International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-4, February 2019

COMPACT: Classifying Stream Data Optimally Using a Modified Pruning and Controlled Tie-threshold

Gayathiri Kathiresan, Krishna Mohanta, Khanaa VelumailuAsari

Abstract*: Big data mining become important in extracting the potential information from the continuously arriving stream data. By extracting knowledge, the data mining algorithms significantly compute feasible decisions for various applications. The Very Fast Decision Tree (VFDT) classifier is a widely applied incremental decision tree to make better decisions. The VFDT classifier processes the arrival of the new instances, without storing them and updates the existing tree structure. Most of the conventional incremental decision tree based algorithms exploit the hoeffding's bound based on the user-defined tie-threshold to split the tree and to manage the tree growth. Even though the size of the tree tremendously increases when handling the fluctuated and imbalanced stream data, it suffers from the misclassification issue due to lack of capturing the optimal attributes over the incoming stream data and declines the classification accuracy and performance. In order to resolve these issues, this paper extends the VFDT, named as Classifying stream data optimally using a Modified Pruning technique And Controlled Tie-threshold (COMPACT). The COMPACT method includes two components, such as enhanced information gain measurement and tie-breaking threshold based pruning method. In order to improve the VFDT performance without affecting the imbalanced data stream handling, the enhanced information gain measurement effectively identifies an optimal number of attributes for a data stream. In order to avoid the information gain biasing, it utilizes the advantages of enhanced splitting metric in attribute reduction. Instead of randomly selecting the threshold, the tie-breaking threshold based pruning method determines the tie-breaking threshold using a number of breaking points. The tie-breaking threshold based pruning method ensures the optimal tree structure while handling the large-scale stream dataset. Finally, the COMPACT method is evaluated using the weather dataset to demonstrate the efficiency. The proposed method significantly outperforms the existing DTFA approach in terms of recall, Root Mean Square Error (RMSE) rate, and execution time.

***Index Terms*: Big data, stream data, VFDT classifier, bias, information gain, threshold, pruning, imbalanced data, optimal attributes, and decision making.**

Manuscript published on 28 February 2019.
***Correspondence Author(s)
Gayathiri Kathiresan, Research Scholar, Bharath Institute of Higher Education and Research, Chennai, India
Krishna Mohanta, Associate Professor, Kakatiya Institute of Technology and science for woman, Nizamabad, Thlangana, India
Khanaa VelumailuAsari, Dean-Info, Bharath Institute of Higher Education and Research, Chennai, India

I. INTRODUCTION

The term 'Big data' is originated from the massive amount of data, which are arriving from diverse sources with high velocity. A vast amount of streaming data comprises diversified data and valuable knowledge sources [1]. The stream data mining extracts the potential information and discovers the patterns from the data [2]. However, analyzing the continuously arriving large-scale stream data is the most challenging task, which induces the time and memory constraints in the system. Thus, the researchers shifting their attention to the streaming data analysis with the help of classification techniques [3]. Although, effectively classifying the rapid arrival of stream data while enhancing the stream detection performance is a complex task. In order to obtain a better decision while classifying the unseen records of stream data with greater accuracy, several researchers employ the decision tree classification model [4,5]. The decision tree classification model predicts the classes of the unstructured instances with the minimal error rate and lacks to construct the tree structure in considerable time and optimal tree size. Moreover, it often faces the difficulties, while reprocessing the tree structure for the arrival of the new instance. To effectively support the processing of new instances in the data stream within a reasonable time, the existing researchers have employed the VFDT as the potential classification model with the optimal pruning method [6]. In a large-scale environment, the processing of raw data stream without the data preprocessing makes the VFDT classification model as an inefficient method. To ensure the quality and the reliability of the data mining system, the VFDT classification model has implemented along with the data preprocessing phases which significantly improve the decision making. Notably, to handle the data mining issues over the evolving stream data without inducing high complexity, the data stream classification exploits the sampling method for the preprocessing of data [7]. The traditional sampling methods lack in performing when the input data stream has fluctuations in rates. In addition, it fails to handle the data stream from the heterogeneous sources.

Furthermore, the splitting criterion based on the random selection of attributes attempts to reduce the computational complexity of the classifier model.

It is mainly, the selected attributes based on the random selection fails to consider the relevant attributes and frequently occurred attributes which lead to inaccurate results and misclassification issue [8]. It is a necessity to split the node based on tie-threshold based pruning method, which plays a crucial role while selecting the tree size. Thus, in order to mitigate these issues, the COMPACT method exploits the adaptive reservoir sampling technique, which assists to drop out the unnecessary memory usage and also handle the imbalanced stream data. Moreover, it makes use of information gain as the best attribute selection measure by the splitting metric which enhances the performance of the system for accuracy and the computational time.

1.1 Contribution:

The main contributions of the proposed Classifying stream data Optimally using a Modified Pruning technique And Controlled Tie-threshold (COMPACT) approach are as follows.

