Annales Univ. Sci. Budapest., Sect. Comp. 55 (2023) 39–48

EFFICIENCY IMPROVEMENT OF ADAPTIVE RANDOM FOREST USING PRINCIPAL COMPONENT ANALYSIS FOR MINING DATA STREAM

Hayder K. Fatlawi and Attila Kiss

(Budapest, Hungary)

Communicated by András Benczúr

(Received December 5, 2022; accepted January 25, 2023)

Abstract. The rapid increase in the volume of generated data from various digital resources motivated a new trend of data mining techniques that can be trained continuously in parallel with data stream generation. This kind of technique needs to adapt to new data samples and forget the old ones according to some methods such as Adaptive Sliding Window ADWIN. Using ADWIN and Hoeffiding tree classifiers, the Random Forest algorithm was developed to handle the data stream. While the reduction of data samples that were processed in each time moment produced a reduction in the required resources (time and space), the high dimensionality of the data features is still considered a challenge. In this work, ARF-PCA, a stream data classification model, is proposed to improve the efficiency of the Adaptive Random Forest ARF classifier using Principal component analysis PCA. The evaluation of the proposed model based on three real datasets showed a significant improvement in efficiency while preserving the accuracy of the classification.

1. Introduction

Data mining is concerned with extracting useful and nontrivial knowledge from data using many preprocessing and machine learning techniques [15]. The daily rapid increase of generated data made a serious challenge for the typical data mining techniques, which lead to the development of new techniques for

Key words and phrases: Data Mining, Adaptive Random Forest, Principle Component Analysis.

The project has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.3-VEKOP-16-2017-00002.

big data analysis. Another direction was to apply the mining process in parallel with data generation, so the mining techniques depend on stream samples instead of batch files.

The mining of the data stream focuses on adapting the trained classifier in a continuous process while new stream elements have arrived. This process has many constraints in comparison with batch data mining, such as; limitation of available memory, fast response for each new instance, and forgetting the old instances to adapt to the new ones [2]. While typical random forest as an ensemble technique uses random sampling to train many classifiers, Adaptive Random Forest ARF uses adaptive sliding window ADWIN and Heoffman bound to handle the streaming process [5].

Although the reduction of the required time and space due to mining a limited number of instances in a specific time, the high dimensionality still represents a challenge for stream mining techniques. Principal Component Analysis PCA is one of the most popular methods for reducing the number of attributes based on the Eigenvector and Eigenvalue concept [10]. This work aims to improve the efficiency of ARF using PCA, this improvement ensures that the time of the PCA step shouldn't delay the response of the classifier, and it should preserve the accuracy of the classification process.

1.1. Principal Component Analysis

It is a statistical technique that is used to reduce the number of features; hence the data will be converted from high-dimensional to low-dimensional space. The result of this process is a linear approximation of the data. A data stream S in $\mathbb{R}^d$ contains a sequence of samples x1,x2, ..., xN and each x is a vector of d random variables, S can be represented using a linear model with the rank b (where $b \leq d$) [1].

1.2. Adaptive Random Forest

Ensemble modeling aims to build a strong accumulative classifier from many weak classifiers. Adaptive Random Forest is a variation from the typical random forest algorithm for data stream mining tasks. The main idea is to utilize Hoeffding trees, which have the ability to adapt to distribution changes, as the base classifier for the bagging ensemble method [2, 4]. For detecting the change in a data stream, ADWIN is used in this technique. It depends on Online Bagging as a resampling method and a drift monitor for change detection per each tree [2, 4].

1.3. Problem statement

In general, stream mining techniques focus on limiting the number of data instances involved in the learning process, and it isn't concerned with the num-

ber of features. The high number of features leads to a high dimensionality problem resulting in the following problems:

1. Increasing of ensemble model for including all possible subsets of features.
2. Increasing of the number of involving features in the base classifier building, hence, making this process more complex (in computational resource measure).
3. As a consequence for the previous issues, the adapting of the classifiers in case of concept drift will be more complex.

In this work, we propose an efficient classification model for data stream having the ability to solve the three problems by reducing the number of featured using PCA with regard to consideration of the possibility of concept drift.

