Systematic Optimization of Overfitting Problem in Machine Learning

Ziling Zhu*

School of Chien-Shiung Wu, Southeast University, Nanjing, 211189, China

* Corresponding Author Email: 213213248@seu.edu.cn

Abstract. The significant challenge of models in machine learning not being able to generalize sufficiently from seen to unseen data is known as overfitting. Overfitting occurs when a model exhibits perfect accuracy on its training dataset but performs poorly on new, test datasets. This issue not only makes the model overly complex and increases computational costs, but it also leads to its instability. Overfitting can be tackled through various methods, which are extensively discussed in the present article. Specifically, the article focuses on three main strategies: regularization, ensemble learning, and data augmentation. Each of these approaches is introduced with a brief explanation of their foundational principles and techniques. Moreover, the practical impact of these strategies is illustrated through several recent case studies across different fields such as text processing, agriculture, and healthcare. These approaches have been shown to significantly enhance the accuracy of models compared to traditional training methodologies, demonstrating their effectiveness in real-world applications.

Keywords: Machine learning, Overfitting, Regularization, Ensemble learning, Data augmentation.

1. Introduction

The inability of models to generalize adequately from observable data to unobserved data is a challenge in machine learning. We refer to this as overfitting. Overfitting is the reason behind the model's flawless performance on its training set and poor fit on its test set. Rather than discovering the underlying patterns, overfitting models have a tendency to memorize all the data, even the unavoidable noise on the training set [1]. Overfitting will increase the complexity and computational cost of the model, and may also lead to model instability.

There are many ways to solve overfitting. Simplifying model complexity is a common method, such as reducing the number of neural network layers or neurons. Increasing the amount of training data helps the model learn more general patterns, and more training data can be generated through techniques such as data enhancement. Regularization techniques such as L1 and L2 regularization can effectively prevent overfitting and balance training error and generalization ability by penalizing model complexity. The early stopping method stops training in advance by monitoring the performance of the verification set, which can also effectively avoid overfitting. Ensemble learning improves generalization by combining predictions from multiple models. Feature selection helps to remove irrelevant or redundant features and reduce the risk of overfitting.

This article mainly focuses on three methods: regularization, ensemble learning and data augmentation. Their principles and implementation methods are briefly introduced. And through several specific application examples in recent years, the effectiveness of these methods in solving practical problems such as medicine, agriculture, and text processing is demonstrated.

2. Theoretical Basis

2.1. Cause

In machine learning, overfitting is when the algorithm fits too closely to parts of the training data, it cannot fit other data well or make accurate predictions. The fundamental cause of overfitting is when the training algorithm interprets the noise in the training data as information and uses it to inform the model's structure's parameters.

The reason why there is the possibility of overfitting is the difference between the criteria for selecting a model and the criteria for evaluating the model. Overfitting can happen when a model tries to memorize every piece of training data, even the noise in the data set, instead of identifying patterns in the data [1]. When the degree of freedom of parameters or the complexity of the model structure exceeds the information content contained in the data, the fitted model may use any number of parameters, which will reduce or destroy the generalizing ability of the model. The effect of overfitting is shown in figure 1.

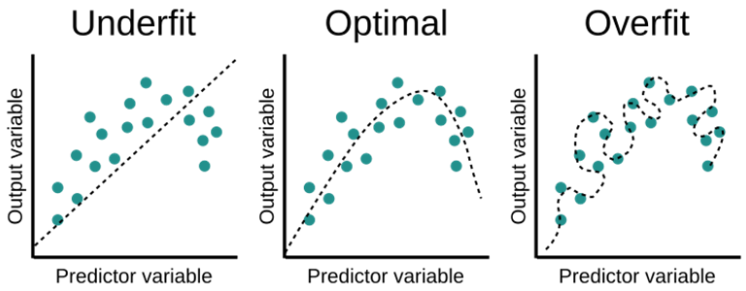


Fig. 1 Underfit, optimal and overfit (Photo/Picture credit: Original)

There are several kinds of situations in machine learning that can lead to overfitting. Because there are insufficient data samples and a short training data set, the potential input data values are represented incorrectly. The training data, sometimes referred to as noise data, contains a significant amount of unrelated information. The model requires an excessive amount of time to train on a single data sample set. Because of the model's complexity, it is possible for the noise in the training set to be learned.