- The proposed COMPACT approach utilizes the VFDT classifier model to ensure the best decision making over stream data while classifying the rapid evolving stream data.
- The COMPACT approach estimates the information gain for all attributes to select the most valuable and optimal attributes over the temporal data stream.
- The COMPACT method eliminates the bias and misclassification issue by confining the information gain deviation based on the enhanced splitting metric in the information gain measurement.
- The proposed method constructs the tree with an optimal size to eliminate the tree explosion issue using the tie-breaking threshold based pruning method that helps to balance the classification accuracy and the efficiency of computation time.
- The experimental framework results illustrate that the performance of the proposed COMPACT approach superior to the existing DTFA approach.

II. RELATED WORK

The development of technology intends to produce the stream of data at a high rate, which makes the complexity while processing and storing of data. In order to mitigate this issue, the conventional data stream management system presented several techniques. However, it lacks to process the numerical attributes which create an impact on efficiency also induces the issues in the construction of decision trees. Thus, research work in [9] processes the numerical attributes based on the numerical interval pruning (NIP) method which assists to reduce the time complexity. Moreover, it employs the entropy along with the Gini index to minimize the sample size, which ensures the accuracy and efficiency of the system. However, it lacks focus on the uncertain data in the stream of data. The work [10] ensures the best attribute selection with the probability specified by the user, which employs Taylor's theorem and the characteristics of the normal distribution to linearize the split-measure function while constructing the decision tree but it is applicable only for the two class issues. To resolve the issue of estimating the confidence interval and the bias error while splitting the nodes of the decision tree, the approach [11] employs the impurity measures such as entropy, Gini index, and the function proposed by Kearns and Mansour to acquire the highest gain. A prototype of spatial entropy-based decision tree [12] exerts the spatial entropy on the ID3 decision tree algorithm based on the spatial autocorrelation phenomenon to classify the data set of air pollution index (API). Moreover, it considers the supportive attribute while splitting nodes in the decision tree which helps to diminish the diversity within the class and to escalate the discrimination of classes that improves the classification process.

Most of the research work in data stream mining often faces challenges while handling the dynamic stream of data. In order to handle the streaming data and manage larger data set within a reasonable period, the decision tree building algorithm developed on a distributed environment, named as Streaming Parallel Decision Tree (SPDT) [13]. It constructs the histogram based decision tree in a breadth-first manner, which facilitates the parallel processing of dataset on multiple processors. In data stream mining, the several decision tree algorithms fail to split the node when the dataset which are having the missing attributes and the uncertainty. In this case, most of the researchers exploit the Very Fast Decision Tree (VFDT) algorithm among the incremental decision tree algorithms. With the target of building the VFDT from the uncertain data, the algorithm [14] employs the hoeffeding bound theory and the Uncertain Naive Bayes classifier (UNB) that assists to enhance the speed of construction and the performance of classification. In order to restrict the unnecessary growth of the decision tree based on the incremental pruning and the adaptive tie-breaking threshold, the optimized VFDT [15] was developed which ensures enhanced predictive accuracy and compact size of the tree. The auxiliary reconciliation control (ARC) method [16] presented to improve the performance of the VFDT by predicting the missing data, replacing noises, and handling the fluctuations of the input data stream which acts as the supportive factor for the VFDT. The framework of random forest algorithm and the VFDT algorithm refereed as Random Forest Based Very Fast Decision Tree algorithm (RFVFDT) introduced to ensure the classification accuracy of data streams. Moreover, it averts the loss of data, process delay while handling the unbounded data stream [17].

Several researchers attempt to provide a better decision with the optimal tree structure by employing the various methods. The algorithm in [18] is developed based on the VFDT for attribute reduction using the impurity measures, including entropy and information gain which helps to reduce the growth of trees and the memory constraints. By applying attribute selection technique to a transaction data stream of a credit card, the research work in [19] constructs the VFDT to improve the accuracy. However, it lacks to retain accuracy when the dataset is larger with the lower illegal data. The Optimized-VFDT (OVFDT) approach in [20] presents the node-splitting mechanism to control the attribute-splitting estimation and to achieve the trade-off between the accuracy of prediction and the size of the tree.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The Moderated-VFDT (M-VFDT) algorithm in [21] employs the adaptive tie threshold technique to control the node splitting in the decision tree based on the incremental computing method. Notably, it exploits the dynamic tie threshold to restrict the selection of attributes in the decision tree. In order to handle the uncertain attributes in the data streams, the model in [22] is developed based on the clustering algorithm which facilitates the classification of leaves to improve the accuracy of the system. The Decision Trees based on the Fractions Approximation algorithm (DTFA) [23] employs the hoeffding bound to acquire the confidence level while splitting the nodes for the decision tree which helps to reduce the size of the tree. However, it fails to consider the bias of the estimate.

Most of the conventional research work based on the VFDT classifier faces significant challenges related to the constraints of unnecessary tree growth, excessive memory usage, and the minimal accuracy when classifying the input stream data based on the user-defined threshold. In order to overcome the shortcomings mentioned above, the proposed model constructs the tree structure based on the deterministic attributes and the enhanced pruning method.

