설정하기 위하여 임계점 이동기법을 활용하였다. 임계점 이동기법은 데이터 불균형 문제의 해결을 위한 알고리즘 수정 기법의 하나로서 인공신경망이나 SVM과 같이 분류경계선을 수학적 함수모형으로 설정하는 분류 모형에 자주 적용된다. 이러한 분류 모형에서 범주 레이블을 결정하는 임계점은 일반적으로 0으로 설정되는데, 범주 불균형 문제에 대하여 임계점을 0으로 고정하는 경우 다수 범주의 특이도는 과대평가되고 소수 범주의 민감도는 과소평가되는 문제가 발생하게 된다. 이러한 문제에 대하여 임계점 이동 기법은 임계점을 0으로 고정하지 않고 데이터 불균형 비율에 따라 임계점을 유동적으로 이동시켜 재설정함으로써 데이터 불균형 분류모형의 성과를 개선한다(Du *et al.*, 2017; Zou *et al.*, 2016; 윤우섭, 김명종, 2021).

본 연구는 UMSVM의 알고리즘에 임계점 이동 기법을 적용하여 소수 범주의 민감도와 다수 범주의 특이도 차이가 가장 적은 지점을 임계점으로 설정하여 기하평균 정확도를 최적화하는 OPT-UMSVM을 제안한다. OPT-UMSVM은 임의로 다수의 임계점을 설정한 후 임계점별로 특이도와 민감도를 계산한 후 특이도와 민감도의 차이가 가장 적은 임계점을 기하평균 정확도를 최적화는 최적의 임계점으로 탐색한다.

UMSVM에서 분류 경계선의 이동이 τ에 의해 결정되는 점을 활용하여 임계점 이동 기법을 적용한 결과, 식 (5)에서 EMSVM의 분류 경계선을 $\frac{1-\tau}{1+\tau}$만큼 평행 이동함으로써 UMSVM에서 최적의 분류 경계선이 설정되는 것을 확인하였다. 평행 이동 전의 EMSVM의 분류 경계선의 결과값(output value)은 0이라는 점을 고려할 때, 평행 이동 후의 UMSVM의 분류 경계선의 결과값은 $\frac{1-\tau}{1+\tau}$이 된다. 본 연구에서는 평행 이동 거리($\frac{1-\tau}{1+\tau}$)를 새로운 임계점 T로 설정하고 이때의 τ가 최적의 마진 파라미터 τ^*가 되며 이 지점에서 소수 범주의 민감도와 다수 범주의 특이도가 근사하게 일치하게 된다. 이를 기준으로 각 모형의 분류 경계선에 대한 관계식을 살펴보면 식 (6)과 같다.

EMSVM 분류 경계선: $< \omega_1^*, x > + b_1^* = 0$ (6)

UMSVM 분류 경계선:

$$< \omega_2^*, x > + b_2^* = < \omega_1^*, x > + b_1^* + \frac{1-\tau}{1+\tau}$$

OPT-UMSVM 분류 경계선:

$$< \omega_3^*, x > + b_3^* = < \omega_1^*, x > + b_1^* + \frac{1-\tau^*}{1+\tau^*}$$
$$= < \omega_1^*, x > + b_1^* + T$$

이제 OPT-UMSVM의 목적식은 UMSVM의 목적식을 만족하는 하이퍼 파라미터 τ를 결정하는 $\arg min$ 함수로 치환될 수 있다.

$$function \ \arg\min_\tau < w, w > + C\sum_{i=1}^{n} \xi \qquad (7)$$
$$constraints \ < w, x^+ > + b \geq +1 - \xi \ \ if \ y_i = +1$$
$$< w, x^- > + b \leq -\tau + \xi \ \ if \ y_i = -1$$
$$\xi \geq 0 \qquad i = 1, \cdots, n$$
$$0 \leq \tau \leq 1$$