1.4. Related works

An extension for Incremental Kernel principal component analysis IKPCA was proposed by [7] to minimize the computational resources by dividing the large data into many small chunks. It utilized the accumulation ratio to choose the most useful data without saving all past data. The result of the extended method showed a high reduction in learning time with preserving the accuracy of recognition. A new approach was presented by [8] that used PCA, information entropy, and support vector regression for anomaly prediction by identifying the outlier data. They utilized Wet Flue Gas Desulfurization dataset to evaluate their approach, and their results showed good efficiency and accuracy. A concept drift detection method was proposed by [10] utilizing PCA and angle optimized global embedding (AOGE) for sensors networks stream learning. The variance and angle of the projection in the subspace were analyzed by the proposed method. The change of subspace for each patch is used to detect the concept drift. [6] presented an improvement for multilinear principal component analysis MPCA to handle the stream data. The total tensor scatter maximizing problem converged by the proposed online MPCA. The Experiment results showed lower dimension reduction but with a little worse recognition accuracy.
Recursive principal component analysis (R-PCA) was utilized by [17] for outlier detection using aggregation of the redundant data. An abnormal squared prediction error (SPE) score is used to determine the potential outliers after the extraction of PCA. The results showed an improvement in recovery accuracy and outlier detection accuracy. Logistic regression was combined with PCA in [12] to minimize the features of a social network dataset. The dataset contained 17 million users' tweets and 159 features.
Improvement on Multivariate Convolution Long Short Term Memory was proposed by [16] for data-driven. The method included utilizing Probabilistic

PCA, probabilistic clustering, and neural networks. The evaluation of their method was applied to 22 million telemetry samples from KOMPSAT-2, and it got 35.8 better performance according to precision measurements in comparison with the best baseline approach. [9] proposed an algorithm for data sampling that had the ability to handle the concept drift of the data stream using probability sampling. In their method, the sampling of data was mentioned as long as the data distribution wasn't changed. Their method had better performance compared with random sampling and Gaussian sampling.

2. Methodology

The main aim of this work is to design and implement a classification model for stream data based on adaptive random forest and principle components analysis. Thereby, there are two main stages; the first one is to apply some of the preprocessing steps to prepare the data for the mining process. The second stage is to build an ensemble classifier which includes many very fast decision trees. Figure 1 illustrates all the steps of our work.

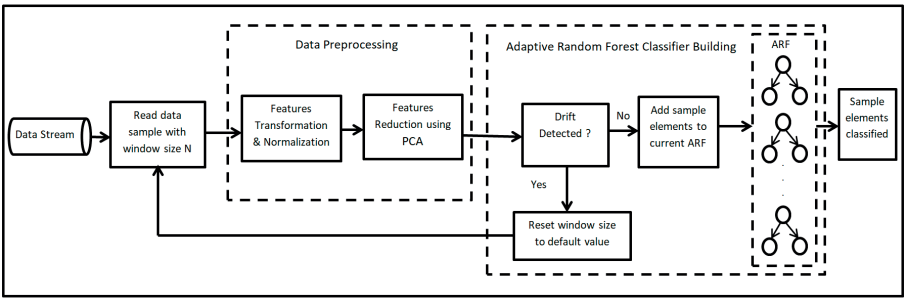


Figure 1. The proposed Classification Model including ARF-PCA

2.1. Stage one: Data preprocessing

Many steps are applied in this stage to provide more useful data for the next stage. In the beginning, all textual values are converted to numerical values; then all values are normalized in the range (-1,1) to avoid any dominance of a specific feature. The previous two steps make changes in the content of data, and the number of instances and features stay the same. The most important step in this stage is to reduce the number of features using PCA. Regardless the original number of features, after this step, only the highest three principle components are used in the next stage.

2.2. Stage two: Building ensemble model

This stage utilized online Bagging of K base classifier [13] as an ensemble model for classification of the reduced data stream. The resampling of ARF

includes choosing different subsets from both features and instances. Poisson(1) distribution is used to choose data samples for each classifier with initial weight w=1, if any data instance is misclassified, its weight is increased before passing it to the next classifier. The final decision is made based on the voting of all K base classifiers with equal weight for all of them. Figure 1 illustrates all the steps of the proposed model.

