The Investigation Related to the Influence of Different Parameters Based on Random Forest for Weather Prediction

Zhuoxing Yu *

Information Management & Information Systems, Southeast University, Nanjing, China

* Corresponding Author Email: guangpu.luo@pumaenergy.com

Abstract. The weather change is important in life for providing useful information that guides people's daily activities. However, in a day that machine learning is used in many fields, the weather prediction using machine learning still needs to be developed. This article uses random forest, a machine learning method that is used in classification to do the weather prediction. By changing the parameters in different values and using the evaluate metrics, this article compared the performance about random forest in different parameters. There are 4 parameters researched in this article. The first parameter is the number of trees. In this situation, the f1-score, recall and precision by the categories have the same trend with accuracy. However, the evaluate metrics about result of 1 is obviously worse than the metrics about result of 0. The second parameter is the size of data used to train. In this situation, the performance when the splitting data size is 0.1 and 0.15 is better than the other. The third parameter is the way of encoding. In this situation, using the order of the probability of rain is better than with no order or using the probability to represent the location. Fourth is to delete the unimportant features by the sort of important score. In this situation, when 4 features are deleted, the performance is worse than any others.

Keywords: Weather prediction, Random forest, Machine learning.

1. Introduction

The weather prediction is using the collected data of weather to predict the weather in future. A lot of features of weather in future can be predicted, such as temperature, humidity, the type of weather and so on. Some features can be used to predict and be predicted. Weather prediction can help to learn the future weather and plan by using it. For example, people will bring an umbrella if they know it will rain. Furthermore, some disasters can be predicted in advance, and thus reduce the loss of casualties and property.

From ancient times to now, human beings used numerous ways to predict weather. The ancients used the shape of clouds and calendar to predict weather. In the last centuries, human beings used mathematical ways to predict weather. For instance, Thompson developed an equation for numerical weather predictions [1]. Lorenc used Bayesian probabilistic arguments to find the best analysis for numerical weather predictions [2]. The discussion in this study included variational techniques, smoothing splines, Kriging, optimal interpolation, successive corrections, constrained initialization, the Kalman-Bucy filter, and adjoint model data assimilation [2]. Lorenz developed an n-layer model about equations to research the general circulation [3]. For now, as the development of Artificial Intelligence (AI), the effectiveness of machine learning algorithms has been demonstrated in many fields [4-6]. A lot of situations about prediction are using machine learning for now. Konstantina et al using machine learning in cancer prognosis and prediction [7]. Jochen employed machine learning methods to estimate and predict risk [8]. Weather prediction is also a type of prediction. A lot of researchers are trying to apply machine learning in weather prediction currently. There are several types of machine learning. Kareem used Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN) to predict weather [9]. Du et al tried to predict weather by using support vector machine and particle swarm optimization [10]. Singh used random forest to predict the weather [11]. However, according to an article written by Schultz, a method of machine learning that suitable for weather prediction still needs to be developed [12]. Although Jaseena compared a lot of methods about machine learning in the field of weather predicting [13], the deepen discussion about important

parameters of a certain model including the number of features, the way of encoding, the size of training data, the parameters of model, is lacked.

This article is focused on the performance of random forests, one method of machine learning, in the field of predicting weather. A dataset of weather in Australia which is in Kaggle will be used to train the model. This data set including a lot of data about weather, such as temperature, humidity, the speed of wind and so on. This article will try to change different parameters to train the model and will use some performance metrics to evaluate the model, including accuracy, f1-score etc. The goal of this article is to compare the performance between different parameters and try to find a better range of these parameters.

2. Method

2.1. Dataset preparation

This article used a dataset of weather in Australia collected from Kaggle [14]. The dataset includes 99, 516 records and 21 features are used to predict whether it will rain tomorrow. The labels 'yes' and 'no' are the target variable for prediction.