Ⅳ. 연구 설계

4.1 표본 수집

본 연구의 표본은 2015부터 2018년까지 4개년 동안에 국내 비금융업 외부 회계감사 법인을 대상으로 수집하였다. 부실기업은 국내 시중은행의 자료를 기초로 500개 부실기업을 선정하였으며, 부실기업에 포함되지 않은 7,500개 기업을 정상 기업으로 선정하여 총 8,000개의 표본을 구성하였다. 표본의 부실율은 6% 내외로 국내 신용평가 전문기관에서 추정한 장기 평균 부도율인 3~6%의 범위 내에 해당한다. 본 연구에서는 불균형 비율에 따라 UMSVM이 EMSVM의 성과 개선에 미치는 효과를 확인하기 위하여 부실기업을 기준으로 불균형 비율이 상이한 5개의 하위 표본군을 구성하였으며, 본 연구에서는 10-flod 교차타당성 분석을 검증 절차로 사용하기 때문에 각 하위 표본별로 학습 표본과 검증 표본을 <표 4>와 같이 구성하였다.

4.2 변수 선정

본 연구에 사용되는 변수를 선정하기 위하여 NICE평가정보에서 제공하는 재무자료를 활용하여 선행연구에서 사용된 비율 및 실무에서 부실예측의 지표로 사용되는 비율을 중심으로 30개의 재무비율을 수집하였다(Altman, 1968; Kim *et al.*, 2015; 박종원, 안종만, 2014). 수집된 재무비율을 7개 재무비율 군(수익성, 부채상환능력, 레버지리, 자본구조, 유동성, 활동성, 규모)으로 분류하였으며, 최종 입력변수는 각 재무비율 군별로 AUC가 가장 높은 7개의 재무비율로 선정하였다(Kim *et al.*, 2015; 윤우섭, 김명종, 2021). <표 5>는 최종 선정된 재무비율에 대하여 AUC, 분산팽창요인(Variance Inflation Factors, VIF) 분석 결과 및 기초통계량을 기술한 것이다. 일반적으로 VIF가 4와 10 사이에 있으면 다중공선성이 민감하고 VIF가 10보다 높으면 매우 심각한 다중공선성이 있다고 본다. 선택한 입력변수의 VIF가 모두 4 미만이므로 실질적인 다중공선성이 나타나지 않음을 보여준다.

〈표 4〉 학습 표본과 검증 표본의 구성

Datasets(IR)	Training set			Test set		
	Normal	Bankrupt	Total	Normal	Bankrupt	Total
A (1:1)	450	450	900	50	50	100
B (1:2)	900	450	1,350	100	50	150
C (1:4)	1,800	450	2,250	200	50	250
D (1:10)	4,500	450	4,950	500	50	550
E (1:15)	6,800	450	7,250	700	50	750

〈표 5〉 재무비율 선정 지표 및 기초 통계량

Group	Variable	AUC	VIF	기초 통계량			
				mean	std	min	max
수익성	총자산경상이익율	54.3	1.32	-14.60	83.92	-893.90	155.07
부채상환능력	EBITA/이자비용	53.1	2.15	151.20	1055.92	3702.10	14327.62
레버리지	자기자본비율	51.7	1.78	118.13	210.09	0.16	3541.66
자본구조	이익잉여금/총자산	51.3	2.52	0.22	0.18	0.00	1.99
유동성	현금비율	48.4	1.36	0.22	0.60	-4.10	5.25
활동성	재고자산회전율	33.4	1.51	4.10	3.59	0.15	38.63
규모	총자산	23.7	1.35	14.90	0.57	7.33	12.59

Ⅴ. 연구 결과

본 연구에서는 입력 변수를 고차원의 입력 벡터로 전환하기 위하여 RBF 커널을 활용하였으며 식(3)을 기준으로 작성된 UMSVM을 기본 학습 알고리즘으로 활용하였다. 하이퍼 파라미터로서 RBF 커널의 분산(γ)과 손실계수(C)는 그리드 탐색(grid search) 기법을 적용하여 <표 6>과 같이 설정하였다.

