面向高频交易中的标签不平衡问题的深度学习算法研究

Zhichong Lyu

School of Economics and Management
Southwest Jiaotong University
zhichonglyu@gmail.com

2024 年 2 月 4 日

西南交通大學
Southwest Jiaotong University

高频交易

- 高频交易（High-Frequency Trading，HFT）是一种利用自动交易系统在极短的时间内捕捉并从市场微小波动中获利的交易策略。
- 高频交易通常采用电脑算法来执行大量高速证券交易，以赚取买卖价格之间的差价。这种交易策略在金融市场中具有很高的竞争力，因此需要不断优化算法和技术设备以保持领先地位。
- 高频交易的优点包括能够利用极小的价格波动来获利、增加市场流动性和使市场价格更加有效。然而，HFT 存在争议，因为它与市场波动性增加、潜在的市场操纵行为以及创造“ghost liquidity”相关联。批评者认为 HFT 为大公司提供了不公平的优势，可能会破坏市场平衡，对长期投资者造成损害(Kelejian & Mukerji, 2016; Yang et al., 2020; Korolev et al., 2015)。

标签不平衡

- 高频交易中的标签不平衡问题指的是数据集中类别（例如，买入或卖出订单）分布不均的问题。这种不平衡可能会显著影响基于此类数据训练的机器学习模型的性能，导致模型偏向多数类的预测。在 HFT 背景下，这可能意味着预测模型在准确预测较不频繁发生的市场事件方面效果不佳，而这对于盈利交易至关重要。
- 标签不平衡主要由市场行为的自然倾斜和数据收集的局限性造成。市场上的交易行为并非总是均衡发生的，而且在收集和处理数据时可能存在偏差，导致某些类型的交易或事件被过多或过少地记录。这使得模型在学习过程中可能过于关注多数类，而忽视了少数但可能更重要的类别，从而影响预测的准确性和模型的泛化能力。

产生原因

- 标签不平衡问题在高频交易中出现的原因很多，主要包括：
  - **市场行为的自然倾斜**：金融市场中的事件，如价格上涨或下跌，并不总是均衡发生的。在某些时期，市场可能会显示出对某一方向的偏好，导致交易信号在不同类别之间分布不均。
  - **数据收集的局限性**：交易数据的收集和处理可能受到技术和方法论的限制，导致某些类型的事件被过多记录，而其他类型则较少。
  - **金融市场的复杂性和动态变化**：金融市场受到各种宏观经济因素、政策变化、市场情绪等多种因素的影响，这些因素的不断变化导致了数据生成过程中的不均衡。

总述

- 为深入解决高频交易中的标签不平衡问题，深度学习模型可以采用多种策略，如利用复杂的网络结构来自适应地学习不平衡数据的特征，或者通过修改损失函数来强化对少数类的识别能力。此外，也可以结合使用数据层面的技术（如**重采样**）和算法层面的调整（如**代价敏感学习**和**集成学习方法**）来提高模型对不平衡数据的处理能力。具体实现时，可以考虑使用深度学习网络中的特定结构（如注意力机制）来增强模型对关键特征的捕捉能力，从而提高对少数类别的预测准确性。

重采样

- 在高频交易中解决标签不平衡问题的重采样技术具体使用方法包括过采样（增加少数类样本的数量）和欠采样（减少多数类样本的数量）。
- 以下方法均旨在创造一个更平衡的训练数据集，使得深度学习模型能够更有效地学习到所有类别的特征，提高对少数类的识别能力。
  - **过采样**：如**SMOTE(Synthetic Minority Over-sampling Technique)**，通过生成少数类样本的合成样本来增加其数量，使数据集更平衡，这种方法有助于深度学习模型更好地学习到少数类别的特征(Chawla et al., 2002; Fernández et al., 2018; Han et al., 2005)[备注：SMOTEBoost]。

