데이터 불균형 문제에서의 SVM 앙상블 기법의 적용

강필성 이형주 조성준[0]
서울대학교 산업공학과
{xfeel80, impatton, zoon[0]} @snu.ac.kr

SVM Ensemble Techniques for Class Imbalance Problem

Pilsung Kang, Hyoung-joo Lee, Sungzoon Cho[0]
Dept. of Industrial Engineering, Seoul National University

요 약

대부분의 기계학습 알고리즘은 학습 데이터에서 각각의 범주간의 비율이 동일하거나 비슷하다는 가정 하에 문제를 풀게 된다. 그러나 실제 문제에서는 그 비율이 동일하지 않으며 매우 큰 차이를 보이기도 하는데, 이는 분류 성능을 저하시키는 요인이기도 하다. 따라서 본 논문에서는 이러한 데이터의 불균형 문제를 해소하는 방안으로 SVM 앙상블 기법을 적용한 샘플링을 제안하고 이를 실제 불균형 데이터에 적용함으로써 제안된 방법이 기존의 방법들에 비해 향상된 성능을 나타내는 것을 보였다.

1. Introduction

데이터 불균형(class imbalance)문제는 데이터 셋에서 한 범주에 속하는 패턴의 수가 다른 범주에 속하는 패턴의 수보다 매우 적거나 많은 경우를 말한다. 대부분의 기계학습 알고리즘은 범주들의 비율이 동일하다는 가정 하에 적용된다. 그러나, fraud detection[1], oil spill detection[2], response modeling[3]에서 나타나는 것처럼 실제 문제에서는 이러한 가정이 맞지 않는 경우가 빈번히 존재한다.

이와 같은 문제는 분류 성능을 저하시키는 요인으로 작용하는데, 이를 해결하기 위하여 여러 가지 연구들이 진행되었다. Japkowicz[4]는 1차원의 분포로부터 가상의 불균형 데이터를 생성하고 불균형을 해소시키는 두 가지 샘플링 방법 (oversampling, undersampling)을 비교하였다. Kubat and Matwin[5]은 Tomek Link를 사용하여 학습에 필요한 패턴을 다수 범주로부터 샘플링하여 다수 범주에 속한 패턴의 수를 감소시키는 방법을 제안하였다. Chawla et al.[6]은 k-NN기법을 사용하여 소수 범주 주변에 인공적으로 데이터를 생성하는 방법으로 불균형을 해소하는 방법을 제안하였다. Shin and Cho[3]는 오분류된 패턴에 대하여 소수 범주와 다수 범주에 서로 다른 페널티를 부과함으로써 샘플링을 하지 않고 불균형을 해소하는 방법을 제안하였다.

본 논문에서는 undersampling된 다수 범주의 학습 데이터 셋에 앙상블 기법을 적용하여, 샘플링 과정에서 나타나는 분포의 왜곡을 줄이고 일반화 성능을 향상시키는 방법을 제안하고, 이에 대한 실험방법 및 결과를 기술하였다.

본 논문의 순서는 다음과 같다. 2장에서는 가상 데이터를 통하여 불균형 데이터가 분류 성능에 미치는 영향에 대해서 논의하고 3장에서는 본 논문에서 제안하는 방법을 소개한다. 4장에서는 제안된 방법으로 실험을 수행하고 그에 대한 결과를 해석하고, 5장에서는 연구에 대한 토의와 추후 연구 과제에 대한 논의한다.