〈표 6〉 하이퍼 파라미터 설정

DataSets(IR)	C	γ
A (1:1)	1	1
B (1:2)	1	1
C (1:4)	400	1
D (1:10)	400	1
E (1:15)	400	1

<표 7>은 데이터의 불균형 비율을 기준으로 조정한 하위 표본을 대상으로 UMSVM의 τ를 다양하게 조정하여 분류한 결과를 제시하고 있다. UMSVM은 마진의 비대칭성이 증가할수록 소수 범주의 경계영역이 확대되어 소수 범주의 민감도가 개선되므로 범주 불균형 문제를 완화할 수 있다. 마진의 비대칭성을 나타내는 마진 파라미터 τ는 0과 1 사이의 실수 값에서 결정되는데, τ의 값이 1일 경우 UMSVM과 EMSVM은 동일한 분류 경계선을 가지며 동일하게 대칭 마진을 가지게 된다. 반면, τ의 값이 0에 가까울수록 다수 범주의 마진은 축소되고 소수 범주의 마진은 확대되어 마진의 비대칭성이 증가하게 된다.

<표 7>의 분석 결과를 요약하면 첫째, 모든 데이터 표본에서 τ가 낮을수록 특이도는 감소하는 경향을 보이고 있다. 이는 τ가 낮을수록 UMSVM의 분류 경계선의 이동 거리가 커져 다수 범주의 경계영역이 축소되고 이에 따라 다수 범주의 특이도가 감소함을 의미한다. 특히, 주목되는 점은 E(1:15)에서 τ를 1에서 0.25로 조정하여도 특이도는 100%로 동일하며, 0으로 조정한 경우에만 93%로 낮아지는데 이는 E(1:15)의 경우 다수 범주가 대부분의 소수 범주 경계영역을 침범하여 소수 범주의 경계영역이 극단적으로 축소된 상태로서 τ를 일정 구간에서 조정하여도 특이도가 감소하지 않음을 의미한다. 둘째, 모든 데이터 표본에서 민감도는 τ가 낮을수록 크게 개선되고 있으며, 특히 범주 불균형 상황이 심각할수록 그 효과가 크게 나타나는 것으로 분석되었다. 예를 들면, A(1:1)군에서 τ가 1에서 0.001로 낮아짐에 따라 민감도의 증가는 4%p에 제한되는 반면, 범주 불균형이 심화된 D(1:10)과 E(1:15)에서 민감도의 증가는 각각 41%p와 33%p로 민감도 개선효과가 크게 나타나는 것으로 분석되었다. 특히 E(1:15)에서 τ를 1에서 0.5로 조정하여도 민감도는 1%로 미미하게 반응하는데 이는 범주 불균형이 심할 경우 소수 범주의 경계영역이 극단적으로 축소된 상태에서는 τ을 크게 조정하여야 민감도가 개선될 수 있음을 의미한다. 셋째, 정확도는 τ가 작아질수록 지속적

으로 완만하게 감소하는 경향을 보인다. 이는 정확도가 범주 간 정확도를 동시에 고려하지 못하고 다수 범주의 특이도에 크게 의존하고 소수 범주의 민감도에 대해서는 민감하게 반응하지 않는 특성을 가지고 있기 때문이다. 마지막으로, 기하평균 정확도는 데이터 표본별로 일관되지 않은 결과를 보이는데 이는 τ의 증가에 따라 특이도는 감소하는 반면, 민감도는 증가하는 상충효과가 범주 불균형에 따라 상이함을 의미한다. 예를 들어, A(1:1)과 B(1:2)의 범주 불균형이 미약한 데이터 표본에서 특이도의 감소가 민감도의 증가보다 크기 때문에 τ가 작아질수록 기하평균 정확도는 감소하는 경향을 보인다. 반면 C(1:4), D(1:10), E(1:15)에서 τ가 낮아질수록 특이도의 감소보다 민감도의 증가가 크기 때문에 기하평균 정확도는 지속적으로 증가하는 것을 나타났다. 특히 기하평균 정확도 증가의 폭은 불균형 비율이 높을수록 커지는데 τ가 1에서 0.001로 낮아질 때의 그 증가폭은 C(1:4), D(1:10), E(1:15)에서 각각 1%p, 27%p 및 46%p로 증가하는 것으로 나타났다.

<표 7>을 통하여 UMSVM의 범주 불균형 해소 효과는 마진의 비대칭성이 커짐에 따라 감소하는 다수 범주의 특이도와 증가하는 소수 범주의 민감도의 상대적인 개선효과의 차이에 따라 결정된다는 점을 알 수 있다. 특히 소수 범주의 민감도 개선효과가 기하평균 정확도의 개선에 가장 크게 기여하고 있는 점은 범주 불균형 문제에서 소수 범주 데이터에 대한 판별력을 강조한 선행연구의 주장과 일치한다(Kang and Cho, 2006).

