扩展阅读

线性模型的多分类扩展与类别不平衡问题研究

引言

线性模型作为机器学习的基础工具，在处理多分类问题时面临两大挑战：一是如何将二分类模型有效扩展到多分类场景，二是如何解决现实数据中普遍存在的各个类别数量不平衡问题。本文系统性地探讨了多分类学习的三种基本策略（OvO、OvR、MvM）及其数学基础，深入分析了类别不平衡问题的产生机理与解决方案，并延伸讨论了代价敏感学习和稀疏表示等相关技术。

一、多分类学习的基本框架与实现策略

1.1 问题形式化

给定训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_m, y_m)\}$，其中 $y_i \in \{1, 2, ..., N\}$，多分类学习的目标是基于二分类线性模型构建决策函数 $f(\boldsymbol{x}) \to \{1, 2, ..., N\}$。

1.2 一对一策略（OvO）

- **算法原理**：构建 $C_N^2 = \dfrac{N(N-1)}{2}$ 个二分类器，每个分类器专门区分特定两类
- **决策机制**：采用投票法，$H(\boldsymbol{x}) = \arg\max\limits_{1 \le i \le N} \sum\limits_{j \ne i} sign(f_{ij}(\boldsymbol{x}))$
- **优势分析**：单个分类器训练效率高，适用于大规模特征数据集
- **局限性**：存储开销为 $O(N^2)$，预测阶段需要调用所有分类器

1.3 一对多策略（OvR）

- **算法原理**：构建 N 个二分类器，第 i 个分类器将第 i 类作为正例，其余作为反例
- **决策机制**：$H(\boldsymbol{x}) = \arg\max\limits_{1 \le i \le N} f_i(\boldsymbol{x})$
- **理论基础**：基于"最大间隔"原则，$f_i(\boldsymbol{x}) = \boldsymbol{w}_i^T \boldsymbol{x} + b_i$
- **缺陷分析**：训练样本分布不平衡（正负样本比1：N-1），易导致分类偏差

1.4 多对多策略（MvM）与ECOC框架[1]

1.4.1 编码矩阵设计

定义编码矩阵 $M \in \{-1, +1\}^{N \times L}$，其中每行对应一个类别的码字，每列定义一个二分类任务。最优编码矩阵应满足：

- 行分离性：$\min\limits_{i\neq j} d_H(\boldsymbol{m}_i, \boldsymbol{m}_j)$ 最大化
- 列分离性：$\min\limits_{j\neq k} d_H(\boldsymbol{m}^j, \boldsymbol{m}^k)$ 最大化

1.4.2 解码策略

对于测试样本 $\boldsymbol{x}$，预测码字 $\boldsymbol{s} = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_L(\boldsymbol{x}))$，解码函数：

$$H(\boldsymbol{x}) = \arg\min_{1\leq i\leq N} d(\boldsymbol{s}, \boldsymbol{m}_i)$$

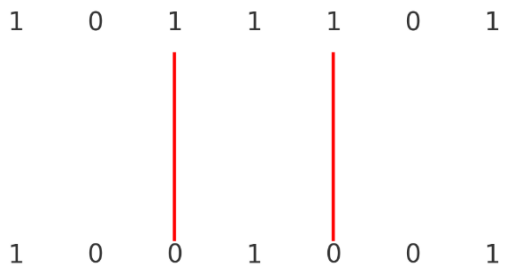
其中 d 为距离度量函数（常用海明距离或欧氏距离）

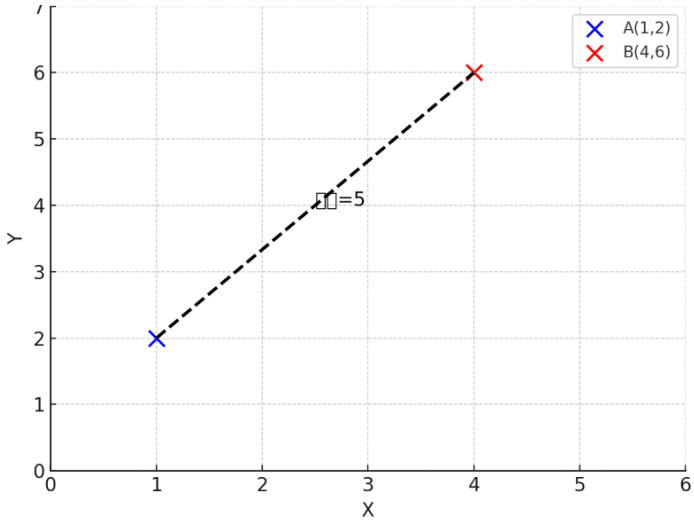
1.4.3 纠错能力分析

设最小海明距离 $d_{min} = \min\limits_{i\neq j} d_H(\boldsymbol{m}_i, \boldsymbol{m}_j)$，则ECOC可纠正 $\lfloor \frac{d_{min}-1}{2} \rfloor$ 个分类器的错误

注：

- **左边**是海明距离，红线标出了两个二进制串在不同位置上的差异。
- **右边**是欧式距离，在平面上展示了两点 (1,2) 和 (4,6)，它们之间的直线距离为 5。





二、类别不平衡问题的系统化解决方案[2]、[3]

