Improved Random Forest Algorithm Performance For Big Data

Yousif Abdulsattar Saadoon 1, Riam Hossam Abdulamir 2

1 Misan University, College Medicine, Misan, Iraq
yousif@uomisan.edu.iq
2 Ministry of Education, Misan Directorate, Misan, Iraq
riam.saffar@gmail.com

Abstract

In this paper, the effectiveness of using random forest algorithm in the big data is studied. The reason for choosing this algorithm is because of its effective results in many previous studies, so it was chosen. The random forest algorithm was applied to the big data Internet of Things (IoT) dataset, with size 150,000 instances. After applying the algorithm, it gave an Accuracy score of 99.976%, and this indicates the effectiveness of the random forest algorithm in the big data. This result made the researcher search for one of the ways that increases the value of Accuracy, knowing that it is an excellent result. And use a filter to remove frequent values, and this helps reduce data volume and keep related data. After applying the filter with the random forest algorithm, the Accuracy value appeared at 99.998%, which indicates that it improved the random forest algorithm's performance in the big data.

Keywords: Random Forest, Big Data, Precision, Recall, F-measure, True Positive Rate, True Negative Rate.

1. Introduction

Data science has become the focus of our times, and many researchers are trying to research and develop technologies used in this field. The data of all kinds increases day after day, which makes the process of retrieving it difficult. Therefore, researchers are constantly researching and developing the techniques used to retrieve data properly. One of the operations that helps greatly in retrieving data is the classification process, as it is about creating a series or group of data, whether text, digital, images, or other things. As each group shares the same characteristics and characteristics. This process is used to deal with the big data.
resulting from daily movements across the Internet, which are stored in databases continuously. Whereas, the concept of big data is a very large group of data, complicated in the process of storing, retrieving and processing in normal ways. In this research one of the classification algorithms will be used which is Random Forest (RF), in addition to one of the techniques will be used to improve its performance and compare and analyze the results.

2. Literature Review

(Nancy et. al., 2020) her study aimed to establish a system that performs an effective and accurate prediction of heart disease. The researcher used to verify his dataset study from the UCI e-learning repository. It was tested with ANOVA technology using the RF algorithm using the Weka tool. In addition, in order for the researcher to obtain the best results enter the feature selection step in the process. The results showed that the RF algorithm gives consistently better accuracy than other algorithms.

(Gao et. al., 2019) they proposed in their study entitled “An Improved RF Algorithm Predicting Employee Turnover” A novel method based on an improved RF algorithm. The proposed RF algorithm is applied to employee turnover data with high-dimensional unbalanced characteristics.

(Christy & Gopakumar, 2019) Their study included analyzing an industrial company’s problem with the “IDA 2016” challenge and as part of the challenge it was to find an improved solution to the problem. The researchers used the algorithm of RF with 50 trees as a primary workbook and used the RF with 25 trees as a secondary workbook. Training samples were used to obtain a quality estimate for the test samples and 10 Fold Cross-Validation was used. The results indicated that the average false positive rate (FPR) was 240, while the false negative rate (FNR) was 11. The results were compared using precision, f-measure and Matthews’s correlation coefficient (MCC). Moreover, compared to the top seed in the challenge, and gave good results in solving the problem.
(Wu et. al., 2019) this paper introduces a new approach to disaggregate mixed signals, determining the independent load. A classification method is proposed using a RF multi-label as an algorithm for determining the non-intrusive load. In addition, a multi-label classification can be used to specify the data for the categories to which it belongs. This classification can help determine the operation states of independent loads. That is through mixed signals without detail. The experiments are conducted in a real environment using a generic dataset. Several features are identified to build the classification model. These features are compared to determine the most appropriate features for a classification according to the importance of the feature. As the accuracy reaches 97% and 98%, respectively.

(Octaviani & Rustam, 2019) it has been suggested that a RF algorithm be used to predict breast cancer. One of the techniques that are effective in the classification process is RF. In addition, used with big data. Whereas, the algorithm was applied to data for cancer patients, and the results proved the accuracy of using RF in the process of predicting breast cancer, as the accuracy reached 100%.

(Zhang et. al., 2017) they used the RF algorithm in the machine learning process, for problems that fall within the predication. Whereas, the RF algorithm has been improved and used within the field of data mining, which is regression. This method has been called the Regression-Enhanced RFs (RERFs). It is based on improving RFs by borrowing the strength of penalized parametric regression. In addition, parameters are chosen to adjust the RERFs. After testing the improved algorithm using real data, the results show better predictive performance of the improved RFs algorithm in the Regression process. It ensures a relationship between response and prediction, and provides reliable predictions in data outside of the training data.