2.1 Problem statement

The traditional decision tree-based classification algorithm constructs the tree by exploiting the entire dataset to make a better decision. It reprocesses and rebuilds the existing tree structure to update the new incoming data instances. The decision tree based classifier is not always suitable to handle the unrecorded and the heterogeneous stream data. In order to overcome these shortcomings, the incremental approach based VFDT algorithm has developed. It dynamically updates the existing decision tree without the reprocessing of earlier data. However, the conventional VFDT algorithm still faces the difficulties while constructing the tree using the unbounded data stream. It fails to retain the classification accuracy under the progressive number of data instances. It necessitates the attribute reduction technique to handle the massive evolving stream data. On the other hand, there is a demand for the stream data mining, which is concerned with the imbalanced stream data. It impairs the accuracy of the VFDT classifier model through the misclassification issue. In addition, it experiences the bias error due to the diversified value of data instances. In the VFDT construction, the attribute selection is an important task. Several existing methods fail to focus on the frequently occurred and also the relevant attributes while randomly selecting the attributes for the classification. It misleads the classification process and leads to the tree explosion issue. Therefore, designing the VFDT based on the deterministic attribute selection and enhanced pruning has to be concentrated more in the future to accomplish the trade-off between the accuracy and the efficiency of the system.

III. OVERVIEW OF PROPOSED METHODOLOGY

The proposed COMPACT method focuses on the better decision by exploiting the optimal attribute set of the stream data and the VFDT classifier model. With the target of attaining the better decision using the optimal attribute set, the proposed approach exploits the adaptive reservoir sampling method. It employs the VFDT classifier to classify the data stream with the tie-threshold based pruning method which mitigates the size of the decision tree model. Moreover, the COMPACT methodology provides the better decision within the reasonable time without compromising the performance using the two significant phases such as data preprocessing phase and the mining phase.

Data preprocessing phase:

The proposed model focuses on the selection of an attribute set in order to ensure the performance of the VFDT classification model. To obtain the optimal number of attribute set, the COMPACT exploits the information gain measurement that measured with the aid of entropy and the splitting measure. In the splitting metric, discriminating the high occurrence attributes with the minimum distinct values and low occurrence attributes with the maximum distinct value is essential. Hence, the proposed COMPACT approach focuses on the difference of attribute set value rather than the occurrence to improve the performance through the splitting metric. In addition, the proposed method attempts to facilitate the process by considering only the retrieved samples rather than the entire data set that allows reducing the computational complexity. To obtain the optimal attribute set, it extracts the attribute from the retrieved sample. By employing the information gain measure and the splitting metric, the COMPACT captures the reduced attribute set among the stream data. Figure 1 shows the decision making using the optimal attribute set based on the adaptive reservoir sampling and the VFDT.

Mining Phase:

The construction of an optimal tree structure for the large-scale dynamic data stream is the most challenging task. Furthermore, it induces the computational complexity while making the decision which intends in lacking the trade-off between the accuracy and efficiency. The existing research works employ the decision tree classifier for mining the data stream. However, there is a requirement of the process the rapid arrival of the data stream without storing and reprocessing the arriving instances. In order to overcome these shortcomings, the COMPACT approach employs the VFDT classifier model for the mining of stream data. The random method based threshold and the redundant measure of information gain are inadequate for the data stream classification. Thus the proposed model decides the tie-threshold based on the breaking point to estimate the necessity of new leaf in the VFDT. Hence, the proposed work classifies the incoming stream data, without corrupting the VFDT classifier accuracy.

IV. EXTENDED VFDT BASED STREAM DATA MINING

Use In order to tackle the constraints of large-scale data, the proposed method enhances the performance of the VFDT classifier using the enhanced splitting measure and the tie-breaking threshold based Pruning method. In addition, the proposed method exploits the adaptive reservoir sampling method to extract the relevant sample over stream data.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

By employing the information gain measure and the enhanced splitting metric, the COMPACT approach facilitates the tree construction with the optimal attribute set, which assists in constructing the optimal tree with the balance between the classification accuracy and the efficiency.

4.1 Stream Data Preprocessing using Adaptive Reservoir Sampling

In order to avert the computational complexity without compromising the efficiency of stream data mining, the COMPACT approach focuses on the preprocessing of data. The proposed COMPACT approach attempts to reduce the size of data and ensure the optimal tree size using the data preprocessing techniques. To make better decisions with the high quality, the data preprocessing involves various techniques, and among them, the proposed approach exploits the sampling method. In the sampling method, the number of tuples is decreased from the large-scale data to acquire the optimal results. The sampling techniques assist in extracting the small data to represent the result of an entire data set.