For being suitable to the training model, the dataset went through the preprocess procedure. Two features and some records are removed since they have numerous missing values. The locations are encoded from 0 to 48 while the direction of wind is encoded by the angel. For example, the north is 0 and the east is 90. The value 'NA' is encoded to -1, and the speed of wind in that record is set in 0 to refer that there is no wind at that day. The result 'yes' and 'no' are encoded in 1 and 0.

2.2. Machine learning-based weather prediction model

2.2.1. Introduction of machine learning and workflow

Machine learning is a method that uses data as an input and gives a result as an output by these data using a specific method. There are 3 types of machine learning, which represent the different ways the method is used to learn. The first type is supervised machine learning. Supervised machine learning uses data that contains the input and results to train, which is called training data. After the training finished, some data that remove the result, which is in order to simulate the actual use, would be used to test the performance of the model. Random forest, which will be introduced later, is supervised machine learning. The second type is unsupervised machine learning. Different from supervised machine learning, unsupervised machine learning uses the data that does not contain label. An example of unsupervised machine learning is cluster analysis. Cluster analysis is a method that classifies a lot of data by calculating the distance of each data. The third type is semi-supervised machine learning. Semi-supervised machine learning is between supervised machine learning and unsupervised machine learning just like the name. The semi-supervised machine learning uses the data contained label and non-label.

Machine learning work is usually composed of 5 steps. The first step is data collection. The second step is data preprocessing. The two steps are mentioned in the previous part. The third step is model building. The fourth step is using data to train the model. The last step is testing and evaluating the trained model. The three steps will be mentioned later. Figure 1 shows the workflow of machine learning.

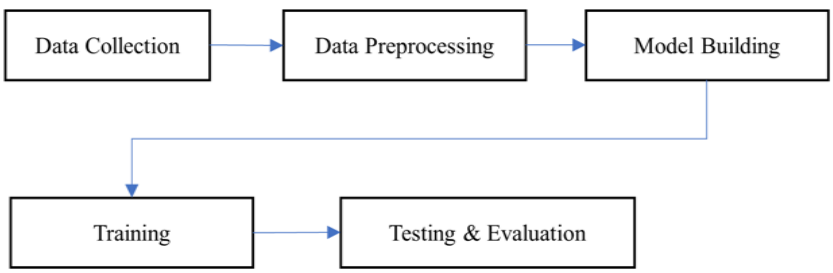

Figure 1. Machine Learning Workflow (Photo/Picture credit: Original).

2.2.2. Random forest

Random forest is a machine learning method used in classification. It is first proposed by Salzburg in 1993 [15]. In 1995, Ho developed the idea of random forest [16]. The current training method of random forest is developed by Leo Breiman [17]. In his article, Leo combined randomized node optimization and bagging, using classification and regression trees to build a forest [17].

Classification and regression trees are two types of decision trees. Decision tree is a method of machine learning. It can predict the result by using the given data. Like decisions made by human beings, the decision tree decides the result by judgement the condition. Figure 2 shows an example of part of a decision tree. For the given data, the decision tree will judge condition A to see whether it is bigger than 10. If A is bigger than 10, the decision tree will predict result 1. Otherwise, it will predict result 2.

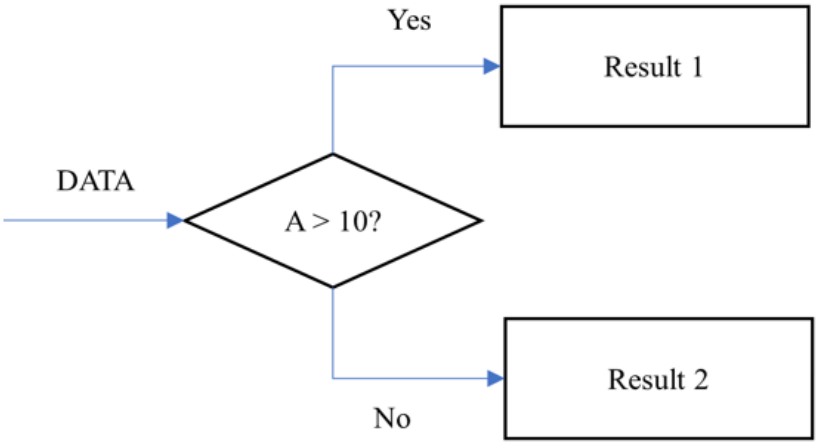

Figure 2. An example of part of decision tree (Photo/Picture credit: Original).