3. Data Description
The Internet of Things (IoT) dataset was used, with data collected from nine commercial (IoT) devices by Mirai and Bashlite (2018). The complete Dataset consists of 7062606 instances, 150,000 instances of which were used in this research. The data was divided into 60% for training process and 40% for testing process in order to test the RF algorithm (Meidan et. Al., 2018).
4. Performance Measure

4.1. Overview

In this section, the performance measure, through which the comparison will be made, will be mentioned for the results of the RF algorithm before and after using the Remove Frequent Values Filter. 14 criteria will be found: Correctly Classified Instances (CCI), Incorrectly Classified Instances (ICI), Kappa Statistic (KS), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRAE), Total Number of Instances (TNoI), Precision, Recall, F-measure, True Positive Rate (TPR), True Negative Rate (TNR), Time and Accuracy. Five of them will be explained and detailed in the field of classification and these measures are Precision, Recall, F-measure, True positive rate, True negative rate, Time and Accuracy.

4.2. Precision

Counted as one of the very important performance evaluation measures and the most common measure of performance of text classification in general and Arabic text classification in specific. It is the ratio of the number of correctly detected rows to the overall number of detected rows, and is calculated using below Equation:

\[
\text{Precision} = \frac{\text{Number of correct positive predications}}{\text{Number of positive predications}}
\]

4.3. Recall

The regarded as one of the most popular measures of performance of text classification in general and Arabic text classification in particular. It is ratio of number of the correctly detected rows to the total number of the correct rows, which can be computed using below Equation:

\[
\text{Recall} = \frac{\text{Number of correct positive predications}}{\text{Number of positive examples}}
\]

4.4. F-measure

Counted as one of the most popular measures of performance of text classification in general and Arabic text classification in specific. It is a combination of recall
and precision. In effect, it is the harmonic average of these two performance measures. The greater the value of the F-measure is, the higher are the precision and efficiency of the classifier. It is calculated using below Equation:

$$F\text{-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

### 4.5. True Positive Rate

It is the probability that the actual positive result is (positive), and the equation below represents the method of calculating it:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

### 4.6. True Negative Rate

It is the probability that the actual negative result is (negative), and the equation below represents the method of calculating it:

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

### 5. Methodology

This search includes two phases for completing the results. The first phase is RF results. It starts by taking the dataset and sending it to the RF algorithm. Results are then presented using the following performance measures: (CCI), (ICI), (KS), (MAE), (RMSE), (RAE), (RRAE), (TNoI), Precision, Recall, F-measure, (TPR), (TNR), Time and Accuracy. Figure 1 shows the methodology for using the RF algorithm.
As for the second phase, which is the stage of showing RF results in the big data, but with the use of a filter based on removing frequent values from the dataset. Where this filter was chosen to keep the data one time without duplication, in order to reduce the steps of passing over the data. Figure 2 shows RF methodology using the filter.

6. Results and Discussion

Table 1 shows the RF result. Which (CCI), (ICI), (KS), (MAE), (RMSE), (RAE), (RRAE), (TNoI), Precision, Recall, F-measure, (TPR), (TNR), Time and Accuracy.

Table 1: RF Results.
| No. | Measure | Value      |
|-----|---------|------------|
| 1   | CCI     | 149963.9837|
| 2   | ICI     | 36.0163    |
| 3   | KS      | 0.9997     |
| 4   | MAE     | 0.0002     |
| 5   | RMSE    | 0.0067     |
| 6   | RAE     | 0.1191 %   |
| 7   | RRSE    | 2.3312 %   |
| 8   | TNoI    | 150000     |
| 9   | Precision | 1.000   |
| 10  | Recall  | 1.000      |
| 11  | F-measure | 1.000    |
| 12  | TPR     | 1.000      |
| 13  | TNR     | 1.000      |
| 14  | Accuracy | 99.976 %  |

Figure 3 below shows the results of RF were used for Precision, Recall, F-measure, TPR, TNR and Accuracy.
Table 2 shows the RF result with remove frequent values filter. Which includes (CCI), (ICI), (KS), (MAE), (RMSE), (RAE), (RRAE), (TNoI), Precision, Recall, F-measure, (TPR), (TNR), Time and Accuracy.

Table 2: RF Results with Filter.

| No. | Measure | Value       |
|-----|---------|-------------|
| 1   | CCI     | 49249       |
| 2   | ICI     | 1           |
| 3   | KS      | 1           |
| 4   | MAE     | 0           |
| 5   | RMSE    | 0.002       |
| 6   | RAE     | 0.0085 %    |
| 7   | RRSE    | 0.9238 %    |
|   |   |   |
|---|---|---|
| 8 | TNoI | 49250 |
| 9 | Precision | 1.000 |
| 10 | Recall | 1.000 |
| 11 | F-measure | 1.000 |
| 12 | TPR | 1.000 |
| 13 | TNR | 1.000 |
| 14 | Accuracy | 99.998 % |

Figure 4 below shows the results of RF were used for Precision, Recall, F-measure, TPR, TNR and Accuracy.
Table 3 shows the RF algorithm and RF algorithm with remove frequent values results, which includes (CCI), (ICI), (KS), (MAE), (RMSE), (RAE), (RRAE), (TNoI), Precision, Recall, F-measure, (TPR), (TNR), Time and Accuracy.