2. The effect of class imbalance

2.1 성능 척도 (Performance Measure)

기계학습에서 분류 문제의 성능 척도로는 [표1]에서 (TP+TN)/(TP+FN+FP+TN)으로 계산되는 정확도(accuracy)가 사용된다. 그러나 불균형 데이터의 경우에는 전체 정확도가 다수 범주에 대한 정확도에 많은 영향을 받게 된다. 따라서 소수 범주와 다수 범주의 정확도를 모두 고려하는 geometric mean을 성능 척도로 사용한다. 소수 범주에 대한 정확도(A_+)는 TP/(TP+FN), 다수 범주에 대한 정확도(A_-)는 TN/(FP+TN)으로 계산한다. 그리고 geometric mean은 다음과 같이 계산할 수 있다.

$$GeometricMean = \sqrt{A_+ \cdot A_-}$$

표 1 성능 척도

		Predict	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

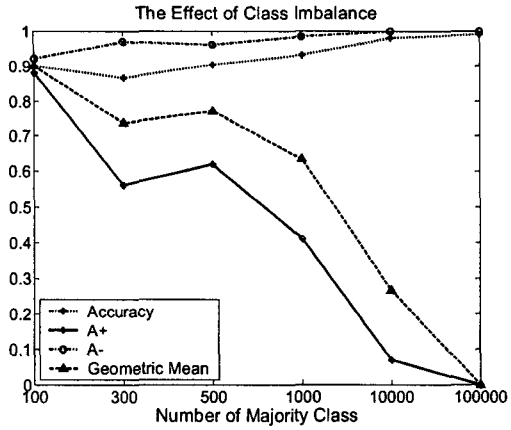

그림 1. The Effect of Class Imbalance

2.2 가상 데이터에서의 불균형 데이터의 영향

불균형 데이터가 실제로 분류 성능에 어떠한 영향을 끼치는지 알아보기 위하여 가상 데이터를 생성하여 실험하였다. 2차원 평면상에서 평균이 [0,0]이고 분산이 [1,1]인 패턴 100개를 소수 범주로 생성하고, 평균이 [3,0]이고 분산이 [2,2]인 데이터를 다수범주로 하여 100, 300, 500, 1000, 10000, 100000개의 패턴을 생성하고, support vector machine(SVM)을 기본 분류기로 하여 실험하였다. [그림 1]에서 알 수 있듯이, 비교적 단순한 2차원 데이터임에도 불구하고, 불균형이 심해질수록 전체 정확도와 다수 범주에 대한 정확도는 증가하지만 소수 범주에 대한 정확도와 geometric mean은 감소하는 것을 알 수 있다. 특히 다수 범주의 패턴이 100,000개인 경우에는 모든 패턴을 다수 범주로 분류하여, 소수범주에 대해서는 전혀 분류를 못하고 있음을 알 수 있다.

3. Ensemble with undersampling

앞서 알아본 바와 같이 데이터 불균형 문제는 분류 성능을 저하시키는 요인이기 때문에, 이 문제를 해결한다면 분류 성능을 향상시킬 수 있다. 본 논문에서는 undersampling과 앙상블을 결합함으로써 불균형 문제를 해결하고자 한다. Japkowicz[4]의 연구에서는 undersampling과 oversampling 모두 데이터 불균형을 해소하는 데 효과적인 방법임이 밝혀졌다. 또한 실증적으로도 두 가지 샘플링 방법 모두 효과가 있음을 알 수 있다[5,6]. 그러나, 소수 범주의 데이터를 다수 범주의 데이터 수만큼 생성하는 oversampling은 학습에 상당히 긴 시간이 소요되기 때문에, 시간 복잡도를 고려한다면 효율적인 방법이라고 할 수 없다. 반면, undersampling은 다수 범주의 데이터 전체를 사용하지 못함으로써 정보의 손실과 분포의 왜곡을 가지고 온다는 단점이 있다. 앙상블 기법은 많은 인구수(population)을 가지고 있기 때문에 이러한 단점을 보완해 줄 수 있다. 그리고, 이 방법은 앙상블에 결합되는 개별 분류기의 성격을 다양화할 수 있기 때문에, 앙상블에 적합한 방법이라고 할 수 있다.

본 연구에서는 다음과 같은 방법으로 undersampling과 앙상블을 결합하였다.

