GS-WSVM 논문 개요

1. Introduction

2. Related work

아래는 논문 내용 전개에 필수적인 관련 논문들입니다. 본문에서는 더 많은 논문을 언급할 계획입니다.

1. Veropulos et al.(1999) WSVM을 제안
2. Akbani et al.(2004) WSVM의 boundary 왜곡 문제를 규명하고 SMOTE를 결합하는 해결법을 제안
3. Lin et al.(2002) Oversampling + WSVM의 asymptotic optimality를 보임
4. Liu et al.(2010) Bayesian decision theory의 관점에서 cost-sensitive learning과 oversampling이 이론적으로 같은 결과를 냄을 보임
5. Bang and Kim(2020) GMC-SMOTE를 제안
6. Mathew et al.(2018) synthetic sample과 original sample에 오분류 비용 차등 적용하는 WSVM을 제안

본 연구의 의의

1. WSVM의 boundary 왜곡 문제를 해결하기 위한 oversampling 적용 시 SVM의 asymptotic optimality가 잘 유지되도록 GMC-SMOTE를 도입
2. Finite sample에서 발생할 수 있는 synthetic sample의 sampling bias의 영향을 완화하기 위해 original sample과 synthetic sample에 차등적으로 적용할 오분류 비용을 계산하는 구체적인 방법을 제시하고, 이러한 차등 적용시에도 SVM의 asymptotic optimality가 유지됨을 보임

3. Combining cost-sensitive learning and oversampling for imbalanced classification

이 챕터에서는 GS-WSVM을 정당화합니다. 주요 내용은 아래와 같습니다.

1. WSVM과 oversampling을 합치는 이유: WSVM을 쓰면 KKT condition 때문에 positive support vector의 수가 감소하여 decision boundary가 왜곡되기 때문
2. Oversampling에 GMC-SMOTE를 사용하는 이유: Oversampling을 해도 SVM의 optimality를 유지하려면 synthetic sample이 원본 분포와 비슷한 분포에서 뽑혀야 하는데, GMC-SMOTE가 SMOTE보다 원본 분포를 더 잘 근사하기 때문
3. Synthetic sample에 weight를 다르게 주는 이유: finite sample 상황에서는 GMC-SMOTE sample이 원본 분포와 차이가 있을 수 있기 때문에 synthetic sample의 영향력을 control할 필요가 있음

내용

이 논문에서 사용할 SVM의 form을 제시. Asymptotic result를 위해 form이 일반 SVM과 약간 다름. ($C = 1/2n\lambda$ 인 C-SVM과 동일)

$$\min_{h \in H_k} \frac{1}{n} \sum_{i=1}^{n} [(1 - y_i f(\mathbf{x}_i)]_+ + \lambda \|h\|_{H_K}$$

3.1.

1. **[WSVM 설명]** Imbalanced classification에서 SVM이 낮은 성능을 보이는 이유는 positive와 negative의 오분류 비용이 같으므로 objective function의 구조상 적은 수의 positive sample을 전부 오분류하고 margin을 키우는 것이 더 이득이기 때문이다[2]. 그래서 왼쪽 항의 hinge loss에 positive sample에는 $L(1)$, majority sample에는 $L(-1)$을 곱해($L(1) > L(-1)$ 오분류 비용을 차등 적용하여 cost-sensitive learning을 구현하는 WSVM이 나왔다[1].

2. **[WSVM의 단점]** WSVM은 결과적으로 positive sample의 Lagrange multiplier α_i 를 증가시키는데, KKT condition $\sum_{i=1}^{n} \alpha_i y_i = 0$ 때문에 positive support vector의 개수가 상대적으로 감소하게 된다. 이 때문에 SVM decision function $f(\mathbf{x}) = sign(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b)$ 에 소수의 positive support vector가 강한 영향력을 미치게 된다. 결과적으로 decision boundary가 소수의 positive support vector의 분포를 따라 구불거리게 되어 generalization ability가 감소한다[2].

3. **[WSVM과 SMOTE의 결합]** 이를 해결하기 위해 SMOTE로 오버샘플링 후 WSVM을 적용하는 방안이 제시되었다[2]. SMOTE로 생성한 synthetic sample들은 주변 original sample들의 정보를 반영하고 있으므로, synthetic support vector가 decision function에 영향을 미치는 것은 original support vector의 영향력의 범위를 확대하는 것과 같다. 이는 positive sample에만 kernel의 γ parameter를 더 작게($k(x, x') = \gamma|x - x'|^2$ parameterization 기준) 주는 것과 비슷한 효과를 낸다.

