Meteorological Application Based on Big Data Streaming Computation Method

Xu Wang\textsuperscript{1,2,*}, HongTao Cheng\textsuperscript{1,2}, XiaoYan Liu\textsuperscript{1,2}, XiaoJie Wang\textsuperscript{1,2}

\textsuperscript{1}Key Laboratory of South China Sea Meteorological Disaster Prevention and Mitigation of Hainan, Hainan,570203
\textsuperscript{2}Hainan Province Meteorological Information Center, Haikou, Hainan, 570203

*Corresponding author e-mail: wangxu@163.com

Abstract. Big data analysis in the form of streaming computation has always been a problem to be solved at present, with relatively few research results and practical experience. The random forest method is currently the most extensively used classification algorithm. However, in the application scenario of streaming computation, real-time, volatility, and disorder features presented by data will lead to a gradual reduction in the accuracy of the algorithm. In this paper, the characteristics of the random forest algorithm are analyzed, and the idea of random forest pruning based on the accuracy of the decision tree is proposed. Meanwhile, to adapt to the changes in meteorological data, the concept of accuracy interval is combined to propose a method for the generation, verification, and supplementation of a new decision tree. Finally, a random forest that can be constantly updated with data is established to meet the requirements of the streaming big data environment for the algorithm. Actual meteorological data are used to verify the feasibility of the improved method. The results show that the new method has a higher classification accuracy in the real streaming big data scenario.

Keywords: Decision Tree, Random Forest Method, Big Data, Streaming Computation, Classifier, Pruning, Distributed System

1. Introduction

In each application scenario, the big data calculation mode can be divided into two types: batch calculation and streaming calculation. Batch calculation refers to the collection and memory of data first, and then the centralized calculation of the static data that has been stored is performed to identify the value of data. In the streaming calculation, it is impossible to determine the order and time of data arrival or store all historical data. Instead, the data are directly calculated in real-time in the memory after they flow in, with the output of valuable information.\cite{1-2} The research on big data batch
calculation technology is relatively more mature. For example, Google’s MapReduce model and other widely used systems are based on batch computation [3]. For scenarios where the accuracy and comprehensiveness of the output results are more significant, batch computation is preferred. In scenarios with higher real-time requirements where the data flow is high, the data flow is uncertain, and data accuracy requirements are slightly lower, streaming computation has apparent advantages. Compared with a large number of batch computation technology studies, there is less research on streaming computation. Early streaming computation studies focus on streaming data computation in a database environment [4]. However, with the growing demand for big data on the Internet, the streaming computation system that meets the requirements of real-time, burst and infinite analysis starts to emerge [5].

In this paper, the characteristics of streaming computation scenarios are analyzed [6]. The main technical characteristics of streaming computation technologies are discussed, and the random forest methods are combined to perform experiments in the meteorological industry to verify the practical feasibility of the method.

2. Algorithms in the streaming big data environment

2.1. Ideas to improve the method
Combine random forest with computation frameworks such as Hadoop and MapReduce to implement a distributed random forest method and improve the processing efficiency of the algorithm.

Preprocessing the data to reduce the imbalance of the data set, thereby improving the accuracy and classification performance of the algorithm on the unbalanced data set.

2.2. Improved random forest method
First define the accuracy $A_h$ of the decision tree h in the random forest, as shown in equation (1):

$$A_h = \frac{n_r}{n}$$

(1)

Where $n_r$ represents the number of times that the decision tree h gives correct results, and n represents the volume of all data processed by the decision tree h. The accuracy indicates the proportion of a tree with the correct results within a specific time.

In the regression problem, if the classification result given by the decision tree h is consistent with the final result, the decision tree is considered to have obtained the correct result. The difference between the result $x_i$ given by the decision tree h and the final result is calculated and taken. The standard deviation is used as the accuracy of h, as shown in equations (2) and (3):

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

(2)

$$A_h = \frac{1 - \sigma}{1}$$

(3)
Accuracy measures the accuracy of a tree's decision over time.

To filter out samples from the data set that is more useful for generating new decision trees, the margin is introduced as follows: The interval refers to the overall decision accuracy of a random forest on a given sample of data \((x, y)\), and is defined as As shown in equation (4):

\[
\text{margin} (x, y) = \frac{1}{m} \sum_{i=1}^{m} I(h_i(x) = y) - \max_{j \neq y} \frac{1}{m} \sum_{i=1}^{m} I(h_i(x) = j)
\]

(4)

Where \(\frac{1}{m} \sum_{i=1}^{m} I(h_i(x) = y)\) represents the average function and \(I(\cdot)\) represents a metric function. If the correct results are obtained from most decision trees in random forests for the sample \((x, y)\), margin \((x, y)\) is greater than zero. If margin \((x, y)\) is less than zero or a certain threshold value, it means that the sample was misidentified by most decision trees, and the algorithm reached the wrong conclusion for this sample.

The random forest method is applied to the data set \(S'\), a new decision tree set \(\{h'(x, \theta_k), k = 1, \cdots\}\) is obtained. The data set \(S'\) represents only a part of the data in the entire data set. A certain proportion of decision trees are filtered in \(S'\), and the original decision tree is added. In collection

The number of decision trees to be screened is determined based on the ratio of \(S'\) to \(S\), as shown in equation (5):

\[
N_{\text{new}} = N_{\text{all}} \times \frac{\text{number}(S')}{\text{number}(S) - \text{number}(S')}
\]

(5)

The flow of the improved random forest method is shown in Figure 1 as follows.

