Novel continual learning techniques on noisy label datasets

Monica Millunzi[1,2,*], Lorenzo Bonicelli[1], Alberto Zurli[2], Alessio Salman[2], Jacopo Credi[2] and Simone Calderara[1]

[1]*AImageLab - Università di Modena e Reggio Emilia, Modena, Italy*
[2]*Axyon AI, Modena, Italy*

Abstract

Many Machine Learning and Deep Learning algorithms are widely used with remarkable success in scenarios whose benchmark datasets consist of reliable data. However, they often struggle to handle realistic scenarios, particularly those in the financial sector, where available data constantly vary, increase daily, and may contain noise. As a result, we present an overview of the ongoing research at the AImageLab research laboratory of the University of Modena and Reggio Emilia, in collaboration with AxyonAI, focused on exploring Continual Learning methods in the presence of noisy data, with a special focus on noisy labels. To the best of our knowledge, this is a problem that has received limited attention from the scientific community thus far.

Keywords

deep learning, finance, noisy label, noise, classification, continual learning

1. Introduction

The use of artificial neural networks, enabled by the availability of large datasets, has yielded remarkable results across various domains, including time series and, consequently, finance. Recently, an increasing trend has emerged to enhance investment performance via the use of Deep Learning methods to forecast market price fluctuations. This provides asset managers with valuable assistance in assessing investment risk and return levels, as well as analyzing portfolio performance under uncertain circumstances. However, these applications are often affected by the high complexity and volatility of the financial market, which makes them susceptible to noisy data. One major reason is the unpredictable and often irrational behaviour of market participants that can cause sudden and unexpected changes(see the latest event at SVB [1]). Furthermore, financial data is often subject to measurement errors, data gaps, class overlap, and contexts where labels are ambiguous or subjective. Additionally, other external factors such as geopolitical events and economic policies can significantly contribute to adding noise to the data. Therefore, the prediction problems in this domain involve datasets with several challenges: (i) uncertain labels, (ii) imbalanced classes, and (iii) data that change significantly over time, leading to the well-known problem of concept drift [2]. While DNNs represent a strong tool for simple classification or prediction problems, financial data usually do not fit appropriately in standard deep learning scenarios for two main reasons: data comes continually from non-stationary environments and may change with high frequency, causing the data drifting phenomenon. This leads to the urgency of periodically retraining networks to learn from the most recent data. Indeed, as new data arrive as a stream, the network naturally adapts to the last-seen ones, making it prone to forget the previously acquired knowledge. This aspect is usually not considered, assuming that old financial trends do not affect future ones. Conversely, market trends are often cyclical; hence, making financial models retain the knowledge of past data while learning from the most recent ones may make them more reliable in already-experienced market conditions. By using Continual Learning techniques, models can adapt to changing financial data and improve their predictive power over time. Therefore, the application of CL to finance has the potential to enhance investment performance by providing more accurate and reliable predictions in the face of noisy data.

The research under discussion endeavours to explore complex CL scenarios that involve training data with incorrect and noisy labels.

2. Related Works

There is a growing body of work on the application of machine learning techniques to financial problems. The interested reader can find a comprehensive overview of financial machine learning techniques in [3, 4]. In this

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

[*]Corresponding author.
monica.millunzi@unimore.it (M. Millunzi); lorenzo.bonicelli@unimore.it (L. Bonicelli); alberto.zurli@axyon.ai (A. Zurli); alessio.salman@axyon.ai (A. Salman); jacopo.credi@axyon.ai (J. Credi); simone.calderara@unimore.it (S. Calderara)
0000-0001-7090-7159 (M. Millunzi); 0000-0002-9717-5602 (L. Bonicelli); 0000-0003-2656-2696 (A. Salman); 0000-0003-2656-2696 (J. Credi); 0000-0001-9056-1538 (S. Calderara)

work, however, we are specifically interested in the problem of training supervised models with noisy financial labels in a continual learning scenario. Both problems have been studied independently, see e.g. [5] for the former and [6, 7] for the latter.

Let us now delve more deeply into the two problems separately to show an overview of the state-of-the-art contribution in both fields, which will then allow us to deal with the current existing works that address both continual learning ad noisy labels simultaneously.