본 연구에서 제안한 최적화 알고리즘으로서 OPT-UMSVM을 EMSVM 및 UMSVM(Li and Shawe-Taylor, 2003)의 분류 성과와 비교한 결과는 <표 8>에 제시되어 있다. EMSVM의 τ는 모든 데이터 표본에서 1이며, Li and Shawe-Taylor(2003)에서 적정 τ는 데이터의 불균형 비율에 맞추어 설정되기 때문에 데이터 표본별로 τ은 A(1:1)=1, B(1:2)=0.5, C(1:4)=0.25, D(1:10)=0.1, 및 E(1:15)=0.07이다. 반면 OPT-UMSVM의 τ는 OPT-UMSVM의 최적화 과정에 따라 민감도와 특이도의 차이가 가장 적은 임계점에 도달할 때 기하평균 정확도가 최대화되는 점을 활용하여 τ을 탐색한 결과로서 각 데이터 표본별로 A(1:1)=1, B(1:2)=0.5, C(1:4)=0.1, D(1:10)=0.001, 및 E(1:15)=0.001로 설정되었다. 이렇게 설정된 을 기준으로 세 가지 모델의 정확도

〈표 7〉 τ에 따른 데이터 표본별 분류 성과 비교

SPE						SEN					
DataSets/τ	0.001	0.25	0.5	0.75	1	DataSets/τ	0.001	0.25	0.5	0.75	1
A (1:1)	0.88	0.93	0.96	0.97	0.98	A (1:1)	0.99	0.97	0.96	0.96	0.95
B (1:2)	0.89	0.94	0.96	0.97	0.98	B (1:2)	0.96	0.94	0.94	0.92	0.91
C (1:4)	0.90	0.95	0.96	0.97	0.98	C (1:4)	0.85	0.82	0.81	0.78	0.77
D (1:10)	0.90	0.96	0.98	0.98	0.99	D (1:10)	0.68	0.44	0.34	0.30	0.27
E (1:15)	0.93	1	1	1	1	E (1:15)	0.34	0.03	0.01	0.01	0.01
ACC						GM					
DataSets/τ	0.001	0.25	0.5	0.75	1	DataSets/τ	0.001	0.25	0.5	0.75	1
A (1:1)	0.93	0.95	0.96	0.96	0.96	A (1:1)	0.93	0.95	0.96	0.96	0.96
B (1:2)	0.91	0.94	0.95	0.96	0.95	B (1:2)	0.92	0.94	0.95	0.95	0.94
C (1:4)	0.89	0.91	0.92	0.92	0.92	C (1:4)	0.87	0.87	0.87	0.87	0.86
D (1:10)	0.88	0.92	0.92	0.92	0.92	D (1:10)	0.78	0.65	0.58	0.55	0.51
E (1:15)	0.90	0.94	0.94	0.94	0.94	E (1:15)	0.53	0.16	0.09	0.08	0.07

주) τ=1로 설정된 UMSVM은 EMSVM의 분석결과와 동일함.

〈표 8〉 EMSVM와 UMSVM 및 OPT-UMSVM의 정확도 비교

DataSets(IR)	EMSVM			UMSVM			OPT-UMSVM		
	τ	ACC	GM	τ	ACC	GM	τ	ACC	GM
A (1:1)	1.00	0.96	0.96	1.00	0.96	0.96	1.00	0.96	0.96
B (1:2)	1.00	0.95	0.94	0.50	0.95	0.95	0.50	0.95	0.95
C (1:4)	1.00	0.92	0.86	0.25	0.91	0.87	0.1	0.90	0.88
D (1:10)	1.00	0.92*	0.51*	0.10	0.91*	0.71*	0.001	0.88	0.78
E (1:15)	1.00	0.94*	0.07*	0.07	0.93*	0.24*	0.001	0.90	0.53

주) *는 1% 수준에서 유의.