**Figure 1.** Flow chart of the improved random forest method

1. The initial training data set \(S\) are used to generate the initial random forest \(H\);
2. Random Forest \(H\) is used to classify the current dataset \(S_i\) to be processed:
a) use each tree $h_j$ in the random forest to classify each piece of data $x_j$ in $S_i$;

b) Record the classification result of each tree and each piece of data, and calculate the margin value $(x_j, y)$ of the classification result of this piece of data;

c) If margin $(x_j, y)$ is less than the given threshold, then add $x_j$ to the new training data set $S_i$.

3. After $S_i$ classification, calculate the accuracy of each tree and perform pruning;
4. Perform the random forest method on the new training data set $S$ to generate a new random forest $H$;
5. Perform pruning on the new random forest, and $H$ is combined with $H$ after pruning to form a new random forest $H$;
6. Empty the training data set $S$ and start processing the next batch of data.

3. Experiment and analysis

3.1. Test data set
The test data set is meteorological information data from the Internet industry. The amount of data is 60,000, and the time span is 5 years. The data amount of each quarter is 120, for a total of 100 cities.

To verify the effectiveness of the improved random forest algorithm, 10,000 pieces of data from Q1 of the first year were used as the initial training set, and these data were used to establish the initial random forest with the number of trees $n_{tree} = 100$.

3.2. Effect of data pattern change on the algorithm
12,000 data in Q1 of the first year are taken as the initial training set to establish a random forest, and the established random forest is used to process the subsequent streaming customer data. Each quarter is a cycle. Changes, the original random forest will gradually become unsuited to the new data, and the accuracy of customer value ratings will decrease with time, from 93% in Q1 to 81% in the fifth year.

The number of pruning is set to 50%, and the improved random forest method is used to generate a new decision tree, and add it to the original random forest. The performance of the improved random forest method is shown in Figure 2.

![Figure 2. Comparison of the accuracy of the improved random forest method](image-url)
The results show that compared with the original random forest method, the improved random forest method has significantly better adaptability to changes in data patterns, and the entire forest is also slowly updated with changes over time.

The accuracy of the post-method has remained stable above 90%.

In the method for screening new decision trees, screening is performed using the accuracy of the new sample data set S’ (S’ screening method) and screening using the ratio of the margin mean to the margin of each tree (margin screening method). Excellent results are obtained, and the differences between the modes are small. The method using the data set S for screening (S screening method) can also improve the accuracy, but its accuracy gradually decreases to 87%, which is not as good as the other two methods.

Suppose that regardless of the memory limit, use another 2 modes to update the random forest after the end of each quarter:

Mode 1: All data from the beginning to the current time are used to retrain and generate a random forest;

Mode 2: All data from the previous quarter are used to retrain and generate a random forest each time;

Mode 3: The new random forest method is used to train and generate random forests;

The calculation results show that the accuracy of mode 3 is optimal, the accuracy of mode 2 is slightly worse, but it is also relatively stable, and the accuracy of mode 1 is the worst, which decreases faster as the data volume increases.

The memory required by each of the three modes is shown in Figure 3.

**Figure 3.** Comparison of the required memory for random forest methods in different modes

The results show that the memory required by Mode 1 grows linearly with time and data volume. The required space is the largest. The memory required by Mode 2 is related to the total data volume of each cycle, which is much less than that of Mode 1. When the traffic changes little, it will remain stable. Mode 3 requires the least space, which is much smaller than Mode 2, and its fluctuation range is related to the current accuracy of the algorithm only and irrelevant to the total data flow.
4. Conclusions

In this paper, a new random forest method that adapts to the streaming big data environment is proposed based on the original random forest method for applications in the meteorological industry. The new method works when the streaming big data only pass through the classifier once, where the memory and scanning of massive historical data are not required. Hence, the requirements for memory are minimal. It can also make self-adjustment according to changes in the streaming big data and adapt to the new data, to ensure the accuracy of data processing while keeping the data throughput and processing efficiency.

Acknowledgement
1. Science Foundation for Young Scientists of Hainan Meteorological Bureau. "Research on the quality control of radiation data in Hainan Province" (HNQXQN201801).
2. Science Foundation for Young Scientists of Hainan Meteorological Bureau."Research on the automatic processing of auditing monthly reports of ground weather observations" (HNQXQN 201802).

References
[1] Qingchen Li, Guangxi Cao, Wei Xu. Relationship research between meteorological disasters and stock markets based on a multifractal detrending moving average algorithm[J]. International Journal of Modern Physics B, 2018, 32(1): 267-278.
[2] Hongxing Zhang, Yunbin Yuan, Wei Li. GPS PPP-derived precipitable water vapor retrieval based on Tm/Ps from multiple sources of meteorological data sets in China[J]. Journal of Geophysical Research, 2017, 122(8):12-23.
[3] Xiangyong Lu, Kaoru Ota, Mianxiong Dong. Predicting Transportation Carbon Emission with Urban Big Data[J]. IEEE Transactions on Sustainable Computation, 2017, 2(4):333-344.
[4] Peigen Lin, Robert S. Pickart, G.W.K. Moore. Characteristics and dynamics of wind-driven upwelling in the Alaskan Beaufort Sea based on six years of mooring data[J]. Deep Sea Research Part II Topical Studies in Oceanography, 2019, 162:79-92.
[5] A A Saveliev, A M Domnina. Recovery of continuous field of ecological-climatic indices based on the meteorological stations data[J]. IOP Conference Series Earth and Environmental Science, 2018, 107(1):93-106.
[6] Liu Tianlei, Wang Huan, Qian Xueming, Zhao Guoshuai, Hou Tao, & Hou Xingsong. (2017). Finding optimal meteorological observation locations by multi-source urban big data analysis. Proceedings - 2016 7th International Conference on Cloud Computing and Big Data, CCBD 2016.