Suppose there is a training set that can be predicted well by a linear function. Obviously, this function has only three parameters: one intercept and two slopes. Replacing this function with a more complex quadratic function, or a more diverse linear function, may lead to overfitting. Occam's razor states that the predictions of any more complex function are less reliable than those of a simple function. When fitting the training data, the benefits brought by the more complex function do not offset the increase in model complexity, the complex function has overfitted the data. At this point, the complex function may perform worse on the test set even while its performance on the training set is now on par with or even better than the simple one. The connection between the mistake and the complexity of the model is shown in figure 2.

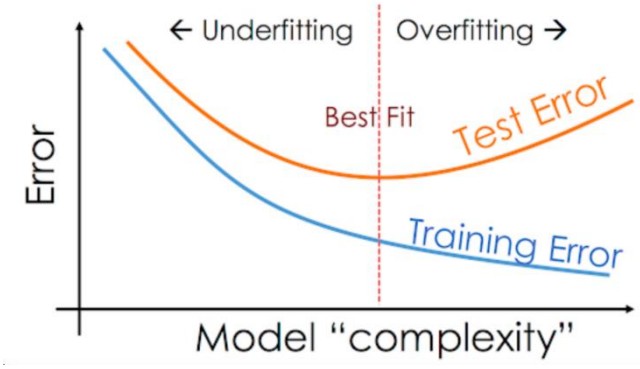


Fig. 2 Model complexity and error (Photo/Picture credit: Original)

However, there is research indicating that when there is no strong relationship between the set of predictor variables and the outcome, Overfitting may pose an issue for data that is low-dimensional and less complex [2].

2.2. Performance and Influence

Overfitting is evident when there are low error rates and significant variance. It is customary to have a portion of the training data set named "test set" in order to check the overfitting and stop the

tendency. If the training data has a low error rate and the test data has a high error rate, overfitting is likely to have occurred.

Take a look at a use case where a machine learning model needs to distinguish between the looks of men and women based on photo analysis. An overfitting model may collect not only face features but also hair features. It will associate a man's face with short hair and a woman's face with long hair. Therefore, both women who have short hair and men who have long hair are unable to be identified correctly.

The most obvious consequence of overfitting is poor performance on the test set. Also, compared to a properly fitted model, am overfitted model tends to obtain more information from each sample in the validation set. Collecting this unnecessary information may be costly or error-prone. This cost is especially obvious when this information requires manual observation or annotation. Complex models that are overfitted are less transferable than simpler models. Extremely complex models can only be reproduced on the original data set, which makes it difficult to reuse the model and reproduce the theoretical research.

3. Solution

3.1. Regularization

Regularization is a way to avoid overfitting. Any adjustment to an algorithm that lowers generalization mistakes as opposed to training mistakes is referred to by this term. An effective regular function can greatly reduce the variance without increasing the error.

Regularization scores features according to their importance, thereby eliminating factors that do not affect the prediction results. The feature with the least impact is assigned a penalty value based on a mathematical computation. Think about a statistical model that aims to forecast a city's future home prices. In the regularization, attributes like average annual income and population growth are assigned a lower penalty value, while the city's average annual temperature is assigned a greater penalty value. The effect of regularization is shown in figure 3.

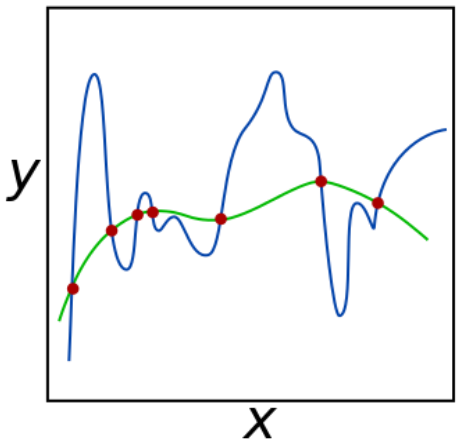


Fig. 3 Regularization (Photo/Picture credit: Original)

For machine learning, there are Two types of regularization.