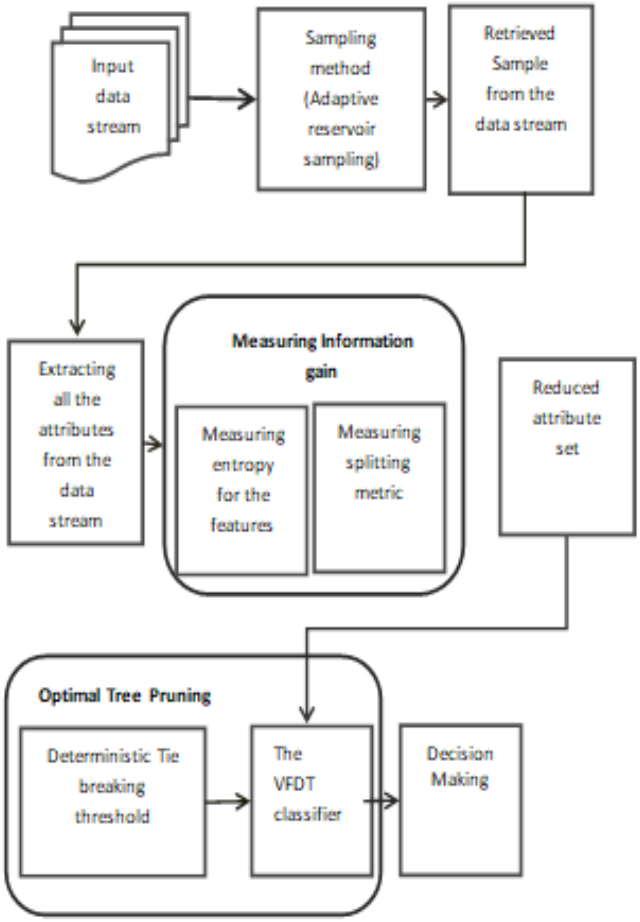

Fig. 1 The extended VFDT with the enhanced splitting measure

The algorithm of adaptive reservoir sampling method:

The conventional reservoir sampling method outperforms when the input stream data arrives from the single source of data instead of heterogeneous sources. It is because of, the reservoir sampling algorithm still maintains the fixed size of data for the entire stream data, which makes the impact on the quality of sampling. Thus, several researchers focus on the adaptive reservoir sampling algorithm to tackle the rapid stream data from the disparate sources [24]. The adaptive reservoir algorithm allocates the optimal reservoir size for the different size of data. Moreover, it estimates the importance score for each data to accomplish the accurate sampling of the data stream. The steps of the sampling algorithm are discussed as follows.

Step 1: Considering the input data stream size as IsI, whereas the substream of the input data stream as $s \epsilon s_1, s_2 ... s_n$, and the reservoir size as r_i whereas i=1 to n.

Step 2: Calculating the size of the reservoirs using the given equation

$$|r_i| = |r| \times \frac{\sigma_i \sum_{j=1}^{|R_i|}(y_{ij})^q \Big/ \sum_{j=1}^{|R_i|}\frac{(y_{ij})^q}{R_i}}{\sum_{k=1}^{n}\sigma_k \sum_{j=1}^{|R_k|}(y_{kj})^q \Big/ \sum_{j=1}^{|R_k|}\frac{(y_{kj})^q}{|R_k|}} \quad (1)$$

In this equation, y_{ij} represents the j^{th} sampling attribute value in R_i, σi indicates the standard deviation in R_i, and 'q' denotes the power of allocation.

Step 3: Checking the size of the reservoir when the arrival of the input data stream to initialize the sampling method. If it becomes complete, then initialize the sub-reservoir.

Step 4: Selecting the samples from the stream data 's' in the reservoir r_i based on the adaptive sampling algorithm.

Step 5: Reallocating the reservoir in optimal size among sub-streams

Step 6: Repeating the above steps when the reservoir size varies based on the input stream data

From the adaptive reservoir sampling method, the proposed COMPACT method extracts the relevant sample over the large-scale stream data. The outcome of the adaptive reservoir sampling method only undergone the further classification process rather than the entire stream data to facilitate the classification.

4.1.1 Attribute Reduction using Information Gain And Splitting Metric

In the stream data classification, the large stream of unbounded data requires efficient analysis and classification to attain better the accuracy. The entire data stream is explored for classification is an arduous task. Moreover, the selection of valuable attributes while constructing the decision tree based classification model is a more complicated process, when the input data stream is having 'n' number of attributes. By exploiting the random method for the attribute set selection in splitting the classesleads to the inefficient results and the lower accuracy. With this intention, the proposed COMPACT approach attempts to reduce the attribute size, which significantly improves the classification accuracy and reduces the errors. Furthermore, it averts the shortcomings of a random selection of attributes by exploiting the information gain measure that estimates the embraced information about each attribute.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

To evaluate the information gain, it exploits the entropy measure and the splitting metric that helps to capture the best attribute among extracted attributes. Correspondingly, the proposed COMPACT method computes the information gain for each attribute (A) by exploiting the equation (2).

$$\text{Information gain}=\text{Entropy}(A)-\left(\sum_{j=1}^{n}\frac{A_j}{|A|}\right)\times\text{Entropy}(A) \quad (2)$$

Whereas the number of attributes 'A' varies from j=1,..,n, and the term Entropy (A) represents the entropy value of each attribute that computed using the equation (3).

$$\text{Entropy}(A)=-\sum_{i=1}^{k}P_i\log_2(P_i) \quad (3)$$

In the equation (3), Pi represents the probability of acquiring the eight data instance, over the j^{th} attribute whereas the instances vary from i=1 to k. To find the attribute that best partition the class with the highest rate of enhanced information gain is essential to ensure the reduced attribute set. Thus, the proposed approach applies the enhanced splitting metric to identify the higher rate split-measure attribute of the decision tree over the attribute set based on the equation (4).

$$\text{Enhanced splitting metric}=(A_j\times\alpha)+\left(\frac{\beta(2\times OC(A_j)\times OC(A_{j-1}))}{OC(A_j)+OC(A_{j-1})}\right) \quad (4)$$