(ACC)와 기하평균 정확도(GM)를 비교한 결과, 본 연구에서 제안한 OPT-UMSVM은 모든 표본 군에서 EMSVM 및 UMSVM와 비교할 때 같거나 우수한 분류 성과를 보여주고 있으며, 특히 범주 불균형 비율이 증가할수록 EMSVM 및 UMSVM과 비교하여 OPT-UMSVM의 성과차이가 증가하고 있음을 확인하였다.

<표 8>에서 제시한 세 가지 모델의 성과차이의 통계적 유의성을 검정하기 위하여 각각의 정확도별로 Mann-Whitney의 U 검정을 수행하였다. 분석결과, OPT-UMSVM는 EMSVM 및 UMSVM과 비교하여 범주 불균형이 미약한 데이터 표본 A(1:1), B(1:2) 및 C(1:4)에서 일부 개선된 성과를 가지지만 통계적으로 유의적인 차이가 없는 것으로 분석되었다. 반면, 범주 불균형이 심화된 데이터 표본 D(1:10) 및 E(1:15)에서는 1% 수준에서 유의적인 성과 차이를 가지는 것으로 분석되었다. OPT-UMSVM과 UMSVM을 비교 분석의 경우에도 A(1:1), B(1:2) 및 C(1:4)에서 유의적인 성과 차이가 없는 것으로 분석된 반면, D(1:10) 및 E(1:15)에서는 1% 수준에서 유의적인 차이를 가지는 것으로 분석되었다. 비록 <표 8>에 제시하지는 않았지만, UMSVM과 EMSVM 역시 D(1:10) 및 E(1:15)에서는 유의적인 성과 차이를 가지는 것으로 분석되었다. 결과적으로 OPT-UMSVM은 다른 두 모델과 비교하였을 때 가장 우수한 성과를 보이고 있고 이는 OPT-UMSVM가 적정 τ을 효과적으로 탐색한 결과로서 범주 균형 및 범주 불균형 데이터 군 모두에서 강건한 일반화 능력을 확보하고 있음을 의미한다.

다만 OPT-UMSVM은 E(1:15)에서 기하평균 정확도는 0.53으로 다소 낮은 예측 성과를 보여주고 있는데 이는 UMSVM에서 τ는 0과 1사이에서 설정되는 관계로 마진의 이동거리가 크게 제약되는 문제점이 발생하기 때문에 예측 성과가 크게 개선되지 못하였음을 의미한다. 이러한 UMSVM의 구조적 한계점을 보완하기 위해서는 전체 마진을 초과한 범위까지 분류 경계선을 이동시킬 수 있는 Cost-sensitive SVM가 고려될 수 있으나, Cost-sensitive SVM 역시 적절한 손실계수를 선택하기 어렵고 분류 경계선의 이동 범위를 조절하기 어렵다는 단점이 있음을 유의해야 한다.

Ⅵ. 결 론

SVM과 같은 대부분의 분류 모형은 범주 간 데이터가 균형적으로 분포하고 있음을 가정하여 개발되기 때문에 범주 불균형이 존재하는 경우 다수 범주의 경계영역은 확대되고 소수 범주의 경계영역은 축소되는 문제가 발생한다. 이에 따라 범주 간 분류 경계선이 소수 범주로 편향되어 분류 모형의 성과는 현저하게 낮아진다. 범주 불균형 문제가 분류 모형의 성과저하에 미치는 부정적인 영향으로 인하여 범주 불균형 문제의 해결은 기계학습 및 데이터 마이닝 분야에서 주요 연구과제로 인식되어 왔다.

본 연구는 범주 불균형 환경에서 SVM의 분류 성과 개선을 위하여 소수 범주에 편향된 분류 경계선을 다수 범주로 재이동시켜 소수 범주의 민감도를 개선하기 위하여 EMSVM의 알고리즘을 수정한 UMSVM을 기업부실 예측모형에 적용하였다. 분석 결과 UMSVM은 범주별 데이터 분포가 유사한 균형 데이터에서는 EMSVM의 성과 개선 효과에 유의적이지 못하지만, 범주 불균형이 심화된 데이터에서 EMSVM의 성과를 크게 개선함을 확인하였다. 본 연구에서 제안한 OPT-UMSVM은 EMSVM 및 UMSVM과 비교하여 범주 균형 및 불균형 문제 모두에서 보다 우수한 성과를 보이고 있으며, 특히 범주 불균형 문제가 심화될수록 성과 차이가 유의적인 것으로 분석되었다.