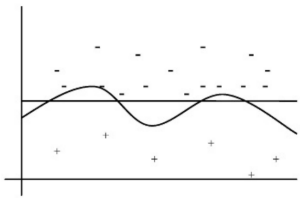

Fig. 4. The learned boundary (*curved line*) in the input space closely follows the distribution of the positive instances. The ideal boundary is denoted by the horizontal line

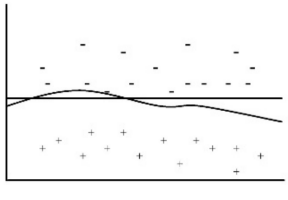

Fig. 5. After using SMOTE, the positive instances are now more densely distributed and the learned boundary (*curved line*) is more well defined

Akbani et al.의 figure

3.2.

statistical decision theory(Bayesian decision theory)를 사용하여 oversampling이 WSVM과 어떻게 상호작용하는지 파악하고 둘의 결합을 통계학적으로 정당화할 수 있다.

1. **[decision theory 설명]**... π^+, π^- : 모집단에서 positive, negative sample의 비율, false negative cost = c^+, false positive cost = c^- 인 상황을 가정

2. **[WSVM의 optimality]**oversampling 없는 WSVM은 decision boundary를 Bayes rule 방향으로 이동시킨다

   근거: Hinge loss에 $L(1) = c^+, L(-1) = c^-$를 곱한 WSVM의 decision function $f(\mathbf{x})$는 $n \to \infty$ 이고 적절한 λ 값일 때, Bayes rule $sign\left(Pr(Y = 1|\mathbf{X} = \mathbf{x}) - \frac{c^-}{c^- + c^+}\right)$로 수렴)[3].

3. **[oversampling이 Bayes rule에 미치는 영향]**positive sample을 $k \times 100\%$ oversampling하고 $c^+ = c^-$ 인 equal cost인 상황에서의 Bayes rule은, oversampling 하지 않고 cost ratio가 $\frac{c^-}{c^+} = \frac{1-(1+k)\pi^+}{(1+k)\pi^+(1-\pi^+)}$ 인 unequal cost 상황에서의 Bayes rule과 같다. 단, synthetic sample이 정확히 원본 분포에서 뽑혔다는 가정이 필요하다. 즉, Bayes rule에 미치는 영향의 측면에서 이상적인 oversampling은 unequal cost와 동일한 효과를 지닌다[4].

4. **[oversampling + WSVM의 optimality]** 따라서 oversampling 실시 후 Bayes rule이 받은 영향만큼 $L(1), L(-1)$을 조정한다면, 2번에서 단독으로 WSVM 한 것과 똑같은 Bayes rule로 수렴할 것이라고 예측할 수 있다.실제로 positive sample을 $k \times 100\%$ oversampling 한 다음

$$L(-1) = c^-\pi^-(1+k)\pi^+$$
$$L(1) = c^+\pi^+\left(\frac{1-(1+k)\pi^+}{1+\pi^+}\right)\pi^-$$

로 설정하면, WSVM의 decision function $f(\mathbf{x})$ 는 Bayes rule

$$sign\left(Pr(Y_s = 1|\mathbf{X_s} = \mathbf{x_s}) - \frac{L(-1)}{L(-1)+L(1)}\right) = sign\left(Pr(Y = 1|\mathbf{X} = \mathbf{x}) - \frac{c^-}{c^- + c^+}\right)$$

로 수렴한다($Y_s, \mathbf{X}_s$는 synthetic sample을 포함한 전체 sample이 뽑힌 분포)[3].