In the equation (4), OC (A_j) denotes that the occurrence of j^{th} attributes over the 'n' attributes, OC(A_j-1) represents that the previous attribute of a j^{th} attribute over the 'n' attributes. α and β indicate that the weighting parameters, whereas α+β=1, 0<α,β≤1, α and β =0.5, which provides equal importance to the value of the instance and its occurrences.

$$\text{Enhanced Gain (EG)}=\frac{\text{Information gain}}{\text{Enhanced splitting metric}} \quad (5)$$

By employing the equation (5), the proposed method significantly identifies the potential attributes based on the enhanced information gain value using the equation (2) and (4). After the computation of enhanced gain for each attribute, the proposed method sorts them based on the attributes which are having the highest enhanced gain (EG_{high}). For instance, the attribute A_1 having the highest information gain (EG_{high}) value compared to attribute A_2, then the attribute A_1 receives more priority than the attribute A_2 is, i.e., an attribute A_1 incorporates the first position in the sorted list of ordered attributes while constructing the decision tree classifier.

4.2 Stream data mining using extended VFDT

In the stream data mining, the control of tree growth is a most challenging task, due to the nature of dynamically arriving stream data. The decision tree construction based on the relevant attributes offers the optimal tree size that creates the necessary tree growth. Hence, the elimination of irrelevant attributes while classifying the data stream improves the classification accuracy. The relevant attribute extraction alone not reduces the computational complexity. In order to keep the size of the decision tree structure and reduce the complexity, the proposed model exploits the tie-breaking threshold method.

4.2.1 VFDT pruning using a tie-breaking threshold

In order to avert the memory and time constraints, the COMPACT approach neglects the attributes, which are having the high splitting rate with the lower enhanced information gain. The proposed methodology captures the relevant attributes based on the tie-breaking threshold value. It decides the attributes by estimating their importance beyond the tie-breaking threshold score, which helps to reduce the irrelevant attributes over the optimal attribute set. In addition, the COMPACT method resolves the bias issues by restraining the deviations of information gain. When the two attributes have identical enhanced gain value, the COMPACT method splits based on the threshold parameter. Moreover, in the classification of stream data, the classifier model often faces the tree explosion issue that makes the negative impact on the performance of the classifier. The larger size of the decision tree constrains the accuracy of prediction and consumes more considerable computation time. Hence, the proposed method exploits the pruning method based on the tie-breaking threshold to limit the growth of the decision tree. It restricts the dispensable growth of the decision tree by adopting the deterministic tie-breaking threshold based pruning method rather than the user predefined threshold.

$$\text{Tie-breaking threshold}(\epsilon)=\left(\frac{A_{1_{|bp|}}-A_{2_{|bp|}}}{A_{1_{|bp|}}+A_{2_{|bp|}}}\right) \quad (6)$$

The tie-breaking threshold is measured using the equation (6). The threshold value relies on the number of breaking point between the two attributes which effectively control the attribute splitting in the decision tree. Subsequently, the proposed method takes advantage of hoeffding bound to split the attributes of the decision tree to compare the tie-breaking threshold against the hoeffding bound. The hoeffding bound identifies the confidence level using the equation (7) to split the attributes in the VFDT.

$$\text{Hoeffding bound}=\sqrt{\left(\frac{R^2\ln\frac{1}{\delta}}{2m}\right)} \quad (7)$$

Let A_{1bp}, A_{2bp} are the number of breaking point belongs to the attribute 1 and 2 respectively. If the tie-breaking threshold (Ɛ) greater than the hoeffding bound with 'm' instances and the distribution of class range R with the probability δ, then the proposed method attempts to split the attribute. This splitting process allows only the optimal attributes to construct the optimal tree structure.

4.3 Stream data classification using extended VFDT

In order to deal with the dynamic stream data with the imbalanced data and the larger tree structure, the proposed COMPACT method exploits the VFDT classifier with the enhanced splitting measure and the deterministic tie-breaking threshold method that illustrated using the figure 2. The extended VFDT classifier model is trained for the arrival of each new stream data. The extended VFDT gets updated when the evolution of data occurs.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Thus, the dynamic update on the proposed approach ensures the accuracy of the classification over the continuously evolving stream data.

The entire procedure for the proposed COMPACT approach explained in algorithm 1.