본 연구는 경영분야의 범주 불균형 문제에 대한 UMSVM의 적용 가능성을 실증하고 최적화를 통하여 분류 경계선의 최적의 이동거리를 자동 결정하여 SVM의 일반화 능력을 최적화하기 위한 심도 있는 이론적 배경과 실증 자료를 제공함으로써 범주 불균형 문제에 대한 SVM의 강건성을 제고한다는 공헌점이 있다.

그러나, 본 연구의 한계점과 관련하여 다음과 같은 향후 연구방향을 제시하고자 한다. 첫째, 본 연구는 SVM에서 분류 경계선의 섬세한 이동을 통하여 범주 불균형 문제를 해결하고자 하였다. 하지만 불균형 비율이 극히 높은 경우에는 마진의 비대칭 정도를 극단적으로 설정하여 이동 거리를 크게 하더라도 UMSVM의 효과가 미미함을 실증하였고 이를 해결하기 위하여 향후 연구에서는 Cost-sensitive SVM과 UMSVM의 알고리즘을 결합한 새로운 연구를 진행하고자 한다. 둘째, 본 연구는 SVM의 마진을 다수 범주 측면과 소수 범주 측면으로 구분하여 조절한 것으로 이범주 분류에 한하여 적용될 수 있다. UMSVM을 다범주 분류(multi-class classification) 모형에 적용하기 위해서는 전체 마진의 분포를 고려할 필요가 있다. 최근 Zhang and Zhou(2020)의 연구에서는 범주별 평균 마진은 높이는 동시에 분산은 낮추는 방법을 통하여 SVM의 일반화 능력을 개선하기 위한 기법을 보고하였다. 이러한 기법을 본 연구의 OPT-UMSVM과 결합하여 다범주 분류 모형에서 최적화된 SVM에 대한 연구를 진행하고자 한다.

참 고 문 헌

[1] 박종원, 안성만, “재무비율을 이용한 부도예측에 대한 연구: 한국의 외부감사대상기업을 대상으로”, *경영학연구*, 제43권, 제3호, 2014, pp. 639-669

[2] 윤우섭, 김명종, “AUROC기반의 부도예측 앙상블 모형”, *중소기업금융연구*, 2021, 제41권 제3호 pp. 41-60

[3] Abd Elrahman, S. M. and A. Abraham, “A review of class imbalance problem”, *Network and Innovative Computing*, Vol.1, 2013, pp. 332-340.

[4] Altman, E. I., “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, *The Journal of Finance*, Vol.23, No.4, 1968, pp. 589-609.

[5] Barboza, F., H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction”, *Expert Systems with Applications*, Vol.83, 2017, pp. 405-417.

[6] Beaver, W., “Financial ratios as predictors of failure”, *Journal of Accounting Research*, Vol.71, No.4, 1966, pp. 71-111.

[7] Burges, C. J. C., “A tutorial on support vector machines for pattern recognition”, *Data Mining and Knowledge Discovery*, Vol.2, 1998, pp. 121-167.

[8] Cao, J., H. Lu, W. Wang, and J. Wang, “A loan default discrimination model using cost-sensitive support vector machine improved by PSO”, *Information Technology and Management*, Vol.14, No.3, 2013, pp. 193-204.

[9] Chen, M. Y., “Bankruptcy prediction in firms

with statistical and intelligent techniques and a comparison of evolutionary computation approaches", *Computers and Mathematics with Applications*, Vol.62, No.12, 2011, pp. 4514-4524.

[10] Cortes, C., V. Vapnik, and L. Saitta, *Support-Vector Networks Editor*, Machine Learning, Kluwer Academic Publishers, 1995, pp. 273-297.

[11] Dal Pozzolo, A., O. Caelen, Y. A. le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective", *Expert Systems with Applications*, Vol.41, No.10, 2014, pp. 4915-4928.

[12] Davis, J. and M. Goadrich, "The relationship between Precision-Recall and ROC curves", *Proceedings of the 23rd International Conference on Machine Learning - ICML '06. 2*, 2006, pp. 33-240.