Regularization using L_1 norm:

$$J(\theta) = \frac{1}{2m}\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}|\theta_j|\right) \quad (1)$$

Regularization using L_2 norm:

$$J(\theta) = \frac{1}{2m}\left(\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right) \quad (2)$$

Where there are m data points $\left(x^{(i)}, y^{(i)}\right), i = 1,2, \ldots, m$; λ is called the regularization parameter; $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$ is the hypothesis function; $J(\theta)$ is the loss function.

In these two regularization methods, to regulate the "attenuation of the weight" and reduce the weight, the regularization parameter term is appended to the goal function. When the parameter is larger, the punishment (normative) will be greater, and the more normative effect it can play. However, if the regularization parameters selected are too large, all parameters will be minimized, resulting in underfitting. Therefore, the selection of λ needs to be reasonable.

The following are some applications of regularization in machine learning. To regularize the model, researchers propose Graph-based Local Resampling of perceptron-like neural networks with random projections (RN-ELM). Using the graph-based method, resampling can be done directly in the margin area rather than thoroughly throughout the input space. The goal is to use structural information retrieved from a planar graph to identify the excess noise in the margin region, which is used to train neural networks. The comparison of RN-ELM with ELM and ELM-REG was conducted for a series of datasets containing real-world categorization difficulties. Regularization did not improve performance on networks with few hidden neurons. Regularization, however, becomes important when there are more than 100 hidden neurons, as is the case with most datasets, where RN-ELM and ELM-REG perform better than ELM. In addition, it can be seen that the suggested approach produced outcomes that were comparable to those of ELM-REG [3].

The weight penalty standard can be chosen anywhere between $\{L\,1,\ L\,2\}$; it depends on the specific problem. The Bridgeout stochastic regularization approach is proposed and shown by researchers, who also show that it is equal to a Lq penalty on the weights. The norm q is a hyperparameter that can be learnt from the data. Datasets MNIST, Fashion-MNIST, and CIFAR10 are used in the paper's evaluation of Bridgeout for picture classification. When evaluated on neural network-based image classification tasks, Bridgeout's flexibility and sparsity-inducing qualities exceed Dropout and Shakeout in the classification accuracy [4].

To lessen the overfitting issue, researchers suggest optimizing the deep CNN's feature boundary using a two-stage training approach (implicit regularization training process and pre-training procedure). To extract the visual representation for anomaly detection, they construct a network model in the pre-training phase. To make the feature boundaries more regular and bring the network to convergence at the desired position during the implicit regularization training phase, they retrain the network using the anomaly detection findings. They use the MNIST, CIFAR10, and CIFAR100 datasets for two sets of tests. When comparing the outcomes of two training methods on three datasets, It is evident that the two-stage training approach outperforms common training methods in terms of classification accuracy for nearly all initialization scenarios [5].

3.2. Ensembling

Ensemble learning is a method of generalizing data by letting multiple models learn. Because they frequently provide erroneous findings, certain models are referred to as poor learners. All weak learners are combined using ensemble techniques to provide more accurate results. These techniques examine sample data using several models and choose the most accurate outcomes.

There are two methods of integration: bagging and the boosting. Bagging algorithm learns from multiple models simultaneously and averages the prediction results to improve the performance of generalization of the model prediction results. Boosting algorithms improve generalization performance by generating corresponding models based on the model's prediction results. Additionally, there is the stacked generalization approach, which is an ensemble learning framework that combines the predictions of two or more ensemble members by training a different machine learning algorithm [6].