3. Implementation and experimental results

The implementation of the proposed method included utilizing many frameworks such us Massive Online Analysis MOA, Waikato Environment for Knowledge Analysis (Weka), and Anaconda framework. Also, three medical datasets are used to verify the performance which were EEG Eye state, Hypothyroid, and Breast Cancer datasets. According to validate the proposed model, three real medical datasets are used. The first one was EEG Eye State dataset [11] that contains 15 features and 14980 data instances. The second dataset was Hypothyroid [11] had 26 features and 3163 instances. The third was Breast Cancer [3] dataset and it had 11 features, 280659 instances. Table 1 describes the properties of three datasets where the three datasets had binary class, i.e. it has two values only; also, in the first dataset, the values of the class are approximately balanced, while they were imbalanced in the other two datasets.

Dataset Name	Rows No.	Features No.	Class 1 ratio	Class 2 ratio
EEG Eye State	14980	15	55%	45%
Hypothyroid	3163	26	5%	95%
Breast Cancer	40000	12	1%	99%

Table 1. Datasets Description

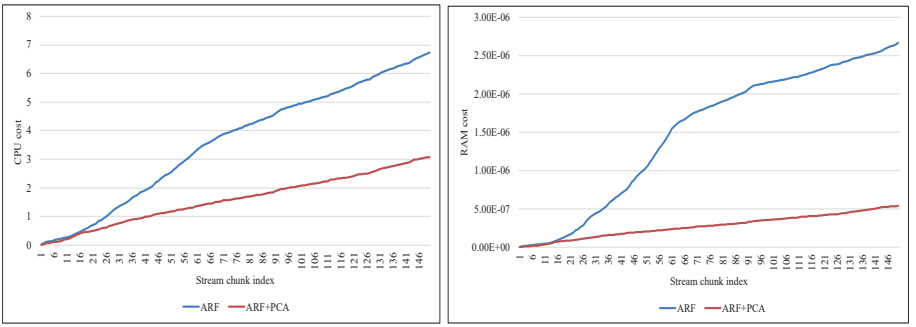
3.1. Data Analysis Platform

In this work, three major tools were utilized to perform the comparison. Weka Platform is an open-source software for data analysis tasks, it was utilized in this work for preprocessing operations (Transformation and Normalization). MOA Platform is an improvement to the Weka platform for the mining data stream, in our comparison, it performed the data streaming and implementation of classification techniques. Anaconda included multiple tools for Python programming such as the Sklearn library that was used in this work for applying PCA to the data.

3.2. Applying ARF-PCA proposed model

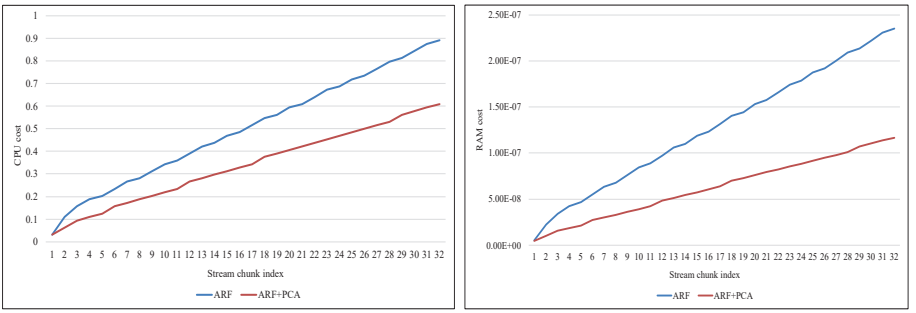
In the implementation of the ARF-PCA model, a data stream was provided by generating sequences from the three datasets batch files. Multiple configurations were used according to the total number of data instances in each dataset. In the first and third datasets, 100 instances per second were used

while 20 instances per second were used in the second dataset.
The reduction of the computational resources i.e. the processor and memory utilization for the three datasets was illustrated in figures 2,3, and 4. The best result was obtained with the EEG dataset with a 55% CPU cost reduction ratio and an 80% RAM cost reduction ratio.