[Step 1] Partitioning training data: 학습데이터를 소수 범주 데이터와 다수 범주 데이터로 분리
[Step 2] Undersampling majority class data: 다수 범주 데이터에서 소수 범주 데이터의 수만큼 비복원 랜덤 샘플링한 데이터 셋을 앙상블 인구수만큼 생성
[Step 3] Construct ensemble training data: [Step 2]에서 생성한 각각의 데이터셋을 소수 범주 데이터와 결합하여 범주 비율이 1:1이 되는 학습 데이터를 앙상블 인구수만큼 생성

4. Experiment

4.1 Data description

실험에 사용된 데이터는 기계학습에서 주로 사용되는 데이터인 UCI Repository[7]에서 범주간 불균형이 나타나는 네 가지 분류 문제를 선택하였다. 원래 데이터의 범주 수가 세 개 이상인 데이터들은 하나의 범주를 소수 범주로 선택하고 나머지 범주들을 다수 범주로 배정하는 방법을 사용하였다. 소수 범주와 다수 범주의 비율은 최소 1:6 (Image)에서부터 최대 1:12 (Ann-Tyroid)까지이다. 실험에 사용된 데이터들은 [표 2]와 같다.

4.2 Experimental Setting

본 연구에서는 기본 분류기로는 SVM을 사용하였다. SVM의 커널은 가우시안 커널을 사용하였다. SVM의 두 모수인 Cost와 가우시안 커널의 넓이는 학습 데이터에서 10-fold cross validation을 하여 가장 좋은 geometric mean을 보이는 조합을 선택하였다. 그리고, 선택된 두 모수의 조합으로 전체 학습 데이터를 학습하고, 테스트 데이터에 적용하였다. 앙상블의 인구수는 50으로 설정하였고 결과값은 다수결 투표 방법으로 결정하였으며, 모든 실험은 30회 반복하였다.

제안된 앙상블 기법과 결과를 비교하기 위하여, original SVM, undersampling SVM, oversampling SVM, modifying-cost SVM 네 가지 방법을 실험하였다. Original SVM은 불균형한 학습데이터를 그대로 사용하는 모델이다. Undersampling SVM은 다수 범주에서 소수 범주의 패턴 수만큼 랜덤 샘플링하여 학습데이터를 구성한 모델이며, oversampling SVM은 소수 범주에서 다수 범주의

표 2. 실험에 사용된 데이터

Domain	Original Class	TrainData		Test Data	
		Minor	Major	Minor	Major
Image	7	30	180	300	1800
Vowel-context	11	48	480	42	420
Ann-Tyroid	3	284	3488	250	3178

패턴의 수만큼 랜덤하게 복제하여 학습데이터를 구성한 모델이다. Modifying-cost SVM은 불균형한 학습데이터를 그대로 사용하되, 소수 범주와 다수 범주의 사전확률의 역수를 오분류 페널티로 적용한 모델이다.

4.3 Experimental Results

네 가지 데이터에 대한 실험 결과는 [표 3~6]에 정리되어 있다. Ann-Thyroid의 modifying-cost SVM을 제외하고는, 모든 방법들이 original SVM에 비하여 향상된 성능을 나타내어 데이터의 불균형을 해소하는 것을 알 수 있다. 또한 본 연구에서 제안된 앙상블 기법이 모든 데이터에서 가장 좋은 결과를 보이고 있음을 알 수 있다. 소수 범주에 대한 정확도인 A_+의 측면에서 볼 때, 앙상블 기법이 다른 기법들에 비해 높은 정확도를 보임으로써, 소수 범주를 정확히 판별하는 것이 중요한 이슈가 되는 실제 문제들에 적용할 경우 우수한 성능을 보일 것으로 기대된다. [표 7]에서, 30회 반복 실험의 분산이 앙상블을 함으로써 작아진다는 것은 분포의 왜곡을 감소시켜 안정적인 성능 향상이 가능하다는 가정을 뒷받침해 준다.