[13] Du, J., Vong, C. M., C. M. Pun, P. K. Wong, and W. F. Ip, "Post-boosting of classification boundary for imbalanced data using geometric mean", *Neural Networks*, Vol.96, 2017, pp. 101-114.

[14] Fawcett, T., "An introduction to ROC analysis", *Pattern Recognition Letters*, Vol.27, No.8, 2006, pp. 861-874.

[15] Feng, W., W. Huang, and J. Ren, "Class imbalance ensemble learning based on the margin theory", *Applied Sciences, Switzerland*, Vol.8, No.5, 2018, p. 815.

[16] Galar, M., A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches", *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, Vol.42, No.4, 2012, pp. 463-484

[17] Geng, M., Y. Wang, Y. Tian, and T. Huang, "CNUSVM: Hybrid CNN-uneven SVM model for imbalanced visual learning", *Proceedings - 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016*, 2016, pp. 186-193.

[18] He, H. and E. A. Garcia, "Learning from imbalanced data", *IEEE Transactions on Knowledge and Data Engineering*, Vol.21, No.9, 2009, pp. 1263-1284.

[19] Horak, J., J. Vrbka, and P. Suler, "Support vector machine methods and artificial neural networks used for the development of bankruptcy prediction models and their comparison", *Journal of Risk and Financial Management*, Vol.13, No.3. 2020, p. 60.

[20] Kang, P. and S. Cho, "EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems", *ICONIP 2006: Neural Information Processing*, 2006, pp. 837-846.

[21] Kim, M. J., D. K. Kang, and H. B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction", *Expert Systems with Applications*, Vol.42, No.3, 2015, pp. 1074-1082.

[22] Kim, T. and H. Ahn, "A hybrid under-sampling approach for better bankruptcy prediction", *Journal of Intelligence and Information Systems*, Vol.21, No.2, 2015, pp. 173-190.

[23] Kubat, M., R. Holte, and S. Matwin, "Learning when negative examples abound", *ECML 1997: Machine Learning: ECML-97*, 1997, pp. 146-153.

[24] Kuspriyanto, S. O., D. H. Widyantoro, H. S. Sastramihardja, K. Muludi, and S. Maimunah, "Performance evaluation of SVM-based information extraction using τ margin values", *International Journal on Electrical Engineering and Informatics*, Vol.2, No.4, 2010, pp. 256-265.

[25] Lee, H. K. and S. B. Kim, "An overlap-sensitive

margin classifier for imbalanced and overlapping data", *Expert Systems with Applications*, Vol.98, 2018, pp. 72-83.

[26] Li, Y. and J. Shawe-Taylor, "The SVM With Uneven Margins and Chinese Document Categorisation", *Language, Information and Computation : Proceedings of the 17th Pacific Asia Conference*, Sentosa, Singapore, 2003, pp. 1-3.

[27] Li, Y., K. Bontcheva, and H. Cunningham, "Adapting SVM for data sparseness and imbalance: A case study in information extraction", *Natural Language Engineering*, Vol.15, No.2, 2009, pp. 241-271.

[28] Li, Y., K. Bontcheva, and H. Cunningham, "Using Uneven Margins SVM and Perceptron for Information Extraction", 2005, Available at http://svmlight.joachims.org.

[29] Mazurowski, M. A., P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", *Neural Networks*, Vol.21, No.2-3, 2008, pp. 427-436.

[30] Ni, W., T. Liu, J. Xu, Y. Huang, and J. Ge, "Acronym extraction using SVM with Uneven Margins", *Proceedings - 2010 IEEE 2nd Symposium on Web Society*, SWS 2010, 2010, pp. 132-138.

[31] Olson, D. L., D. Delen, and Y. Meng, "Comparative analysis of data mining methods for bankruptcy prediction", *Decision Support Systems*, Vol.52, No.2, 2012, pp. 464-473.

[32] Shawe-Taylor, J., "Classification accuracy based on observed margin", *Algorithmica*, Vol.22, 1998, pp. 157-172.

[33] Sundarkumar, G. G. and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance", *Engineering Applications of Artificial Intelligence*, Vol.37, 2015, pp. 368-377.