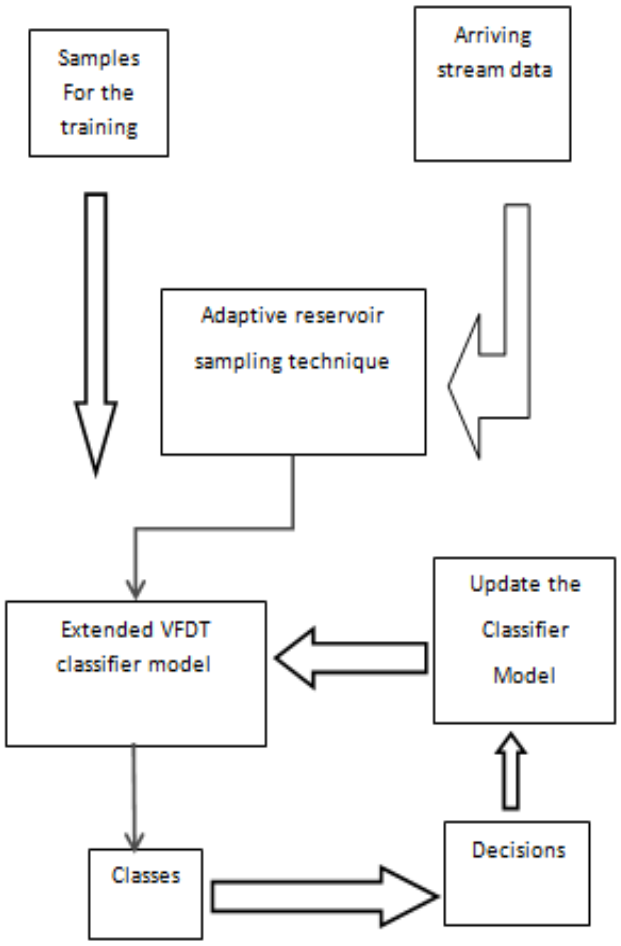

Fig. 2 The classification process of the COMPACT method

V. EXPERIMENTAL EVALUATION

In order to evaluate the performance of the proposed COMPACT approach, the experimental framework compares the proposed approach with the existing DTFA approach using the test data set. The experimental setup evaluates the proposed approach by utilizing the weather dataset. This dataset gathered from the disparate sources of data with different characteristics. It accumulates the 30 number of USA city data from the 18 sources for 45 minutes each. It has common attributes of Timestamp, Temperature (°F), Condition, and location.

5.1 Experimental setup

The proposed COMPACT method runs in Java using the Apache Storm 0.9.4 version as a processing engine. In a storm, the spouts and bolts are mutually compounded to build the topology whereas storm exploits the raw data sources as the input data and transfer them to the bolt. The bolt effectively applies various stream transformation process on the outcome of the spout. Apache Storm does not equip with any Machine Learning (ML) libraries. In order to mine the big streaming data with the machine learning techniques in an efficient manner, the proposed method exploits the Apache SAMOA 0.3.0 version. It is the framework of distributed machine learning techniques that help to boost the development process of the various machine learning algorithms.

Input: N number of attributes (A),
i data instances, i $\in$ k,
Output:VFDT classifier model
While the decision tree have root node only
foreach attribute A **do**
foreach data instance i **do**
Compute the sample s using the adaptive reservoir algorithm
forset of sample S **do**
foreach attribute (A_j) **do**
Compute the information gain using the equation (2)and (3)
Compute the enhanced splitting metric using the equation (4)
Compute the (EG) using the equation (5)
endfor
endfor
endfor
endfor
foreach attribute having the (EG_{High}) **do**
Compute the tie-breaking threshold using the equation (6)
Compute the hoeffding bound using the equation (7)
If(ε>hoeffding bound)then
Add a leaf node
Otherwise
Discard as a outlier
recurse to next node
endif
endfor
endwhile

Algorithm 1: The procedure of COMPACT Method

5.1.1 Evaluation metrics

The experimental evaluation has provided the metrics of accuracy, recall, and Root Mean Square Error (RMSE).

- **Recall:** It is defined as the percentage of the number of correct decisions to the number of decisions that should have been returned.

$$Recall = \frac{Number\ of\ correct\ decisions}{Number\ of\ decisions\ that\ should\ have\ been\ returned}$$

- **Execution time:** It is the time utilized by the COMPACT to classify the continuously evolving stream data using the extended VFDT classifier based on the enhanced splitting metric and the tie-breaking threshold based pruning method.
- **RMSE:** It is a squared root error which calculates the absolute deviation between the predicted and actual value, where N is the total number of instance.

$$RMSE = \frac{\sum_{i=1}^{N}(Predicted\ value - Actual\ value)}{N}$$

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

5.2 Evaluation results

5.2.1 Number of attributes Vs. Recall

Figure3 depicts the percentage of recall for both the proposed COMPACT approach and the existing DTFA approach while varying the number of attributes. The overall recall rate increases with the increasing number of attributes and the COMPACT approach acquire the highest level of recall rate, which shows that it provides better decisions compared to the existing approach. When the number of attributes is equal to 4, the COMPACT method attains the recall rate at 55%, and this value is very close to the DTFA approach recall rate. When increasing the number of attributes to 5, the recall value of COMPACT increases drastically. It is because, the proposed method neglects the irrelevant attributes while classifying the incoming data stream based on the enhanced information gain score, which helps to capture the best attribute from the retrieved samples to achieve greater accuracy. The proposed approach retains the recall rate even when increasing the number of attribute values from 5 to 7, but the existing method fails to maintain the recall rate. The existing approach suddenly decreases the recall rate when the number of attributes increases beyond the certain level due to the lack of attention of optimal attributes among the incoming stream data. Hence, the proposed approach enables the attribute selection based on the deterministic manner rather than random manner, which helps to improve the accuracy.