2.1. Noisy Labels

Learning with noisy labels (**LNL**) is a common and longstanding challenge in many machine learning applications, where the labels assigned to the data points may not be entirely accurate. Consequently, the model performances tend to degrade rapidly with respect to classical scenarios since DNNs can easily overfit noisy labels and show poor generalization performances. In this study, we focus mostly on the supervised classification learning task. In [8] the authors present a comprehensive overview of the existing approaches proposed to enhance the robustness of models against noisy labels. We can distinguish two main types of methods: those that attempt to learn the noise transition matrix and those that prioritize cleaning the data stream before using it to train the model. The former approach focuses on building a robust architecture that is capable to estimate a noise model that captures the label distribution pattern [9, 10]. Instead, the latter approach addresses the issue of keeping training data as clean as possible via various "sample-selection" techniques that try to identify clean examples based on the **small-loss** hypothesis [11, 12, 13] and then apply semi-supervised learning to re-label the leftover and hence train on the whole training set. Notably, some methods [14, 15, 16] feature a two-component Gaussian Mixture Model to model and split the loss distribution of correctly-labeled and incorrectly-labeled examples.

2.2. Continual Learning

The field of Continual Learning (**CL**) seeks to learn from a non-stationary non-i.i.d. stream of data without incurring the forgetting problem. Methods that tackle CL are commonly categorized into three main groups [17]. *Regularization-based* methods introduce regularization terms to prevent significant changes in the performance of the model on past tasks. Elastic Weight Consolidation (EWC) [18] and Synaptic Intelligence (SI) [19] seek to prevent parameters deemed as important for the current task from changing in the future. In EWC, the importance of each parameter is estimated as the Fisher information matrix while SI preserves connections that strongly affected the past task loss. Differently, Learning without Forgetting [20] (LwF) employs Knowledge Distillation (KD) to distill the model learned during past tasks into the current one. *Parameter-isolation* methods explicitly define a separate sub-network per task to avoid interference between the parameters during learning. Notably, in Progressive Neural Networks [21] (PNN) the learner is expanded at each subsequent task, while in Context-dependent Gating [22] (XdG) only sparse, mostly non-overlapping patterns of units are active for any task. While these methods usually achieve high performance, they usually rely on the knowledge of the task identity during inference, thus limiting their applicability to real scenarios. Finally, *Reharsal* methods [23, 24] are based on the idea of interleaving data for the current task with a buffer of data from the previous ones, a strategy commonly referred to as Experience Replay (ER). iCARL [25] combines replay with KD and uses the *herding* strategy to select the most representative samples for each class. DER [26] and X-DER [27] use the past responses of the model as a means for knowledge distillation.

2.3. Continual with Noisy Labels

Bridging the gap between CL and LNL, a preliminary work [28] proposes a self-supervised loss term to learn a representation that is independent of label noise, while distilling a small, purified buffer of samples for later use in fine-tuning and classification. Although this method achieves good initial performance when compared with common CL baselines, its learning objective features a low sample efficiency which limits its real-world applicability. To address this limitation, [29] exploits the loss difference between correct and noisy samples to split the incoming data, using the latter as unlabeled data in a FixMatch [30] objective. Finally, in [31] the authors seek a balance between the purity and the effectiveness of samples stored in the buffer. To achieve this, they promote the removal of samples from the buffer based on a combination of the sample loss with a measure of diversity in the buffer.

3. Datasets with noisy labels

Researchers have developed several datasets specifically designed to address the challenge of LNL. Although these consist of images and therefore differ from the typical financial datasets, we are interested in understanding model response to noise, regardless of the data itself. Here, we briefly describe the dataset commonly used in LNL literature:

- **Clothing1M**: this large-scale dataset contains 1 million images of clothing items from different online stores. The labels for the images were

Table 1
Final Average Accuracy (FAA) of CL baselines with different rates of label noise.

	Method	Split-N-CIFAR-10			
	Noise rate (symmetric)	0%	20%	40%	60%
	Multitask	91.69	82.02	72.04	54.83
	Finetuning	19.66	18.83	18.02	15.99
Offline	ER-ACE [34]	71.15	53.82	37.43	22.87
	ER-ACE w/ oracle	-	51.10	39.06	23.57
	ER-ACE w/ GMM	-	52.90	37.95	24.93
Online	ER-ACE [34]	49.14	36.31	29.61	19.90
	ER-ACE w/ oracle	-	36.60	29.06	20.85
	ER-ACE w/ GMM	-	36.82	30.12	22.04
	SPR [28]	-	43.9	43.0	40.0
	CNLL [29]	-	68.7	65.1	52.8

obtained through a web search and thus may contain incorrect labels.

- **WebVision**: this dataset contains over 2.4 million images from 1000 different categories collected from the web. The dataset has been used for various tasks such as image classification, object detection, and fine-grained recognition under noisy label scenarios.
- **CIFAR-10-N and CIFAR-100-N** [32]: these are variations of the popular CIFAR-10 and CIFAR-100 datasets [33] respectively, with human-annotated real-world noisy labels collected from Amazon Mechanical Turk[1].

