Uma análise empírica do efeito do Ruído de Classe no aprendizado de Redes Neurais Artificiais

Lívia de Azevedo[1], **Filipe Braida**[1]

[1] Departamento de Ciência da Computação
Universidade Federal Rural do Rio de Janeiro (UFRRJ) – Nova Iguaçu, RJ – Brasil

livia_de_azevedo@yahoo.com.br, filipebraida@ufrrj.br

Abstract. *Label noise consists of the class labeling error. It can negatively affect the performance of a model, and it can vary with respect to the model chosen. For this reason, works have emerged that evaluate the natural resistance of Machine Learning models to label noise. Therefore, it would be relevant to evaluate the natural resistance of the artificial neural network to label noise, given its relevance to Deep Learning. The purpose of this work is to perform an experiment to evaluate the influence of label noise on artificial neural networks, training them on noisy databases. The results showed that the network complexity can influence their resistance to label noise.*

Resumo. *O ruído de classe consiste no erro de rotulação da classe. Ele pode afetar negativamente o desempenho de um modelo, podendo variar com relação ao modelo escolhido. Por essa razão, surgiram trabalhos que avaliam a resistência natural dos modelos de Aprendizado de Máquina ao ruído de classe. Sendo assim, seria relevante avaliar a resistência natural da rede neural artificial ao ruído de classe, dado a sua relevância ao Aprendizado Profundo. O objetivo deste trabalho é realizar um experimento para avaliar a influência do ruído de classe nas redes neurais artificiais, treinando-as sob base de dados ruidosas. Os resultados mostraram que a complexidade da rede pode influenciar a resistência delas ao ruído de classe.*

1. Introdução

O Aprendizado de Máquina (AM) se tornou popular por causa da sua capacidade de extrair informação e padrões complexos dos dados para poder tomar uma decisão inteligente, principalmente quando se adentra na subárea denominada Aprendizado Profundo (AP) que contém aplicações em visão computacional e processamento de linguagem natural, por exemplo. A maioria das abordagens de AP usam de Redes Neurais Artificiais (RNAs) como ideia base para as suas soluções, dando a estas redes uma relevância significativa de estudo [Russell and Norvig 2021, Goodfellow et al. 2016].

O processo de aprendizado começa com a observação dos dados com o objetivo em encontrar padrões. Desta maneira, a qualidade dos padrões observados está relacionado diretamente com a qualidade dos dados, tornando-se vital durante esse processo. O processo de aquisição de dados está sujeito a erros e interferências que podem resultar na coleta de dados inconsistentes, o que diminui a qualidade destes dados e consequentemente impacta negativamente o desempenho dos modelos de AM. Na literatura, a inconsistência ou um erro em um dado é chamado de ruído, e o ruído de classe é aquele

em que há um erro de rotulação na classe de um dado, sendo o ruído mais explorado na literatura [Frénay and Verleysen 2013, Han et al. 2020, Song et al. 2020]. As causas do ruído de classe envolvem principalmente fatores humanos durante a rotulação das classes [Frénay and Verleysen 2013]. Com uma grande quantidade de dados a disposição em um curto espaço de tempo, a presença do ruído de classe nos dados torna-se mais relevante. Song et al. [2020] apresentaram alguns trabalhos que mostraram uma estimativa variando entre 8% e 38,5% de ruído de classe nas bases de dados reais.

Frénay and Verleysen [2013] apresentam que uma das formas de lidar com o ruído de classe é considerar a resistência natural do modelo preditivo, ou seja, o modelo ser robusto ao ruído. Na literatura, diversos trabalhos experimentais foram realizados para avaliar a robustez natural de modelos clássicos de AM e também de AP. Estes trabalhos têm a sua importância para reforçar, conhecer, validar e trazer intuições sobre os efeitos do ruído nestes modelos, demonstrando o impacto do ruído na acurácia dos modelos.

Dado a relevância das RNAs para o AP e a importância do estudo da robustez dos modelos ao ruído de classe, o objetivo deste trabalho é oferecer uma análise experimental que forneça uma observação sobre os efeitos do ruído de classe no aprendizado das RNAs, considerando diferentes tipos de ruído de classe em modelos distintos de RNAs.

O artigo está organizado da seguinte forma: a Seção 2 descreve os trabalhos relacionados neste contexto experimental; a Seção 3 descreve a metodologia utilizada para o experimento de avaliação da robustez nas redes; a Seção 4 os resultados do experimento realizado e por fim a Seção 5 trará as conclusões dos resultados obtidos.