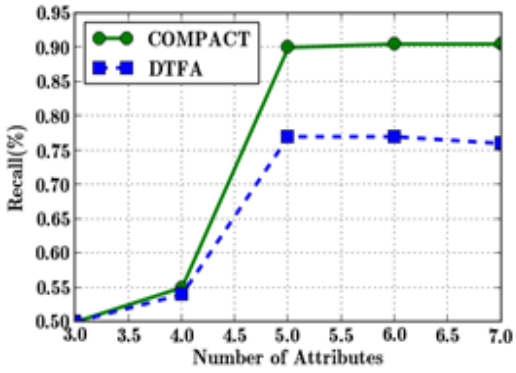

Fig.3 Number of attributes Vs. Recall

Multipliers can be especially confusing. Write "Magnetization (kA/m)" or "Magnetization (10^3 A/m)." Do not write "Magnetization (A/m) × 1000" because the reader would not know whether the top axis label in Fig. 1 meant 16000 A/m or 0.016 A/m. Figure labels should be legible, approximately 8 to 12 point type.

5.2.2 Number of instances Vs. Execution time

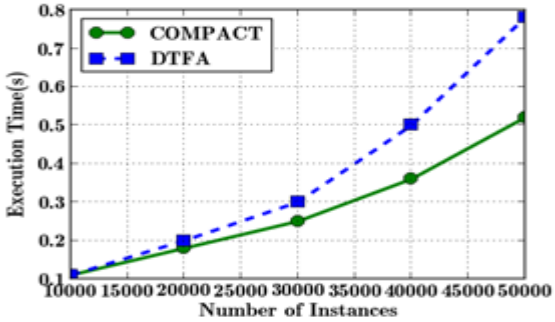

Fig.4 Number of instances Vs. Execution time

5.2.3 Number of instances Vs. RMSE

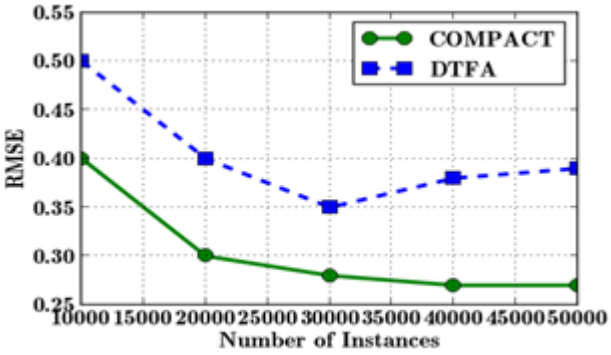

Fig.5 Number of instances Vs. RMSE

The execution time of the COMPACT approach and the DTFA approach is demonstrated in figure 4 when varying the total number of instances over the stream data. The execution time increases with the increase of the number of instances. The proposed approach consumes 0.36 seconds of execution time when dealing with the 40000 numbers of instances. However, the DTFA approach takes 0.5 seconds to classify the streaming data instances. Hence, the COMPACT method applies the tie threshold based pruning method instead of a user-defined threshold, which assists in mitigating the overall execution time through the optimal tree size. The construction of the optimal tree structure takes lesser execution time and the complexity which eases the classification process. When the arrival of new instances, the proposed approach extends the existing optimal tree structure within a reasonable time rather than reprocessing the instances. The proposed approach gradually increases the execution time even the dramatic growth of data instances. In contrast, the existing approach escalates the execution time when the increase of instances due to the absence of pruning method while constructing the tree.

Figure 5 illustrates the RMSE error rate of the COMPACT approach and the DTFA approach with the variation of the number of data instances in the stream data from 10000 to 50000. The minimum RMSE score reflects the remarkable performance of the classifier model. When the data instance size varies from the 30,000 to 50000, the proposed COMPACT approach effectively classifies the instances to attain better decisions within the optimum level, and then it becomes stable beyond 30000 instances in the stream data as the COMPACT method along with the enhanced information gain score enables better decisions than the existing DTFA approach. However, the existing DTFA approach lacks to maintain the stable environment under the progressive number of data instances, which tend to diminish the performance of the existing approach. The COMPACT method potentially captures the samples with the higher enhanced gain value rather than large-scale streaming data to construct the tree structure. It averts the irrelevant attributes in the tree structure that improves the performance of the classifier model while classifying the dynamic stream data. Moreover, the existing approach increases the error rate by 12% than the proposed approach when the number of instances is equal to 50000.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In this regard, the proposed method enhances the system efficiency than the existing DTFA approach.

VI. CONCLUSION

This work presented the COMPACT approach, which is the extended VFDT classifier model for classifying the stream data. With the intention of making better decisions, the proposed COMPACT method has incorporated two components, include enhanced information gain measurement and tie-breaking threshold based pruning method. The proposed method initially applied the enhanced information gain measure to extract the optimal attributes over the large-scale continuously evolving stream data. The proposed method applies the tie-breaking threshold based pruning method for the optimal attribute set for building the optimal and efficient tree structure. It assures the classification model with reduced computational complexity and time constraints without promising the classifier efficiency. Furthermore, it has enhanced the classifier accuracy and decreased the memory requirements of the system. The experimental results have illustrated that the COMPACT approach significantly improves the classifier performance in terms of accuracy and the time. The COMPACT method enables the improvement of 14% higher recall rate than the existing DTFA method when varying the number of attributes. In the future, the proposed scheme has to dynamically select the number of attributes among the stream data to improve the performance of the classifier. Further, the context-aware based feature selection for the classification of diversified stream data helps to make a better decision.