Table 3: All Results.

| No. | Measure  | RF      | RF (Filter) |
|-----|----------|---------|-------------|
| 1   | CCI      | 149963.9837 | 49249 |
| 2   | ICI      | 36.0163   | 1           |
| 3   | KS       | 0.9997    | 1           |
| 4   | MAE      | 0.0002    | 0           |
| 5   | RMSE     | 0.0067    | 0.002       |
| 6   | RAE      | 0.1191 %  | 0.0085 %    |
| 7   | RRSE     | 2.3312 %  | 0.9238 %    |
| 8   | TNoI     | 150000    | 49250       |
| 9   | Precision| 1.000     | 1.000       |
| 10  | Recall   | 1.000     | 1.000       |
| 11  | F-measure| 1.000     | 1.000       |
| 12  | TPR      | 1.000     | 1.000       |
| 13  | TNR      | 1.000     | 1.000       |
| 14  | Accuracy | 99.976 %  | 99.998 %    |

Figure 5 below shows the results of RF algorithm and RF algorithm with remove frequent values used for Precision, Recall, F-measure, TPR, TNR and Accuracy.
After showing the results and comparing them based on several criteria which are (CCI), (ICI), (KS), (MAE), (RMSE), (RAE), (RRAE), (TNoI), Precision, Recall, F-measure, (TPR), (TNR), Time and Accuracy. When using the Train data, it came as follow:

1. In terms of the CCI, RF algorithm were the largest number of correct instances, with 149963.9837 instances.
2. In terms of the ICI, RF algorithm are the largest number of incorrect instances with 36.0163 instances.
3. In terms of KS, the RF algorithm with remove frequent values filter were the highest, which was 1.
4. In terms of the MAE, RF algorithm was the highest, equal to 0.0002.
5. In terms of the RMSE, RF algorithm had the highest ratio, equal to 0.0067.
6. In terms of RAE, RF algorithm had the highest percentage, equal to 0.1191%.
7. As for the RRSE, the highest value for RF algorithm was 2.3312%.
8. In terms of TNoI, RF algorithm had the highest value, equal to 150000 instances.

![Figure 5: All Results](image-url)
9. The RF algorithm and RF with remove frequent values filter had the Precision value at 1.000.
10. As for Recall, the RF algorithm and RF with remove frequent values filter had a value of 1.000.
11. The RF algorithm and RF with remove frequent values filter had the F-measure value at 1.000.
12. As for the TPR, RF algorithm and RF with remove frequent values filter had a value of 1.000.
13. The TNR was value for the RF algorithm and RF with remove frequent values filter which 1.000.
14. In terms of Accuracy, the RF with remove frequent values filter was the highest at 99.998%.

7. Conclusions and Recommendations

7.1. Conclusions

The main Aim of this research is to improve RF algorithm performance when using big data. In order to complete the full results, the first two phases of the RF algorithm were tested, using 60% of the data for the training process and 40% for the testing process, where the number of instances was equal to 150000 instances. After testing, the algorithm results showed that its results are effective, and the Accuracy value is equal to 99.976%.

The second stage was by using the RF algorithm with remove frequent values filter and using the big data consisting of 150,000 instances, the data was divided into two parts: 60% for the training process and 40% for the testing process. After showing the results, it was found that the use of remove frequent values filter with RF algorithm improves the results as the Accuracy value increased to 99.998%, or nearly 100%.

7.2. Recommendations

After showing the results, the researcher formulated a number of recommendations, and they came as follows:

1. Using RF algorithm with big data.
2. Use of techniques that help increase the efficiency of the algorithm, such as filters.
3. Remove frequent values filter is effective on big data.

References

1. Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, D. Breitenbacher, A. Shabtai, and Y. Elovici 'N-BaIoT: Network-based Detection of IoT Botnet Attacks Using Deep Autoencoders', IEEE Pervasive Computing, Special Issue - Securing the IoT (July/Sep 2018).
2. Zhang, H.; Nettleton, D; Zhu, Z. (2017). Regression-Enhanced Random Forests.
3. Octaviani, T.; Rustam, Z. (2019). Random Forest For Breast Cancer Predication. AIP Conference Proceedings.
4. Wu, X.; Gao, Y.; Jiao, D. (2019). Multi-Label Classification Based on Random Forest Algorithm For Non-Intrusive Load Monitoring System. MDPI Journal.
5. Christy, J. Gopakumar, G. (2019). An Improved Random Forest Algorithm for classification in an imbalanced dataset. URSI AP-RASC 2019, New Delhi, India.
6. Gao, X.; Wen, J.; Cheng, Z. (2019). An Improved Random Forest Algorithm for Predicting Employee Turnover. Mathematical Problems in Engineering.
7. Nancy, P., Swaminathan, B., Navina, K., Nandgine, B., Lokesh, P. (2020). Tuned Random Forest Algorithm for Improved Prediction of Cardiovascular Disease. International Journal of Recent Technology and Engineering (IJRTE).