2. Trabalhos Relacionados

Nettleton et al. [2010] efetuaram uma análise experimental do efeito do ruído de classe em diferentes contextos em quatro modelos de AM: Naïve Bayes, k-vizinhos mais próximos (kVP), Árvore de Decisão e Máquinas de Vetor de Suporte (MVS). O Naïve Bayes e o kVP foram os modelos que apresentaram maior resistência ao ruído e com resistência semelhantes, seguidos pela Árvore de Decisão e o MVS como o pior de todos.

Rolnick et al. [2017] utilizaram algumas arquiteturas de AP e de Perceptron de Múltiplas Camadas para realizar os experimentos com dados sob o efeito de ruído de classe, selecionando duas formas diferentes de como o ruído é gerado nas bases de dados. O trabalho observou que os modelos de arquiteturas profundas apresentaram uma maior resistência ao ruído e que quanto mais há exemplos ruidosos na base é preciso uma quantidade maior de exemplos corretos para suprir os exemplos ruidosos.

Rusiecki [2020] buscou verificar empiricamente se o *Dropout* aplicado em RNAs consegue tornar o modelo mais resistente ao ruído de classe. Foram utilizadas duas arquiteturas de AP sob o efeito de um único tipo de ruído. O trabalho concluiu que o uso do *Dropout* poderia ajudar na resistência da rede ao ruído de classe.

Algan and Ulusoy [2020] trabalharam com três tipos de ruído de classe nos experimentos feitos: aleatório, um que depende da classe do exemplo correspondente e outro que depende dos atributos do exemplo. Duas arquiteturas de AP foram usadas para os experimentos. Os resultados mostraram que o ruído que depende do atributo foi mais prejudicial aos modelos de maneira geral do que os demais.

Mesmo com os resultados experimentais relevantes obtidos pelos trabalhos ante-

riores a respeito da robustez dos modelos, seria interessante realizar mais experimentos em outros cenários, a exemplo da variação de parâmetros específicos do modelo, para aumentar a compreensão da robustez dos modelos que, neste trabalho, será voltado para a RNA.

3. Metodologia

Nesta seção será explicada com detalhes a metodologia experimental proposta que envolve as etapas de separação em treino e teste das bases de dados, a forma como o ruído foi gerado artificialmente, a forma de avaliação dos modelos e como foi definido o treinamento dos modelos de RNAs. É comum na literatura gerar artificialmente o ruído, exemplificado pelos trabalhos relacionados anteriores.

Foram usadas duas bases de dados populares de imagens: o MNIST [LeCun 1998] e o Fashion-MNIST [Xiao et al. 2017]. Os conjuntos de treino e teste foram os mesmos disponibilizados pelas referências originais das bases, os quais consistem em 60000 imagens para o treino e 10000 imagens para o teste.

Frénay and Verleysen [2013] definiram uma taxonomia para o ruído de classe, caracterizada em três tipos: *Noise Completely at Random* (NCAR), o qual o erro de rotulação é algo aleatório; *Noise at Random* (NAR), o qual o erro de rotulação depende da classe real do exemplo; e o *Noise not at Random* (NNAR), o qual o erro de rotulação depende da classe real do exemplo e também dos seus atributos. Além disso, Frénay and Verleysen [2013] explicaram que para o caso do NAR é possível representar a distribuição do ruído nos dados como uma matriz de transição, em que a linha representa a classe verdadeira, a coluna a classe observada e o elemento da matriz a probabilidade condicional da classe observada dado a classe verdadeira.

Neste trabalho foram usadas duas formas de gerar ruído: ruído uniforme e o *flip noise*, explicitadas pelas matrizes abaixo. p_e é a probabilidade de existir um ruído, n_y é o número de classes do problema e o *flip noise* depende da escolha de um ou mais pares de classes distintas para a construção da matriz. Mais de um par pode ser representado na mesma matriz de transição no caso do *flip noise*.

$$\gamma_{uniforme} = \begin{pmatrix} 1-p_e & \dots & \frac{p_e}{n_y-1} \\ \vdots & \ddots & \vdots \\ \frac{p_e}{n_y-1} & \dots & 1-p_e \end{pmatrix} \qquad \gamma_{flip} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \dots & \dots & \dots & \vdots \\ 0 & \dots & 1-p_e & \dots & p_e & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & \dots & \dots & 1 \end{pmatrix}$$

No experimento, p_e adotou valores de 0% até 90%, em intervalos de 10%. Para o *flip noise*, os pares escolhidos foram os mesmos usados em Patrini et al. [2017] e Zhang and Sabuncu [2018] no MNIST e no Fashion-MNIST, respectivamente. Para o MNIST, os pares (2,7), (3,8), (7,1), (5,6) e (6,5) foram escolhidos e para o Fashion-MNIST os pares (Bota de Tornozelo, Tênis), (Tênis, Sandália), (Pulôver, Camisa), (Casaco, Vestido) e (Vestido, Casaco) foram escolhidos.