Ensemble learning techniques have demonstrated superior performance in a wide range of applications, garnering significant interest across various academic domains. These uses of ensemble methods—medical diagnosis, fraud detection, and sentiment analysis—have recently taken center

stage in the field of applied machine learning. Traditional machine learning algorithms struggled to get good results in each of these application areas due to various issues. As an example, consider the small sample sizes in medical datasets and the difficulty sentiment analysis has in interpreting the ambiguity of human language [6].

Simplified Acute Physiology Score (SAPS) and Acute Physiology and Chronic Health (APACHE), they are two ICU scores that may be used to assess the severity of an illness or predict death. However, it is not able to predict readmission from the ICU. In this paper, a fog computing system for clinical monitoring—which enables remote patient monitoring and prognosis—is presented. Different machine learning models are used to predict readmission after 30 days after ICU discharge. The recommended stacking models performed the best when contrasted with other models. When compared to complete features and selected features, positive outcomes were obtained using the recommended stacking ensemble model with particular characteristics chosen by POS [7].

Another aspect of application is the categorization of land uses using many machine learning techniques. For the purpose of assessing climatic changes, precise land use or land cover (LULC) classifications are crucial in arid areas. In Xinjiang, researchers suggest using ensemble learning to increase the accuracy of LULC categorization. The study region was first split into Northern Xinjiang, Tianshan, and Southern Xinjiang using multisource geographical datasets. Second, based on the topography and climate of each subregion, several ensemble learning strategies were developed using five algorithms for machine learning: k-nearest neighbor, support vector machine (SVM), random forest (RF), artificial neural network (ANN), and C4.5. Third, to acquire training examples and the best parameters, stratified random sampling was employed. Finally, the performance of each generated strategy was assessed across the province of Xinjiang. Compared to individual machine learning algorithms, the proposed ensemble method improved the accuracy of LULC across Xinjiang by ≥6.85%. Tianshan, Southern Xinjiang, and Northern Xinjiang had increases in accuracy of ≥6.70%, 5.87%, and 6.86%, respectively [8].

3.3. Data Augmentation

The technique of creating artificially generated data from preexisting data is known as data augmentation, and it is mostly used to train new machine learning models. Insufficient data is often encountered in machine learning. However, we don't always need to add new data. The current data set just needs to be slightly modified. For instance, little adjustments like rotations, flips, and translations. This will appear to the network as a separate image.

Suppose we have a dataset of car images, but all the cars have their fronts facing left. When a picture of a car with the front facing right is input for testing, it may not be recognized. To avoid this, an easy way is to flip the picture of the car. We can stop the neural network from learning unimportant characteristics by enhancing the data, which will significantly increase overall performance. The example of overfitting is shown in figure 4.



Fig. 4 Data augmentation (Photo/Picture credit: Original)

Since the augmentation effect in previous systems is frequently global, i.e., the entire image is subjected to one enhancement treatment, the variety of regional characteristics in augmented images is restricted. A novel approach to data augmentation is presented by the researchers, dubbed Local Augment (LA), which dramatically modifies the attribute of local prejudice to produce a wide range of enhanced images and provide a superior augmentation effect for the network. Choose a small number of local patches in a picture, and then give each local patch a distinct augmentation strategy. More thoroughly, the network may learn the local bias characteristic thanks to the locally diverse samples produced by this augmentation procedure. The paper evaluates the suggested LA with other representative augmentation methods using three datasets—ImageNet, STL10, and CIFAR100—in order to confirm the efficacy of the suggested method. The suggested approach performs competitively on CIFAR100 and beats the most recent information augmentation strategies on ImageNet and STL10 [9].

Here is an example that data augmentation is applied to agriculture. Common issues regarding the domain of intelligent farming include high costs and challenges in getting disease samples of sufficient quality. In order to address this issue, scientists developed a dual generative adversarial network (GAN) based high-quality image augmentation (HQIA) technique for producing high-quality photos of rice leaf disease. With this method, the original picture is created using the image creation GAN network, and the quality is then enhanced by applying the image super-resolution GAN network. This technique creates photos with high quality by combining image production with super-resolution GAN networks. For the dataset, Three common diseases of rice—brown spot, leaf blast, and bacterial blight—were chosen. According to the experimental results, compared to not employing data augmentation, the HQIA approach in this work increased 4.1% accuracy with VGG11 and 4.57% with ResNet18. When utilizing ResNet18 and VGG11, respectively, the accuracy increased by 3.08% and 3.55% in comparison to the conventional WGAN-GP data augmentation method [10].