[34] Tao, X., Li, Q., Guo, W., Ren, C., Li, C., Liu, R., & Zou, J., "Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification", *Information Sciences*, Vol.487, 2019, pp. 31-56.

[35] Vapnik, V. N., "An overview of statistical learning theory", *IEEE Transactions on Neural Networks*, Vol.10, No.5, 1999, pp. 988-999.

[36] Veganzones, D. and E. Séverin, "An investigation of bankruptcy prediction in imbalanced datasets", *Decision Support Systems*, Vol.112, 2018, pp. 111-124.

[37] Veropoulos, K., C. Campbell, and N. Cristianini, "Controlling the Sensitivity of Support Vector Machines", *Proceedings of the International Joint Conference on Artificial Intelligence, Stockhol, Sweden (IJCAI99)*, 1999, pp. 55-60.

[38] Wu, G. and E. Y. Chang, "KBA: Kernel boundary alignment considering imbalanced data distribution", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17 No.6, 2005, pp. 786-795

[39] Yan, Q., S. Xia, and F. Meng, "Optimizing cost-sensitive SVM for imbalanced data: Connecting cluster to classification", arXiv:1702.01504, Cornell University, 2017, pp. 1-15.

[40] Zhang, T. and Z. H. Zhou, "Optimal margin distribution machine", *IEEE Transactions on Knowledge and Data Engineering*, Vol.32, No.6, 2020, pp. 1143-1156.

[41] Zhou, L., "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods", *Knowledge-Based Systems*, Vol.41, 2013, pp. 16-25.

[42] Zou, Q., S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification", *Big Data Research*, Vol.5, 2016, pp. 2-8.

Information Systems Review
Volume 24 Number 4
November 2022

Optimization of Uneven Margin SVM to Solve Class Imbalance in Bankruptcy Prediction

Sung Yim Jo* · Myoung Jong Kim**

Abstract

Although Support Vector Machine(SVM) has been used in various fields such as bankruptcy prediction model, the hyperplane learned by SVM in class imbalance problem can be severely skewed toward minority class and has a negative impact on performance because the area of majority class is expanded while the area of minority class is invaded. This study proposed optimized uneven margin SVM(OPT-UMSVM) combining threshold moving or post scaling method with UMSVM to cope with the limitation of the traditional even margin SVM(EMSVM) in class imbalance problem. OPT-UMSVM readjusted the skewed hyperplane to the majority class and had better generation ability than EMSVM improving the sensitivity of minority class and calculating the optimized performance. To validate OPT-UMSVM, 10-fold cross validations were performed on five sub-datasets with different imbalance ratio values. Empirical results showed two main findings. First, UMSVM had a weak effect on improving the performance of EMSVM in balanced datasets, but it greatly outperformed EMSVM in severely imbalanced datasets. Second, compared to EMSVM and conventional UMSVM, OPT-UMSVM had better performance in both balanced and imbalanced datasets and showed a significant difference performance especially in severely imbalanced datasets.

Keywords: Uneven Margin SVM, SVM, Class Imbalance, Bankruptcy Prediction, OPT-UMSVM

* Ph.D Student, School of Business, Pusan National University
** Corresponding Author, Professor, School of Business, Pusan National University

◐ 저 자 소 개 ◑

Sung Yim Jo (reina337@pusan.ac.kr)

She is a Doctoral course student of Division of Business in Pusan National University. She received an undergraduate degree from Pusan National University and a MS degree from Korea University. She published papers related to corporate bankruptcy. Her main research interests are Fin-tech and Data Mining in accounting and finance fields.

Myoung-Jong Kim (mjongkim@pusan.ac.kr)

He is a professor of Division of Business in Pusan National University. He received a BS and MS degree from Sungkyunkwan University, and a PhD from Korea Advanced Institute of Science and Technology in Korea. He has published many papers related to the business applications of Artificial Intelligence. His main research interests are Data Mining and intelligent systems in accounting and finance fields.

논문접수일 : 2022년 08월 04일 게재확정일 : 2022년 10월 28일
1차 수정일 : 2022년 09월 28일 2차 수정일 : 2022년 10월 13일