Para o treinamento, alguns hiperparâmetros foram definidos, sendo eles o otimizador, o tamanho do lote, o número de épocas, a função de custo, um critério de parada prematura do treino e uma porcentagem do conjunto de treinamento para o conjunto de validação, obtida via *holdout*. Para o otimizador, foi usado o *Adaptive Moment Estima-*

tion (*Adam*), por ser um dos mais utilizados. Para o tamanho do lote, o valor foi de 128, o mesmo de Rolnick et al. [2017]. Para o número de épocas, o valor foi de 300, seguindo o valor de 200 no artigo do Rolnick et al. [2017] adicionando mais 100, para garantir o treino no Fashion-MNIST. Para a parada prematura, foi definido um critério com limite de 20 épocas sem melhora no desempenho no conjunto de validação, para evitar a criação de modelos com sobreajuste. Para o conjunto de validação, 10% do treino foi usado.

O protocolo de avaliação do desempenho de uma rede sob um valor de p_e consiste em executar o *holdout* dez vezes, em que cada uma das vezes será calculada a acurácia do modelo no conjunto de teste. Por fim, a média aritmética entre as acurácias dos modelos treinados com os *holdouts* será calculada para determinar a acurácia final da rede. Para auxiliar as análises, o desvio padrão da acurácia final também foi obtido.

O experimento proposto consiste na variação da quantidade de neurônios da RNA. Foram escolhidas duas RNAs bases com boas acurácias nos seus artigos de origem, sendo que uma com uma camada oculta com 800 neurônios do trabalho de Simard et al. [2003] para o MNIST e outra com uma camada oculta de 100 neurônios para o Fashion-MNIST em Xiao et al. [2017], e a partir desses valores base a quantidade de neurônios foi variada. Foram adotadas as funções de ativação ReLU e Softmax para a camada oculta e de saída, respectivamente, para suprir a falta de informações na fonte com relação a esses parâmetros. Para o MNIST, foram usadas as quantidades de neurônios: 10, 15, 25, 35, 50, 75, 100, 200, 400, 600, 800, 1000, 1200, 1400 e 1600. Para o Fashion-MNIST, foram usadas as quantidades de neurônios: 10, 25, 50, 75, 100, 300, 500, 700 e 900.

4. Resultados

As Figuras 1 e 2 mostram o resultado do experimento da variação de neurônios para o MNIST e o Fashion-MNIST, respectivamente.

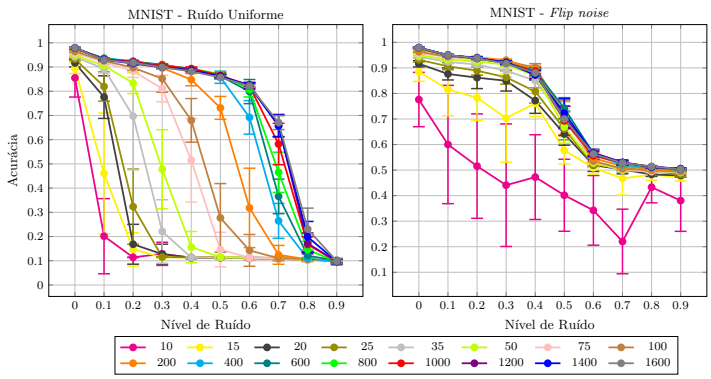

Figura 1. Acurácias resultantes da variação dos neurônios a partir de uma RNA de 800 neurônios para o MNIST.

Pode-se observar um comportamento bem interessante nos gráficos do ruído uniforme: a medida que se aumenta o número de neurônios, percebe-se que os gráficos dos desempenhos dos modelos vão se tornando menos íngrimes com relação ao gráfico anterior a ele, dado a ordem de crescimento dos neurônios. Nesta ideia, se tornar menos íngrime significa que a diferença do erro das acurácias entre os níveis de ruído adjacentes vai diminuindo. Isto significa que o modelo vai ficando mais robusto a medida que o número de neurônios aumenta para as bases de dados consideradas no ruído uniforme.