5. **[SMOTE 대신 GMC-SMOTE 쓰는 이유]** SMOTE는 positive sample이 구성하는 convex hull 안에서만 synthetic sample을 생성하므로 샘플이 뽑히는 분포가 원본 분포와 차이가 클 가능성이 높다. 또한 sample이 여러 subgroup으로 나뉘어 있는 경우 subgroup 사이에 새로운 sample을 잘못 생성하여 원본 분포에서는 나오지 않을 sample을 뽑을 가능성도 있다. Oversampling을 최대한 원본분포와 비슷한 분포에서 하기 위해 GMC-SMOTE[5]를 사용했다. Gaussian mixture는 충분히 큰 k에서 any smooth density를 approximate할 수 있기 때문에, $n \to \infty$ 인 상황에선 원본분포와 거의 비슷해질 것이라고 생각할 수 있다

6. **[synthetic sample에 다른 weight 주는 이유]** 그러나 실제 데이터 분석시 항상 very large sample을 가지고 있는 것은 아니므로 finite sample 상황에서 GMC-SMOTE sample은 원본 분포에서 뽑은 것과 차이가 있을 가능성이 있다. 따라서 synthetic sample이 알고리즘에 미치는 영향력을 조절할 필요가 있다. 이를 위해 synthetic minority sample과 original minority

sample에 서로 다른 오분류 비용을 부여하는 방법이 제안되었으나, 해당 연구에서는 synthetic sample과 original sample의 weight 차이를 어떻게 정할지에 대해서는 논의하지 않았다[6]. 본 연구에서는 synthetic과 original minority sample에 오분류 비용을 차등 적용하는 구체적 계산 방법을 제시하고, 이렇게 오분류 비용을 차등 적용해도 SVM의 asymptotic optimality가 유지됨을 보였다.

4. Proposed GS-WSVM algorithm

1. 알고리즘 소개
2. asymptotic property 증명

5. Simulation study

데이터 형태

- Gaussian mixture로 만든 데이터 사용
  - GMC SMOTE가 효과를 잘 발휘할 수 있음
  - Bayes rule을 analytic하게 계산할 수 있음
- Positive와 negative sample의 subgroup들이 섞여 있도록 하고 imbalance ratio를 높여 일반 WSVM을 사용했을 때 decision boundary가 구불구불한 상태가 되도록 함
  - GS-WSVM의 우수성이 plot에서 잘 드러나도록 함
  - Bang and Kim(2020)에 나온 것과 같은 checkerboard 데이터 사용 고려

비교 대상

- SVM, WSVM, SMOTE SVM 등

Figures

- Bang and Kim(2020) 4장에 나온 것과 같은 테이블 하나
- Lin et al.(2002) 4장에 나온 것과 같은 2차원 데이터 scatterplot, 주어진 상황에서의 Bayes rule과 각 분류기의 decision boundary가 그려진 plot 하나

비교 항목

- GS-WSVM이 다른 방법들보다 test set에서의 성능이 더 좋음을 g-mean 수치를 통해 보임
- GS-WSVM이 SVM보다 Bayes rule에 더 가까움을 decision boundary 그림을 통해 보임
- GS-WSVM이 WSVM보다 support vector imbalance ratio가 낮음을 보이고, 그 결과 boundary가 덜 구불구불함을 decision boundary 그림을 통해 보임

6. Real data study

7. Conclusion

References

[1] K. Veropoulos, C. Campbell, N. Cristianini, and Others, “Controlling the sensitivity of support vector machines,” Proc. Int. Jt. Conf. Artif. Intell., pp. 55–60, 1999

[2] R. Akbani, S. Kwek, and N. Japkowicz, “Applying Support Vector Machines to Imbalanced Datasets,” in Machine Learning: ECML 2004, 2004, pp. 39–50.

[3] Y. Lin, Y. Lee, and G. Wahba, “Support Vector Machines for Classification in Nonstandard Situations.,” Mach. Learn., vol. 46, no. 1–3, pp. 191–202, 2002

[4] A. Liu, C. Martin, B. La Cour, and J. Ghosh, “Effects of Oversampling Versus Cost-Sensitive Learning for Bayesian and SVM Classifiers,” in Data Mining: Special Issue in Annals of Information Systems, R. Stahlbock, S. F. Crone, and S. Lessmann, Eds. Boston, MA: Springer US, 2010, pp. 159–192.

[5] S. Bang and J. Kim, “Sampling Method Using Gaussian Mixture Clustering for Classification Analysis of Imbalanced Data,” Korean Data Anal. Soc., vol. 22, no. 2, pp. 565–574, Apr. 2020

[6] J. Mathew, C. K. Pang, M. Luo, and W. H. Leong, “Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines,” IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 9, pp. 4065–4076, Sep. 2018