It is more complex in practice. The decision tree will use more than 1 condition. The order of these conditions is decided by a metric called 'entropy'. It can be calculated by a formula. What it measures is the ability to reduce uncertainty.

The random forest is a collection of many decision trees. By sampling randomly from the training set, the model can get a collection of different decisions trees. The algorithm is called bagging. When using the random forest to predict, each tree will predict independently. By calculating the average of each result, the random forest will produce a result to represent all trees.

In this article, the Python library called Sklearn will be used for building the model. Scikit-learn (often abbreviated as sklearn) is a powerful and widely-used open-source machine learning library in Python, designed for simplicity and efficiency. It provides a range of tools for data analysis, including algorithms for classification, regression, clustering, and dimensionality reduction. Scikit-learn is built on top of popular Python libraries such as NumPy, SciPy, and Matplotlib, making it easy to integrate with other data science workflows. Its intuitive API and comprehensive documentation make it accessible to both beginners and experienced practitioners, enabling them to quickly develop and deploy machine learning models in various domains. The evaluation metrics used in this article are also from the same library. To compare the performance of different parameters, the accuracy, one of the evaluation metrics will be used. In the comparation of different n_estimators, a parameter in the function which means the number of trees, the accuracy, the f1-score and the recall will be used to evaluate the model.

Feature importance is a metric which shows the score of importance about each feature. This article will show the feature importance, and compare the different performance when the least 1/2/3/4 important features are deleted. Besides these, this article will also compare the performance in different size of training data and different way to encode the value of location. This article uses 3 different ways to encode the value of location. First is encoding it by the default sort of the dataset. Second is encoding it by the frequency of the rain day in the location in last year. The third way is encoding it by the sort of the frequency.

3. Results and discussion

3.1. The influence of number of trees in model performance

Figure 3 shows the model performance when changing the number of trees. This figure shows not only the accuracy, but also other evaluation metrics including precision, recall and f1-socre. Each evaluation metric, except accuracy, is calculated independently by the category of target variables. As shown in Figure 3, all the evaluation metrics improved as the number of trees increased to 100. After the number of trees increased to 100, the evaluation metrics became stable and did not change dramatically, and all the metrics have the same trend with accuracy. However, the result of 0, which means that it will not rain tomorrow, is more stable than the result of 1, which means that it will rain tomorrow. Furthermore, the result of 0 gets a high score in all evaluation metrics while the result of 1 gets a low score. One possible reason is the data size of the result of 1 is smaller than the data size of the number of 0. There is a total of 67, 776 records that are the result of 0, while only 19, 279 records are the result of 1.

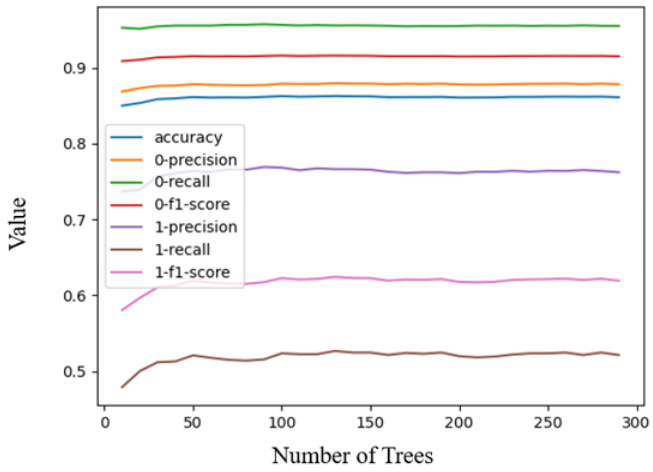

Figure 3. The influence of number of trees in model performance (Photo/Picture credit: Original).