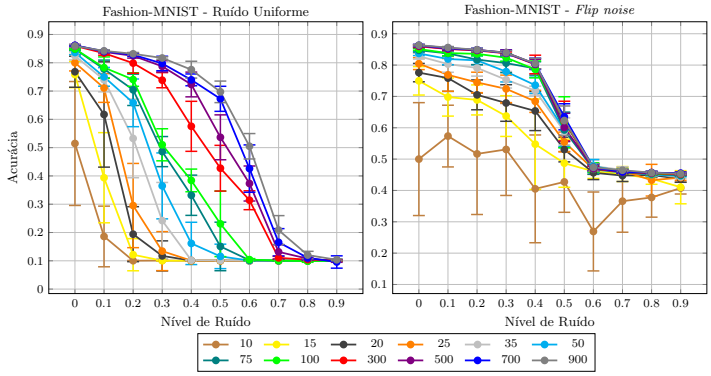

Figura 2. Acurácias resultantes da variação dos neurônios a partir de uma RNA de 100 neurônios para o Fashion-MNIST.

Quando analisamos os gráficos do *flip noise*, assim como no ruído uniforme, pode-se ver um comportamento incremental de melhora nítida na robustez quando a complexidade do modelo vai aumentando desde a quantidade de 10 neurônios até 50 neurônios para o MNIST e 100 para o Fashion-MNIST, devido ao gráfico assumir um formato menos íngrime e constante quando vai aumentando a quantidade de neurônios.

Considerando os gráficos do *flip noise*, é interessante notar o comportamento peculiar da RNA de 10 neurônios do MNIST e do Fashion-MNIST, com quedas e subidas irregulares com o aumento do ruído além dos valores altos dos desvios padrões, mostrando a dificuldade da rede em aprender quando aumenta o ruído aliada a questão da baixa complexidade da mesma. Depois dos 50 neurônios para o MNIST e dos 100 para o Fashion-MNIST, os gráficos começam a se interpolar com maior intensidade, mostrando que suas acurácias são muito próximas e por causa disto não é possível averiguar uma certeza de melhora ou não sempre que a quantidade de neurônios aumenta.

A quantidade de neurônios para uma RNA define a sua complexidade, ou seja, quanto mais neurônios, maior será a sua capacidade de aprender funções mais complexas, pelo aumento de parâmetros livres a serem treinados dado o aumento dos neurônios. Intuitivamente, as fronteiras de decisão que o modelo complexo pode aprender são mais maleáveis a ponto de ainda separar as classes corretamente como se estivesse desconsiderando os exemplos ruidosos. Frénay and Verleysen [2013] pontuaram como uma das consequências do ruído de classe sendo justamente o aumento da complexidade dos modelos preditivos quando há a presença do mesmo nos dados, com trabalhos da literatura mostrando resultados reforçando este fato e agora também o resultado obtido neste experimento. Em suma, este experimento trouxe a seguinte conclusão: quando a complexidade da RNA aumenta, a robustez do mesmo pode melhorar.

5. Conclusões

Neste trabalho, um experimento foi realizado em duas bases de dados bem conhecidas na literatura com o fim de analisar o efeito do ruído de classe em modelos baseados em RNAs, criando e testando estes modelos variando o número de neurônios. O experimento exemplificou que o ruído de classe pode afetar negativamente o desempenho das RNAs, expondo a relevância de tratá-lo nas bases de dados. Ademais, o resultado experimental ajudou a dar um indício de que a mudança de hiperparâmetros específicos da RNA tem a

sua importância na busca de uma robustez natural nas RNAs.

O trabalho possui algumas limitações como a restrição das bases de dados usadas, a falta do uso de um ruído do tipo NNAR e o intervalo de ruído considerado. Sendo assim, os trabalhos futuros são voltados na utilização de outras bases de dados, para permitir a análise em outros contextos de problemas e eliminar mais o viés da análise voltada para as bases de dados usadas, no uso de um ruído do tipo NNAR, para englobar o caso mais geral da taxonomia do ruído de classe e no aumento da granularidade do intervalo de ruído, que poderia ajudar em resultados mais assertivos no caso do *flip noise*.

Referências

Algan, G. and Ulusoy, I. (2020). Label noise types and their effects on deep learning. *arXiv preprint arXiv:2003.10471*.

Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I. W., Kwok, J. T., and Sugiyama, M. (2020). A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*.

Nettleton, D. F., Orriols-Puig, A., and Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33(4):275–306.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952.

Rolnick, D., Veit, A., Belongie, S., and Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Rusiecki, A. (2020). Standard dropout as remedy for training deep neural networks with label noise. In *International Conference on Dependability and Complex Systems*, pages 534–542. Springer.

Russell, S. J. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson, global edition.

Simard, P. Y., Steinkraus, D., Platt, J. C., et al. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Icdar*, volume 3.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. (2020). Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.