REFERENCES

1. Nasereddin, Hebah HO, "Stream Data Mining," *IJWA*, Vol.3, No.2, pp.90-97, 2011.
2. Almalki, EbtesamHamed, and Manal Abdullah, "A survey on big data stream mining," *Journal of Fundamental and Applied Sciences*, Vol.10, No.4S, pp.278-284, 2018.
3. Aggarwal, Charu C, "A Survey of Stream Classification Algorithms," *In Data. Classification: Algorithms and Applications*, pp.245-274, 2014.
4. Song, Yan-Yan, and L. U. Ying, "Decision tree methods: applications for classification and prediction," *Shanghaiarchives of Psychiatry*, Vol.27, No.2, pp.130, 2015.
5. Nguyen, Hai-Long, Yew-KwongWoon, and Wee-Keong Ng, "A survey on data stream clustering and classification," *Knowledge and information systems*, Vol.45, No.3, pp.535-569, 2015.
6. Subrahmanyam, M. V. V. S., and R. V. Venkateswara, "VFDT Algorithm for Decision Tree Generation," *International Journal for Development of Computer Science and Technology (IJDCST)*, Vol.1, No. 7, 2013.
7. Ramírez-Gallego, Sergio, BartoszKrawczyk, Salvador García, MichałWoźniak, and Francisco Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, Vol.239, pp.39-57, 2017.
8. Krempl, Georg, IndrеŽliobaite, DariuszBrzeziński, EykeHüllermeier, Mark Last, Vincent Lemaire, TinoNoack, et al., "Open challenges for data stream mining research," *ACM SIGKDD explorations newsletter*, Vol.16, No.1, pp.1-10, 2014.
9. Jin, Ruoming, and GaganAgrawal, "Efficient decision tree construction on streaming data," *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.571-576, 2003.
10. Rutkowski, Leszek, MaciejJaworski, Lena Pietruczuk, and PiotrDuda, "Decision trees for mining data streams based on the Gaussian approximation," *IEEE Transactions on Knowledge and Data Engineering*,Vol.26, No.1, pp. 108-119, 2014.
11. De Rosa, Rocco, and NicoloCesa-Bianchi, "Splitting with confidence in decision trees with application to stream mining," *In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pp.1-8, 2015.
12. Zhao, Minyue, and Xiang Li, "An application of spatial decision tree for classification of air pollution index," *In Proceedings of the 19th IEEE International Conference on Geoinformatics*, pp.1-6, 2011.
13. Ben-Haim, Yael, and Elad Tom-Tov, "A streaming parallel decision tree algorithm," *Journal of Machine Learning Research*,Vol.11, pp.849-872, 2010.
14. Liang, Chunquan, Yang Zhang, Peng Shi, and Zhengguo Hu, "Learning accurate very fast decision trees from uncertain data streams," *International Journal of Systems Science*,Vol.46, No.16, pp.3032-3050, 2015.
15. Yang, Hang, and Simon Fong, "Optimized very fast decision tree with balanced classification accuracy and compact tree size," *In IEEE3rd International Conference on Data Mining and Intelligent Information Technology Applications (ICMiA)*,pp.57-64, 2011.
16. Naidu, Ch SKVR, and T. Y. Ramakrushna, "Augmentation of very fast decision tree algorithm aimed at data mining," *IJRCCT*, Vol.4, No. 9, pp.684-690, 2015.
17. Dong, Z. J., S. M. Luo, Tao Wen, F. Y. Zhang, and L. J. Li, "Random forest-based very fast decision tree algorithm for data stream," *Res. Paper*,Vol.12, pp. 52-57, 2017.
18. Da Costa, Victor GuilhermeTurrisi, André Carlos Ponce de Leon Ferreira, and SylvioBarbon Junior, "Strict Very Fast Decision Tree: a memory conservative algorithm for data stream mining," *Pattern Recognition Letters*, Vol.116, pp.22-28, 2018.
19. Minegishi, Tatsuya, Masayuki Ise, AyahikoNiimi, and Osamu Konishi, "Extension of Decision Tree Algorithm for Stream Data Mining Using Real Data,"2009.
20. Yang, Hang, and Simon Fong, "Incremental optimization mechanism for constructing a decision tree in data stream mining," *Mathematical Problems in Engineering*, 2013.
21. Yang, Hang, and Simon Fong, "Moderated VFDT in-stream mining using adaptive tie threshold and incremental pruning," *In Springer International Conference on Data Warehousing and Knowledge Discovery*, pp.471-483, 2011.
22. Xu, Wenhua, Zheng Qin, Hao Hu, and Nan Zhao, "Mining uncertain data streams using clustering feature decision trees," *In Springer Int. Conf. on Advanced Data Mining and Applications*, pp.195-208, 2011.
23. Duda, Piotr, MaciejJaworski, Lena Pietruczuk, and LeszekRutkowski, "A novel application of hoeffding's inequality to decision trees construction for data streams," *In IEEE International Joint Conference on Neural Networks (IJCNN)*, pp.3324-3330, 2014.
24. Al-Kateb, Mohammed, and Byung Suk Lee, "Adaptive stratified reservoir sampling over heterogeneous data streams", *Information Systems*, Vol.39, pp.199-216, 2014.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)