표 3. Image Data 실험 결과

Image	G-Mean	Accuracy	A+	A-
Original	0.8578	0.9302	0.7633	0.9639
Under	0.8671	0.8470	0.8981	0.8385
Over	0.8953	0.9179	0.8653	0.9267
ModCost	0.8621	0.8519	0.8768	0.8478
Ensemble	**0.8972**	0.8740	0.9312	0.8645

표 4. Vowel-context Data 실험 결과

Vowel	G-Mean	Accuracy	A+	A-
Original	0.8976	0.9784	0.8095	0.9952
Under	0.9112	0.9548	0.8659	0.9637
Over	0.8290	0.9675	0.6905	0.9952
ModCost	0.9052	0.9697	0.8333	0.9833
Ensemble	**0.9453**	0.9646	0.9191	0.9724

표 5. Ann-Tyroid Data 실험 결과

Ann	G-Mean	Accuracy	A+	A-
Original	0.8521	0.9719	0.7324	0.9914
Under	0.9241	0.9377	0.9085	0.9401
Over	0.9329	0.9702	0.8912	0.9706
ModCost	0.6728	0.7444	0.5986	0.7563
Ensemble	**0.9378**	0.9514	0.9222	0.9537

표 6. Satimage Data 실험 결과

Satimage	G-Mean	Accuracy	A+	A-
Original	0.7992	0.9310	0.6635	0.9625
Under	0.8944	0.8629	0.9365	0.8542
Over	0.8981	0.9015	0.8938	0.9023
ModCost	0.8868	0.8665	0.9147	0.8597
Ensemble	**0.9074**	0.8738	0.9523	0.8643

표 7. Undersampling SVM과 Ensemble의 분산

	Image	Vowel	Ann	Satimage
Under	0.0216	0.0568	0.0077	0.0057
Ensemble	0.0009	0.0267	0.0000	0.0035

Satimage	6	415	4020	211	1789

5. Conclusion

본 연구에서는 데이터 불균형 문제를 해소하기 위하여 앙상블 기법을 사용한 샘플링 방법을 제안하였다. 데이터 불균형 문제가 분류 성능에 미치는 영향을 서술하고 제안된 방법을 실제 데이터에 적용하였다. 네 가지의 UCI 데이터에 대한 실험결과, 기존의 기법들보다 정확도와 안정성 측면에서 향상된 결과를 얻음으로써 제안된 방법이 데이터 불균형 문제에 대한 하나의 해법이 될 수 있음을 보였다. 추후 연구로써는 기본 분류기로써 SVM이 아닌 다른 기계학습 알고리즘을 사용했을 경우에 본 논문에서 제안된 방법의 효과와, 앙상블을 하는데 있어 최적의 인구수 결정이나 결과값 결정 방식의 변경을 통해 보다 향상된 성능을 나타내는 앙상블 기법을 찾는 연구가 진행될 수 있을 것이다.

6. 참고문헌

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

[2] Kubat, M., Holte, R. and Matwin, S., Machine Learning for the detection of oil spolls in satellite radar images, *Machine Learning 30*, 195-215, 1998

[3] Shin, H.J. and Cho, S.Z., Response Modeling with Support Vector Machines, *Data Mining and Knowledge Discovery*, (submitted), 2003.

[4] Japkowicz, N., The Class Imbalance Problem : Significance and strategies. *In Proceedings of the 2000 International Conference on Artificial Intelligence*, 2000

[5] Kubat, M. and Matwin, S., Addressing the Curse of Imbalanced Data Sets : One-Sided Sampling, *In Proceedings of the Fourteenth International Conference on Machine Learning*, 179-186, 1997

[6] Chawla, N., Hall, L. and Kegelmeyer, W., SMOTE : Synthetic Minority Oversampling Techniques, *Journal of Artificial Intelligence Research 16*, 321-357, 2002

[7]http://www.ics.uci.edu/~mlearn/MLRepository. html