2.1 问题形式化与影响分析

设训练集中第 i 类样本数为 m_i，若存在 $m_i \gg m_j$ （$i \neq j$），则称存在类别不平衡。这对OvR和MvM的影响尤为显著：

- OvR中反例样本数远多于正例
- MvM中正反例集合容量可能不对等

导致决策边界向少数类方向偏移

2.2 数据层面解决方案

2.2.1 重采样技术

- 欠采样：从多数类中随机删除样本
  - 改进方法：Tomek links [4]（双边清洗）、ENN [5]（Edited Nearest Neighbors，单边清洗）
- 过采样：向少数类中添加样本
  - 随机过采样：简单复制，易导致过拟合
  - SMOTE算法 [6]：在特征空间中进行插值

$$\boldsymbol{x}_{new} = \boldsymbol{x}_i + \lambda(\boldsymbol{x}_j - \boldsymbol{x}_i)$$

其中 $\boldsymbol{x}_j$为$\boldsymbol{x}_i$ 的 k 近邻样本， $\lambda \in [0,1]$ 为随机数

2.2.2 混合方法

- SMOTE+ENN：先过采样再清理重叠样本
- Borderline-SMOTE：仅在边界区域生成新样本

前者先用 SMOTE 把少数类“补”到大致平衡，再用 Tomek links 或 ENN 把两类边界附近“搅局”的样本剪掉，从而同时缓解“样本稀缺”和“类别重叠”两个问题。后者则是仅限“补”，并且是在边界进行样本补充

2.3 算法层面解决方案

2.3.1 代价敏感学习

定义代价矩阵 $C \in \mathbb{R}^{N \times N}$，其中 C_{ij} 表示将第 i 类误判为第 j 类的代价。优化目标变为：

$$\min \sum_{i=1}^{m} C_{y_i, f(\boldsymbol{x}_i)} + \lambda \Omega(\boldsymbol{w})$$

2.3.2 阈值移动

对于概率模型（如逻辑回归），调整决策阈值：

$$\frac{p(y=1|\boldsymbol{x})}{p(y=0|\boldsymbol{x})} > \frac{m^-}{m^+}$$

其中 m^+、m^- 分别为正负样本数

2.4 集成学习方法

- EasyEnsemble：将多数类样本划分成多个子集与少数类组合训练
- BalanceCascade：使用级联结构逐步淘汰被正确分类的多数类样本

三、扩展技术：稀疏表示与特征选择

3.1 稀疏表示的理论基础

信号 $\boldsymbol{x} \in \mathbb{R}^d$ 在字典 $D \in \mathbb{R}^{d \times k}$ 下的稀疏表示：

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{x} - D\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_0$$

由于 L0 范数非凸，实际常用 L1 范数替代

3.2 LASSO算法 [7]及其扩展

3.2.1 基本LASSO

$$\min_{\boldsymbol{w}} \frac{1}{2m}\|X\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1$$

- 产生稀疏解，实现特征选择
- 解路径的Piecewise Linear特性

3.2.2 扩展变体

- Group LASSO：对特征组进行整体选择

$$\|\boldsymbol{w}\|_G = \sum_{g=1}^{G} \sqrt{|g|}\|\boldsymbol{w}_g\|_2$$

- Adaptive LASSO：赋予不同系数自适应权重

$$\lambda \sum_{j=1}^{p} \frac{|w_j|}{|\hat{w}_j|^\gamma}$$

四、实验分析与应用建议

4.1 多分类策略选择指南

- 小规模数据集（N≤10）：优先选择OvO，准确性更高
- 大规模类别（N>100）：推荐使用OvR或MvM，计算效率更优
- 对分类器错误敏感场景：采用ECOC框架提升鲁棒性

4.2 不平衡处理方案选择

- 轻度不平衡（比例<10:1）：阈值移动或代价敏感学习
- 中度不平衡（10:1~100:1）：SMOTE 或集成方法
- 极度不平衡（比例>100:1）：组合策略（如 SMOTE+集成）

结论

本文系统建立了线性模型处理多分类与不平衡问题的理论框架。多分类学习中，OvO、OvR 和 MvM 三种策略各有适用场景，其中 ECOC 框架通过编码设计提供了误差校正机制。对于类别不平衡问题，需从数据层面（重采样）、算法层面（代价敏感）和模型层面（集成学习）多角度协同解决。稀疏表示技术如LASSO通过特征选择提升了模型可解释性。实际应用中应根据数据特性和任务需求选择适当的技术组合。

References Paper

[1] Dietterich, T. G., & Bakiri, G. (1995). *Solving multiclass learning problems via error-correcting output codes*. Journal of Artificial Intelligence Research, 2, 263-286.

[2] He, H., & Garcia, E. A. (2009). *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[3] Liu, X. Y., et al. (2009). *Exploratory undersampling for class-imbalance learning*. IEEE Transactions on Systems, Man, and Cybernetics, 39(2), 539-550.

[4] Tomek, I. (1976). *Two modifications of CNN*. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(11), 769–772. DOI.

[5] Wilson, D. L. (1972). *Asymptotic Properties of Nearest Neighbor Rules Using Edited Data*. IEEE Transactions on Systems, Man, and Cybernetics, SMC-2(3), 408–421. DOI.

[6] Chawla, N. V., et al. (2002). *SMOTE: synthetic minority over-sampling technique*. Journal of Artificial Intelligence Research, 16, 321-357.

[7] Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B, 58(1), 267-288.

References Book

[8] 周志华. (2016). *机器学习*. 清华大学出版社.

[9] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[10] Zou, H., & Hastie, T. (2005). *Regularization and variable selection via the elastic net*. Journal of the Royal Statistical Society: Series B, 67(2), 301-320.

[11] Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.