3.2. The influence of size of testing dataset in model performance

Figure 4 shows the accuracy when changing the size of test data. The number of trees increased to 200 in this situation. As shown in the Figure 4, the accuracy when the ratio of test data bigger than 0.25 is lower than the ratio smaller than 0.25, and the performance when the ratio of test data is 0.2 is worse than the ratio as 0.1 and 0.15, but it still better than the ratio bigger than 0.25. One possible reason is that when increasing the ratio of test data, the size of data that is used to train becomes smaller and cannot provide the enough data to train.

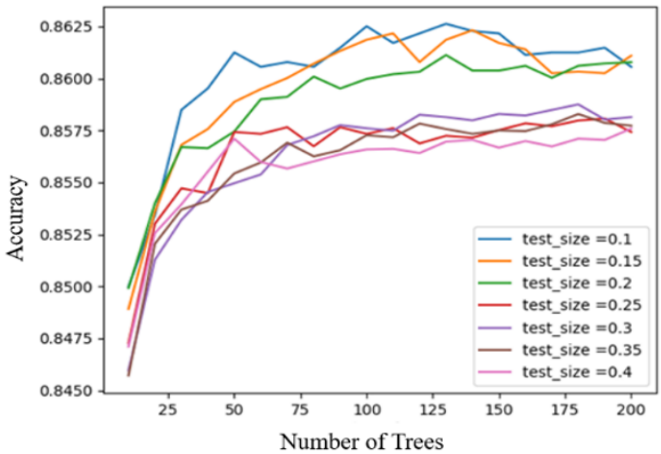

Figure 4. The influence of size of testing dataset in model performance (Photo/Picture credit: Original).

3.3. The influence of ways of encoding in model performance

Figure 5 shows the accuracy when changing the size of the way of encoding the location. In this research, there are 3 ways to encoding to try. The first way is to encode the location without sequencing. The locations are encoded from 0 to 48 by the time they first appear in the data sheet. The second way is to use the probability of rain to refer to the location. The probability of rain is calculated by the days of rain in the last year. The third way is to encode the location by the sequence of the probability calculated in the second way. As shown in Figure 5, their accuracy is close to each other, but the accuracy about third way is better than others in most time. Another noteworthy point is that the second way performed better than the first way in the beginning, but worse when the number of trees increased.

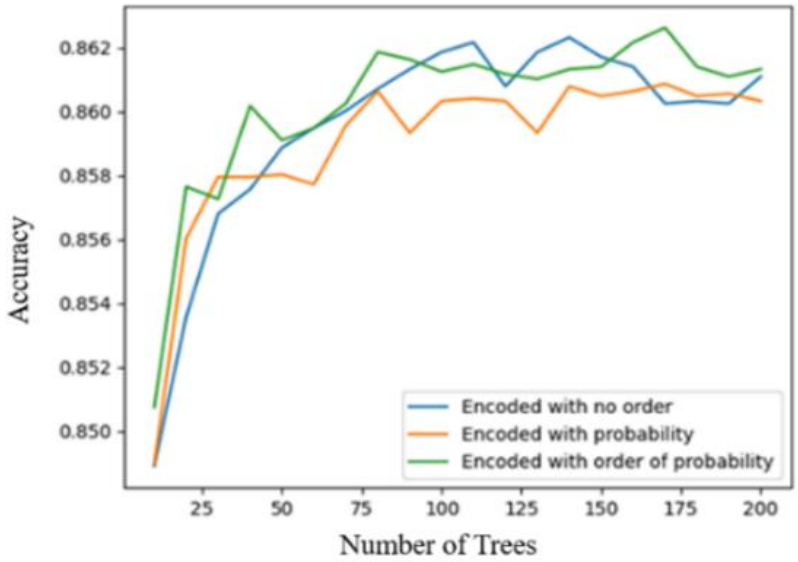

Figure 5. The influence of ways of encoding in model performance (Photo/Picture credit: Original).