Another use example is text data augmentation for word embeddings. Large corpora are a major training resource for modern language models. A model is better able to capture different semantic associations the larger the corpus. The restricted scope of the employed corpora is the problem at hand. Using data augmentation to increase the size of the current corpus can be one possible answer to this issue. A variety of corpus augmentation approaches are included in data augmentation. Scholars conduct a more thorough examination of three methodologies: Reduction of Function Words, Back Translation, and Synonym Replacement. By applying these three methods, we were able to create various iterations of the corpus that was used to train Word2Vec Skip-gram models. Significant statistical disparities between the supplemented and original corpora's classifier outcomes are highlighted in the study. In particular, accuracy is greatly improved by back translation, especially when using Bernoulli Naive Bayes and Support Vector models. Although the initial collection does well in Random Forest classification, Logistic Regression is improved by the Reduction of Function Words (FWD) [11].

4. Conclusion

Three approaches are mostly discussed in this article: data improvement, integration, and regularization. Regularization eliminates elements that have no bearing on the prediction outcomes by assigning a score to each feature based on importance and applying a penalty value to the item with the least impact. Several models can learn from each other to generalize data using ensemble learning. These techniques integrate all of the weak learners to provide results that are more accurate by using numerous models to assess sample data and select the most correct findings. The technique of creating new data artificially from preexisting data is known as data augmentation. The network will see a slight alteration to the current data set—such as flips, translations, rotations, etc.—as a distinct image. Through several specific application examples in recent years, the effectiveness of these methods in solving practical problems is shown. However, solving the overfitting problem still faces many challenges. It is necessary to comprehensively consider the model's complexity, the

characteristics of the data, the selection of regularization parameters, the application of ensemble learning, and feature selection and processing. Hope there will be more developments and applications in the future.

References

[1] Ying X. An overview of overfitting and its solutions. Journal of physics: Conference series. IOP Publishing, 2019, 1168: 022022.

[2] Subramanian J, Simon R. Overfitting in prediction models–is it a problem only in high dimensions?. Contemporary clinical trials, 2013, 36(2): 636-641.

[3] Assis A D, Torres L C B, Araújo L R G, et al. Neural networks regularization with graph-based local resampling. IEEE Access, 2021, 9: 50727-50737.

[4] Khan N, Shah J, Stavness I. Bridgeout: stochastic bridge regularization for deep neural networks. IEEE Access, 2018, 6: 42961-42970.

[5] Zheng Q, Yang M, Yang J, et al. Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process. IEEE Access, 2018, 6: 15844-15869.

[6] Mienye I D, Sun Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 2022, 10: 99129-99149.

[7] Alabdulhafith M, Saleh H, Elmannai H, et al. A clinical decision support system for edge/cloud ICU readmission model based on particle swarm optimization, ensemble machine learning, and explainable artificial intelligence. IEEE Access, 2023.

[8] Du H, Li M, Xu Y, et al. An ensemble learning approach for land use/land cover classification of arid regions for climate simulation: A case study of Xinjiang, Northwest China. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 2413-2426.

[9] Kim Y, Uddin A F M S, Bae S H. Local augment: Utilizing local bias property of convolutional neural networks for data augmentation. IEEE Access, 2021, 9: 15191-15199.

[10] Zhang Z, Gao Q, Liu L, et al. A high-quality rice leaf disease image data augmentation method based on a dual GAN. IEEE Access, 2023, 11: 21176-21191.

[11] Kapusta J, Držík D, Šteflovič K, et al. Text Data Augmentation Techniques for Word Embeddings in Fake News Classification. IEEE Access, 2024, 12: 31538-31550.