Overall, these datasets provide valuable resources for developing and evaluating machine learning models under noisy label scenarios, which are common in many real-world applications.

In addition to these, another popular choice consists in artificially applying noise on the annotations of a clean dataset by flipping the label with a fixed probability. Such a synthetic setting is particularly helpful since the availability of ground-truth labels, coupled with the noisy ones, facilitates monitoring models' behaviour with respect to different noise patterns.

4. Problem formulation

For this proof preliminary analysis, we focus on a C-class image *classification* problem, which we split in T B-fold classification tasks. Let $X \in \mathscr{R}^{N\times d}$ and $Y \in \{1, \dots, C\}$ be the input and ground-truth label space respectively. In a standard CL setting, each task receives data $D_t = \{(x_i, y_i)\}_{i=1}^{N}$, where each pair (x_i, y_i) is independently and identically distributed according to a certain data generating distribution. Instead, in the LNL scenario, data comes from a noisy distribution $\tilde{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, with $\tilde{Y}$ being the noisy label space. Here, we assume that the corruption process that produces $\tilde{Y}$ from Y is independent of the input data; hence, one true label may be flipped into another label with equal probability (*noise rate*). We will refer to this process as "symmetric noise".

5. Experiments

For the sake of simplicity and a fair comparison with other existing works [28, 29], we will conduct experiments on CIFAR-10. This consists of 60000 32×32 images, usually split into 50000 images for the training set and 10000 for the test set. The classification task involves 10 non-overlapping classes and the number of examples per class is uniform. We modify the dataset by adding symmetric noise in each of the classes. To be compliant with Class-IL scenario [17], we split the 10-fold classification into 5 binary tasks and let the model learn from each one sequentially. We denote the version of the dataset obtained by such two modifications as **Split-N-CIFAR-10**.

As emerges from [28, 29, 31], keeping the memory buffer as clean as possible during training is a key aspect of dealing with the label noise issue in a continual learning scenario. Therefore, we decided to test the GMM technique used in [15, 14] to separate exemplars into noisy and clean against a strategy we call *"oracle"*. For the latter, since in this controlled scenario, we know which samples are associated with a noisy label, we can use this information to prevent all the noisy ones from being stored inside the replay buffer. Both techniques are tested using ER-ACE as a base strategy for CL. This

[1] https://www.mturk.com

method extends the common Experience Replay baseline by adding an asymmetrical cross-entropy loss between stream and buffer examples.

We used the ResNet18 [35] model initialized from scratch and trained with Stochastic Gradient Descent (SGD) for 50 epochs per task, in a standard **offline** fashion. Additionally, to allow comparison against [28, 29] we evaluate the **online** (single epoch) scenario.

5.1. Results

In Table 1 we report performances in terms of Final Average Accuracy (FAA) at the end of all tasks. Results are averaged across five runs and the buffer size is set to 500. We provide a lower and an upper bound, respectively finetuning without any countermeasure to forgetting and "Multitask", given by training all tasks jointly. Expectedly, the model is strongly affected by the presence of noise, and indeed its performances decrease as the noise rate increases. When in presence of low noise levels (20%), the standard ER-ACE model and its variants with buffer cleaning techniques perform on par, whereas for noise levels above 40%, the model seems to benefit from replaying a clean buffer. Surprisingly, in some scenarios the adoption of the GMM makes the model outperform the oracle for the same noise conditions, suggesting that having some noisy examples stored in the buffer may reduce generalization error by providing some extra regularization. Finally, while the GMM-based buffer cleaning provides an initial benefit, by looking at results from the online scenario it is clear that using semi-supervised learning on mislabeled examples (as CNLL does) provides a significant boost in terms of performance.

6. Conclusion

In this manuscript, we have investigated the challenging problem of learning from non-stationary and noisy data, which is highly relevant to the domain of financial data analysis. Through a joint effort between the AImageLab research laboratory at the University of Modena and Reggio Emilia and AxyonAI, we have conducted a preliminary analysis to assess the effectiveness of rehearsal-based continual learning baselines equipped with common label noise learning strategies. Our findings shed light on the complex interaction between these challenging fields and provide a proof of concept for future research in this area.

References

[1] M. T. Fennell, Svb financial group: Santa clara, california, Fed. Res. Bull. 107 (2021) 49.

[2] A. Tsymbal, The problem of concept drift: definitions and related work, Computer Science Department, Trinity College Dublin 106 (2004) 58.

[3] M. L. De Prado, Advances in financial machine learning, John Wiley & Sons, 2018.

[4] M. F. Dixon, I. Halperin, P. Bilokon, Machine learning in Finance, volume 1170, Springer, 2020.