CPU cost reduction 55% RAM cost reduction 80%

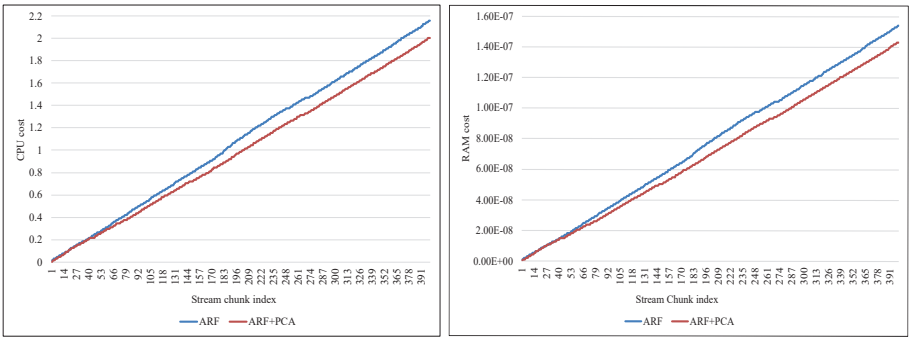
Figure 2. Reduction of CPU and RAM costs applying ARF-PCA on EEG Eye dataset



CPU cost reduction 33% RAM cost reduction 48%

Figure 3. Reduction of CPU and RAM costs applying ARF-PCA F on Hypothyroid dateset

Another observation could be noticed in figure 5 that the utilization of the ARF-PCA model is more stable than the typical ARF. The three highest principle components were chosen for creating the new data stream. The quality of classification was illustrated in figures 6,7, and 8. The classification accuracy was preserved for the three datasets at 98%, 99%, and 99%, respectively.



CPU cost reduction 8% RAM cost reduction 8%

Figure 4. Reduction of CPU and RAM costs applying ARF-PCA on Breast Cancer Dateset