The random forest uses decision trees to predict the result, and the decision tree predicts the result by the judgement of the data. As shown in Figure 2, the decision tree will judge the data by bigger or smaller a specific value. When encode the location in the first way, the value in the data of location is meaningless, and will lead to a lot of branches in the decision tree since the decision tree will try to divide them one by one. However, when the way of encoding used the second and third way, the value of location has meaning. The number of branches about this feature will decrease since the decision tree only needs judgement whether it is bigger than a specific number or not, as what the decision tree does to other features. Hence, the third way is performed better than the first way. However, the second way is not performed better than the first way as expected after the number of trees increased. One possible reason for this strange result is the way of calculating the probability. The data used to calculate is not the data used to train the model. However, the order of the probability is stable each year. Therefore, the third performed better than expected.

3.4. The feature importance and related influence in model performance

Figure 6 shows the importance score of each feature. Figure 7 is the result of accuracy when deleting features from 0 to 4, and the number of trees increased to 200. The feature is deleted by the order of the important score. The best performance of the model before the number of trees of 75 is delete 1 feature, while delete nothing performed better in later until 160. Deleting 1 feature or 3 features performed best after 160. Considering the important score, whether it is raining today is not so important, and can improve the performance in most time through deleting it.

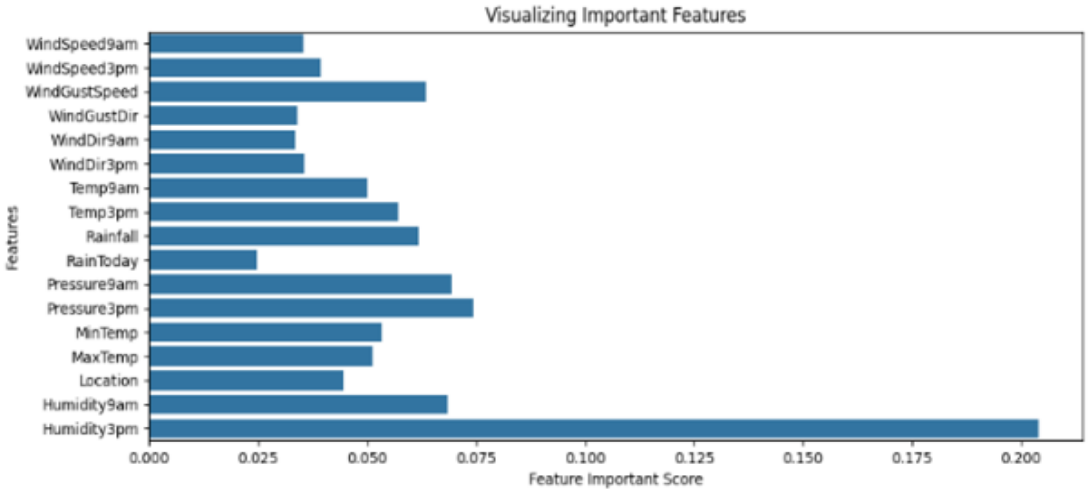

Figure 6. Importance score of features (Photo/Picture credit: Original).

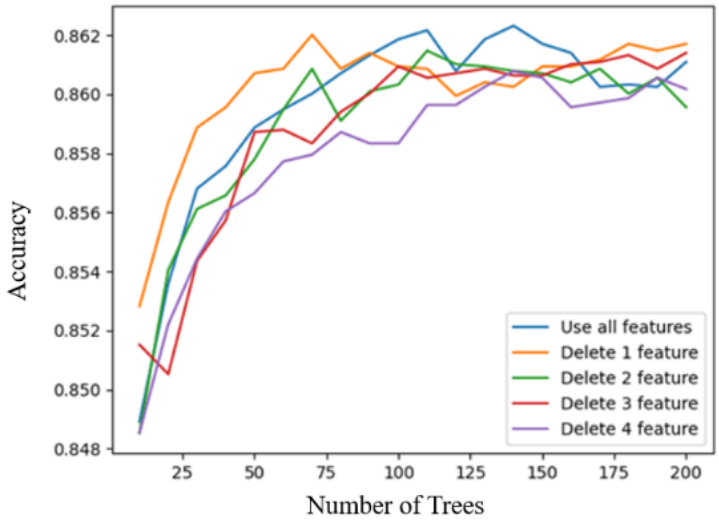

Figure 7. The result when deleting features (Photo/Picture credit: Original).