[5] Y.-H. Lien, Y.-S. Lin, Y.-S. Wang, Uncertainty awareness for predicting noisy stock price movements, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part VI, Springer, 2023, pp. 154–169.

[6] D. Philps, T. Weyde, A. d. Garcez, R. Batchelor, Continual learning augmented investment decisions, arXiv preprint arXiv:1812.02340 (2018).

[7] A. Zurli, A. Bertugli, J. Credi, Does catastrophic forgetting negatively affect financial predictions?, in: Machine Learning, Optimization, and Data Science: 8th International Workshop, LOD 2022, Certosa di Pontignano, Italy, September 19–22, 2022, Revised Selected Papers, Part I, Springer, 2023, pp. 501–515.

[8] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE Transactions on Neural Networks and Learning Systems (2022).

[9] X. Chen, A. Gupta, Webly supervised learning of convolutional networks, 2015. arXiv:1505.01554.

[10] T. Xiao, T. Xia, Y. Yang, C. Huang, X. Wang, Learning from massive noisy labeled data for image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2691–2699.

[11] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, L. Fei-Fei, Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels, in: International Conference on Machine Learning, 2018.

[12] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, M. Sugiyama, Co-teaching: Robust training of deep neural networks with extremely noisy labels, Advances in neural information processing systems 31 (2018).

[13] H. Song, M. Kim, J.-G. Lee, Selfie: Refurbishing unclean samples for robust deep learning, in: International Conference on Machine Learning, PMLR, 2019, pp. 5907–5915.

[14] E. Arazo, D. Ortego, P. Albert, N. O'Connor, K. McGuinness, Unsupervised label noise modeling and loss correction, in: International conference on machine learning, PMLR, 2019, pp. 312–321.

[15] J. Li, R. Socher, S. C. Hoi, Dividemix Learning with noisy labels as semi-supervised learning, in: International Conference on Learning Representations, 2020.

[16] G. Pleiss, T. Zhang, E. Elenberg, K. Q. Weinberger,

Identifying mislabeled data using the area under the margin ranking, Advances in Neural Information Processing Systems 33 (2020) 17044–17056.

[17] G. M. van de Ven, A. S. Tolias, Three continual learning scenarios, in: Neural Information Processing Systems Workshops, 2018.

[18] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al., Overcoming catastrophic forgetting in neural networks, Proceedings of the national academy of sciences 114 (2017) 3521–3526.

[19] F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, in: International Conference on Machine Learning, 2017.

[20] Z. Li, D. Hoiem, Learning without forgetting, IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

[21] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, R. Hadsell, Progressive neural networks, arXiv preprint arXiv:1606.04671 (2016).

[22] N. Y. Masse, G. D. Grant, D. J. Freedman, Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization, Proceedings of the National Academy of Sciences 115 (2018) E10467–E10475.

[23] R. Ratcliff, Connectionist models of recognition memory: constraints imposed by learning and forgetting functions., Psychological Review (1990).

[24] A. Robins, Catastrophic forgetting, rehearsal and pseudorehearsal, Connection Science (1995).

[25] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, C. H. Lampert, iCaRL: Incremental classifier and representation learning, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017.

[26] P. Buzzega, M. Boschini, A. Porrello, D. Abati, S. Calderara, Dark Experience for General Continual Learning: a Strong, Simple Baseline, in: Advances in Neural Information Processing Systems, 2020.

[27] M. Boschini, L. Bonicelli, P. Buzzega, A. Porrello, S. Calderara, Class-incremental continual learning into the extended der-verse, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[28] C. D. Kim, J. Jeong, S. Moon, G. Kim, Continual learning on noisy data streams via self-purified replay, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 537–547.

[29] N. Karim, U. Khalid, A. Esmaeili, N. Rahnavard, Cnll: A semi-supervised approach for continual noisy label learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3878–3888.

[30] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, C. Raffel, Fixmatch: Simplifying semi-supervised learning with consistency and confidence, arXiv preprint arXiv:2001.07685 (2020).

[31] J. Bang, H. Koh, S. Park, H. Song, J.-W. Ha, J. Choi, Online continual learning on a contaminated data stream with blurry task boundaries, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9275–9284.

[32] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, Y. Liu, Learning with noisy labels revisited: A study using real-world human annotations, arXiv preprint arXiv:2110.12088 (2021).

[33] A. Krizhevsky, et al., Learning multiple layers of features from tiny images, Technical Report, Citeseer, 2009.

[34] L. Caccia, R. Aljundi, N. Asadi, T. Tuytelaars, J. Pineau, E. Belilovsky, New insights on reducing abrupt representation change in online continual learning, arXiv preprint arXiv:2203.03798 (2022).

[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. arxiv 2015, arXiv preprint arXiv:1512.03385 14 (2015).