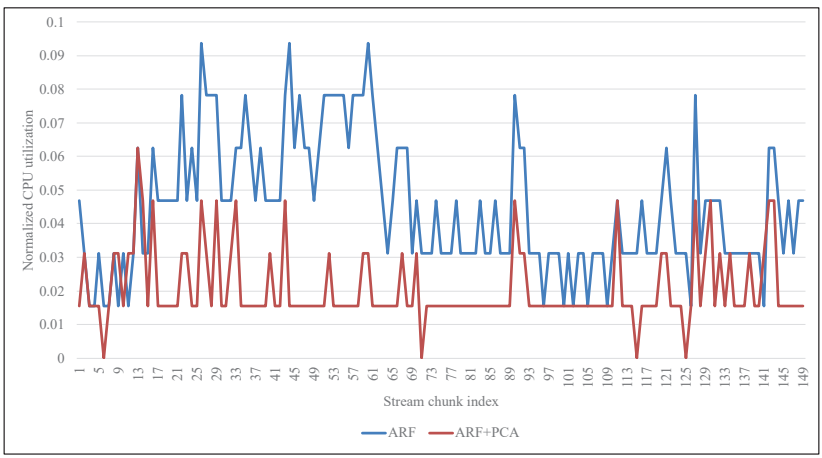


Figure 5. CPU utilization Growth Comparison between ARF and ARF-PCA

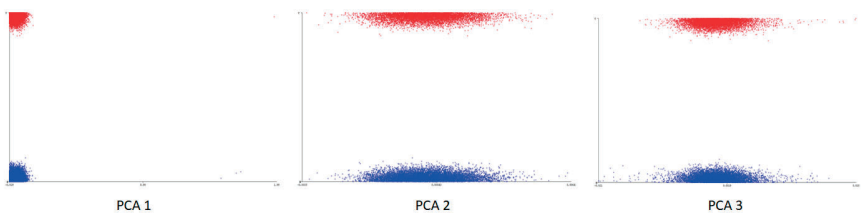


Figure 6. Classes separating virtualization for EEG Eye State Dataset

Figure 7. Classes separating virtualization for Hypothyroid Dataset

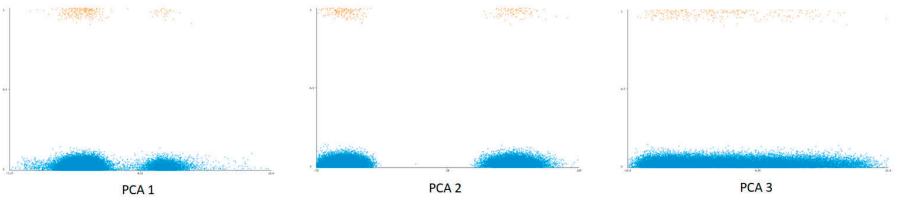


Figure 8. Classes separating virtualization for Breast Cancer Dataset

4. Conclusion

The data stream has many constraints like infinite, fast arrival, and distribution change of data samples. ADWIN aims to detect the change in a data stream by tracking the average of bits in the stream. ARF classification technique utilized ADWIN and Hoeffding trees for adapting to distribution changes. PCA has an interesting ability to improve the efficiency of ARF by reducing the number of features of the data stream. This led to a reasonable reduction of CPU and RAM utilization costs, reaching up to 55% and 80%, respectively, while preserving the accuracy of classification.

References

[1] **Abdi, H. and L.J. Williams,** Principal component analysis, *Wiley interdisciplinary reviews: computational statistics, Wiley Online Library* **2(4)** (2010), 433–459.

[2] **Babenko, B., M. Yang and S. Belongie,** A family of online boosting algorithms, in: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, 2009, 1346–1353.

[3] BCSC Rebosotiry, 2020-03-04. https://www.bcsc-research.org/data/rf/documentation

[4] **Fatlawi, H. K. and A. Kiss,** On robustness of Adaptive Random Forest classifier on biomedical data stream, in: *12th Asian Conference on Intelligent Information and Database Systems*, Phuket, 2020, 332–344.

[5] **Gama, J.,** *Knowledge Discovery from Data Streams*, CRC Press, 2010.

[6] **Han, Le., Z. Wu, K. Zeng and X. Yang,** Online multilinear principal component analysis, *Neurocomputing*, Elsevier, **275** (2018), 888–896.

[7] **Hong, D., D. Zhao and Y. Zhang,** The entropy and PCA based anomaly prediction in data streams, *Procedia Computer Science*, Elsevier, **96** (2016), 139–146.

[8] **Joseph, A. A., T. Tokumoto and S. Ozawa,** Online feature extraction based on accelerated kernel principal component analysis for data stream, *Evolving Systems*, Springer, **7(1)** (2016), 15–27.

[9] **Lin, L., X. Qi, Z. Zhuand and Y. Gao,** Concept drift based multi-dimensional data streams sampling method, in: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Macau, 2019, 331–342.

[10] **Liu S., L. Feng, J. Wu, G. Hou and G. Han,** Concept drift detection for data stream learning based on angle optimized global embedding and principal component analysis in sensor networks, *Computers & Electrical Engineering*, Elsevier, **58** (2017), 327–336.

[11] Machine Learning Repository, 2020-03-04. https://archive.ics.uci.edu/ml/index.php

[12] **Murugan, N. S. and G. U. Devi,** Feature extraction using LR-PCA hybridization on twitter data and classification accuracy using machine learning algorithms, *Cluster Computing*, Springer **22 (6)** (2019), 13965–13974.

[13] **Oza, N.C.,** Online bagging and boosting, in: *2005 IEEE International Conference on Systems, Man and Cybernetics*, Waikoloa, 2005, 2340–2345.

[14] Physionet Rebosotiry, 2020-03-09. https://physionet.org/content/mitdb/1.0.0/

[15] **Tan, P., M. Steinbach and V. Kumar,** *Introduction to Data Mining*, Pearson Education India, 2016.

[16] **Tariq, S., S. Lee, Y. Shin, M.S. Lee, O. Jung, D. Chung and S.S. Simon,** Detecting anomalies in space using multivariate convolutional LSTM with mixtures of probabilistic PCA, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, 2019, 2123–2133.

[17] **Yu, T., X. Wang and A. Shami,** Recursive principal component analysis-based data outlier detection and sensor data aggregation in IoT systems, *IEEE Internet of Things Journal*, **4(6)** (2017), 2207–2216.

H. K. Fatlawi and A. Kiss
Eötvös Loránd University
Faculty of Informatics
Department of Information Systems
Budapest
Hungary
hayder@inf.elte.hu
kiss@inf.elte.hu