重采样

- **欠采样**：随机删除多数类别中的样本来减少其数量，达到类别平衡的目的。
  - **NearMiss**方法通过选择多数类中与少数类最近的样本来减少多数类样本的数量。不仅减少了数据集的不平衡，而且尽可能保留了重要的信息。在高频交易数据分析中，使用 NearMiss 方法可以帮助模型更平衡地学习各类别的特征，从而提高对少数类别的预测准确性(Mani & Zhang, 2003a; Liu et al., 2008; Mani & Zhang, 2003b)。
  - **Tomek Links**方法识别不同类别之间的边界样本，然后删除这些样本中多数类的部分，以此来改善类别之间的分界清晰度。在高频交易数据中应用时，Tomek Links 方法通过减少多数类和少数类之间的重叠，帮助模型更准确地识别和预测少数类事件，从而提高整体预测性能。这种方法特别适用于数据集中存在轻微重叠的情况，能够在不显著减少多数类样本的情况下，提高模型对少数类的识别能力(Batista et al., 2004; Stefanowski & Wilk, 2008)。

代价敏感学习

- **代价敏感学习（Cost-Sensitive Learning）**通过为不同类别的样本分配不同的权重来解决标签不平衡问题的一种方法。在高频交易中，可以通过增加少数类样本的误差代价，降低多数类样本的误差代价，来迫使模型更关注少数类的正确分类。这种方法不需要改变数据集的分布，而是直接通过调整模型训练过程中的损失函数来实现。通过这种方式，模型在训练过程中会更加注重对少数类别的学习，从而提高在不平衡数据上的预测性能。

代价敏感学习

- (Elkan, 2001)讨论了代价敏感学习的基础理论，提出了代价敏感学习相比于传统学习方法的优势和应用场景。
- (Ling & Sheng, 2008)综述文章讨论了类别不平衡问题和代价敏感学习的关系，详细解释了如何在类别不平衡的数据集上应用代价敏感学习方法。
- (Zhou & Liu, 2005)提出了一种训练代价敏感神经网络的新方法，特别关注于解决类别不平衡问题，为高频交易中的应用提供了实际指导。

损失函数

- (Lin et al., 2017)详细介绍了**Focal Loss**设计的动机和数学形式，并通过实验验证了其在处理极端类别不平衡问题时的有效性。Focal Loss 是一种设计用于解决分类问题中极端类别不平衡的损失函数。它通过修改交叉熵损失，降低了对容易分类样本的关注度，增加了模型对难分类（通常是少数类）样本的关注度，公式如下：

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

  - 其中，p_t 是模型对实际标签的预测概率，$alpha_t$ 是针对类别的权重，*gamma* 是调节因子，用于减少容易分类样本的权重。通过这种方式，Focal Loss 能够帮助模型在训练过程中更加关注少数类样本，从而提高在不平衡数据上的性能。

集成模型

- 随机森林通过构建多个决策树并在这些树的结果上进行投票来做出最终决策，对于不平衡数据，它可以通过调整每个树的训练数据或改变错误的权重来提高少数类的识别率(Chen et al., 2004; Khoshgoftaar et al., 2010; Galar et al., 2011)。
- 通过逐步调整数据权重，强制后续模型更加关注之前模型错误分类的样本，对不平衡数据处理，可以通过调整错误权重的策略来增加少数类样本的重要性(Sun et al., 2009; Viola & Jones, 2004; Freund & Schapire, 1997)。
- 通过逐步减少模型残差的方式来构建新的模型，对于不平衡数据的处理，可以通过采样技术或者自定义损失函数来提升少数类样本的预测性能(Natekin & Knoll, 2013; He & Garcia, 2009; Friedman, 2001)。

References I

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*.

Chen, C., Liaw, A., Breiman, L., et al. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence.*

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*.

References II

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). 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)*.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*.

References III

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*.

Kelejian, H. H., & Mukerji, P. (2016). Does high frequency algorithmic trading matter for non-at investors? *Research in International Business and Finance*.

Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*.

Korolev, V. Y., Chertok, A., Korchagin, A. Y., & Zeifman, A. I. (2015). Modeling high-frequency order flow imbalance by functional limit theorems for two-sided risk processes. *Applied Mathematics and Computation*.

References IV

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the ieee international conference on computer vision.*

Ling, C. X., & Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*.

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*.

Mani, I., & Zhang, I. (2003a). knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets.*

Mani, I., & Zhang, I. (2003b). knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets.*

References V

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*.

Stefanowski, J., & Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. In *International conference on data warehousing and knowledge discovery*.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*.

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*.

Yang, H., Ge, H., & Luo, Y. (2020). The optimal bid-ask price strategies of high-frequency trading and the effect on market liquidity. *Research in International Business and Finance*.

References VI

Zhou, Z.-H., & Liu, X.-Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*.

Thanks!