4. Conclusion

This article used random forest, a method of machine learning, to predict whether it will rain tomorrow or not. By changing the parameters, this article shows a better way to train the model to get a better performance of the model. However, there still has some work to do in the future. The data size of the result of 0 and 1 has a dramatic difference and thus get low evaluation metrics in the result of 1. The number of trees increases by 10 each time, and thus the curve is not smooth enough. The way of encoding the location can use the one-way encryption to see whether it will perform better than all the three ways shown in the article. Also, the reason why the humidity in 3pm has a high important score still needs to be determined.

References

[1] JSS. 1962. Numerical weather analyses and prediction. By PD Thompson. New York (Macmillan), 1961. Pp. xiv, 166; 18 Figures; 2 Tables; 49s. Quarterly Journal of the Royal Meteorological Society, 88(378), 560-561.

[2] Lorenc, A. C. 1986. Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society, 112(474), 1177-1194.

[3] Lorenz, E. N. 1960. Energy and numerical weather prediction. Tellus, 12(4), 364-373.

[4] Alzakari, S. A., Maashi, M., Alahmari, S., Arasi, M. A., Alharbi, A. A., & Sayed, A. 2024. Towards laryngeal cancer diagnosis using Dandelion Optimizer Algorithm with ensemble learning on biomedical throat region images. Scientific Reports, 14(1), 1-19.

[5] Popenici, S. A., & Kerr, S. 2017. Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and practice in technology enhanced learning, 12(1), 22.

[6] Srilakshmi, R., BalaKrishnan, S., & Vani, K. S. 2024. Revolutionizing Brain Tumour Detection: Integrating 3D U-Net-R Segmentation with Volume Analysis for high Diagnostic Accuracy. International Journal of Computing and Digital Systems, 16(1), 1-21.

[7] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. 2015. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

[8] Kruppa, J., Ziegler, A., & König, I. R. 2012. Risk estimation and risk prediction using machine-learning methods. Human genetics, 131, 1639-1654.

[9] Kareem, S., Hamad, Z. J., & Askar, S. 2021. An evaluation of CNN and ANN in prediction weather forecasting: A review. Sustainable Engineering and Innovation, 3(2), 148-159.

[10] Du, J., Liu, Y., Yu, Y., & Yan, W. 2017. A prediction of precipitation data based on support vector machine and particle swarm optimization (PSO-SVM) algorithms. Algorithms, 10(2), 57.

[11] Singh, N., Chaturvedi, S., & Akhter, S. 2019, March. Weather forecasting using machine learning algorithm. In 2019 International Conference on Signal Processing and Communication (ICSC) (pp. 171-174). IEEE.

[12] Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., ... & Stadtler, S. 2021. Can deep learning beat numerical weather prediction? Philosophical Transactions of the Royal Society A, 379(2194), 20200097.

[13] Jaseena, K. U., & Kovoor, B. C. 2022. Deterministic weather forecasting models based on intelligent predictors: A survey. Journal of king saud university-computer and information sciences, 34(6), 3393-3412.

[14] Kaggle 2022 Australia Weather Data https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data?select=Weather+Training+Data.csv

[15] Heath, D., Kasif, S., & Salzberg, S. 1993. K-DT: A multi-tree learning method. In Proc. of the Second Int. Workshop on Multistrategy Learning (pp. 138-149).

[16] Ho, T. K. 1995, Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

[17] Breiman, L. 2001. Random forests. Machine learning, 45, 5-32.