TELKOMNIKA, Vol.17, No.6, December 2019, pp.3086~3099
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i6.11711

An adaptive clustering and classification algorithm for Twitter data streaming in Apache Spark

Raed A. Hasan*[1], Royida A. Ibrahem Alhayali[2], Nashwan Dheyaa Zaki[3], Ahmed Hussien Ali[4]
[1]Faculty of Al-Dour Technical institute/ Noerthern Technical University, Mosel, 41002, Iraq
[2]Department of Computer Engineering, College of Engineering, University of Diyala, Diyala, Iraq
[3]University of Information Technology and communications, College of Engineering, Baghdad, Iraq
[4]AL Salam University College Computer Science Department Baghdad, Iraq
*Corresponding author e-mail: raed.isc.sa@gmail.com[1], royida.alhayali@engineering.uodiyala.edu.iq[2], nashwanalani@uoitc.edu.iqcom[3], msc.ahmed.h.ali@gmail.com[4]

Abstract

On-going big data from social networks sites alike Twitter or Facebook has been an entrancing hotspot for investigation by researchers in current decades as a result of various aspects including up-to-date-ness, accessibility and popularity; however anyway there may be a trade off in accuracy. Moreover, clustering of twitter data has caught the attention of researchers. As such, an algorithm which can cluster data within a lesser computational time, especially for data streaming is needed. The presented adaptive clustering and classification algorithm is used for data streaming in Apache spark to overcome the existing problems is processed in two phases. In the first phase, the input pre-processed twitter data is viably clustered utilizing an Improved Fuzzy C-means clustering and the proposed clustering is additionally improved by an Adaptive Particle swarm optimization (PSO) algorithm. Further the clustered data streaming is assessed utilizing spark engine. In the second phase, the input pre-processed Higgs data is classified utilizing the modified support vector machine (MSVM) classifier with grid search optimization. At long last the optimized information is assessed in spark engine and the assessed esteem is utilized to discover an accomplished confusion matrix. The proposed work is utilizing Twitter dataset and Higgs dataset for the data streaming in Apache Spark. The computational examinations exhibit the superiority of presented approach comparing with the existing methods in terms of precision, recall, F-score, convergence, ROC curve and accuracy.

Keywords*: classification, clustering, data streaming, optimization, pre-processing*

1. Introduction

Over recent years, businesses and associations didn't have to store and perform much tasks and analytics on information of the clients [1]. The need to change everything into information is quite engaged to fulfil the necessities of the general population. Along these lines, big data came into image in the real time business examination of processing data. Presently, individuals are communicating their opinions through online blogs, conversation forms and furthermore some online applications like Facebook, Twitter, and so on [2, 3]. In the most recent decade, there has been an enormous development in the utilization of micro blogging stages, like Twitter [4] which is overpowered by astonishing statistics [5].

Widespread information accumulation from news sources and micro blogs has delivered huge literary data information streams that are trying to process and examine. The identification of rising occasions from data streams, for example, Twitter has gotten developing consideration from analysts [6, 7]. Twitter is the enormous online social networking webpage that presumably ended up ordinary surfing sites by a large number of clients. Twitter supports its client to express the sentiments or thinking with respect to certain circumstances of real-world happenings [8]. Twitter investigates the thoughts by utilizing client's posts, blogs, and reviews to support numerous associations which are hook up with Twitter for enhancing the client sentiments and governmental issues, and recommender framework [9-11].

Apache spark is a quick, broadly useful and distributed processing platform that utilizes dispersed memory generalization to process huge volume of information effectively. Apache spark is adaptable and versatile computing framework comprises of effective API and higher

Received November 8, 2018; Revised June 28, 2019; Accepted July 24, 2019

request apparatuses that are good with hadoop [12]. As of late, analysing enormous unstructured information is a business need. Cluster analysis is one of the mining issues utilized for investigation like assessment mining, sentimental investigation and popularity examination [13]. Current systems use devices and advances to process Twitter information which are utilizing event processing and one message at time investigation [14].

A standout amongst the latest studies utilized several learning frameworks [15] such as K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB) [16, 17]. The RFA generates better recall, precision, and F-measure values. SVM all performed similarly by achieving about 93% accuracy in every group. In every one of these prior studies investigations, classification was used for spam discovery on Twitter. The anomaly identification framework improvement is for distinguishing spammers on Twitter utilizing account data and streaming tweets [18, 19].

The main contributions can be stated as: 1) pre-processed utilizing an Improved Fuzzy C-means clustering to viably cluster the atwitter information then the clustering is additionally improved by utilizing an Adaptive Particle swarm optimization (PSO) algorithm, 2) pre-processed information is classified utilizing the modified support vector machine (MSVM) classifier with grid search optimization.This article is presented in different sections as follows: the related previous studies to the proposed system were reviewed in section 2, while section 3 briefly discussed the suggested approach. In section 4, the experimental results were discussed while section 5 presented the conclusion.

2. Research Method

A Hypertext-Induced Topic Search (HITS) was suggested by Leilei et al. [20] based on the Topic-Decision strategy (TD-HITS) and a Latent Dirichlet Allocation (LDA)-based Three-Step display (TS-LDA). The framework was suggested for influential spreaders detection and identification in social media data streams. The proposed TDHITS can easily identify the number of themes as different related posts in a huge number of posts. TS-LDA can identify powerful propagators of trending event based on the client data and the post. On a Twitter dataset, the results showed the efficiency of the suggested methods in events recognition and in distinguishing powerful event propagators.

Shangsong Liang et al. [21] proposed a work for handling the issue of client clustering with regards to their distributed short text streams. To acquire better client clustering performance, they proposed a two-user cooperative interest following models that go for following changes of every client's dynamic point dissemination as a team with their followers' dynamic subject dispersions, based both with respect to the content of current short messages and the recently evaluated conveyances. They also suggested 2 collapsed Gibbs sampling frameworks for the cooperate inducement of the dynamic advantages of the clients for both short- and long-term clustering reliance point models.

Streaming data is one of the considerations accepting hotspots for concept-evolution studies. At the point when another class happens in the information stream it very well may be considered as another idea thus the concept-evolution. Tahseen et al. [22] highlighted the problem by characterizing a new collaborative strategy called "class-based' group which swaps the conventional "chunk-based" method for repetitive class identification. The study discussed the attribute of the 2 different techniques in class-based group in order to provide their detailed analysis and clarification. They also proved the superiority of the "class-based" groups over procedures by means of observational methodology on various benchmark databases comprising web remarks as text mining challenge.

Lekha et al. [23] developed a framework for open-source big data called Apache Spark which is a cloud-based framework that focus on the development of machine learning framework with respect to big data streaming. In this framework, the user tweets his/her health traits and the application get the equivalent progressively, extricates the traits and develop machine learning framework to anticipate client's health status which was then legitimately informed to him/her immediately to make suitable action.

Senthil and Usha [24] worked on categorizing streams of Twitter data based on sentiment analysis using hybridization. The study used a URL-based security device to collect 600 million open tweets while feature selection was applied for sentiment investigation. The ternary classification was performed based on a pre-processing strategy while the results of

the tweets sent by the users are collected. Then, a hybridization approach based on 3 optimization methods (PSO, GA and DT) was applied for classification accuracy using sentiment analysis. The results were compared with previous works, and their developed strategy demonstrates a greater than different classifiers analysis.

3. Proposed Methodology

3.1. Phase 1: Adaptive Clustering for Twitter Data Streams in Apache Spark

The presented technique consists of the subsequent steps: initially, input twitter data is pre-processed using tokenization and stop word removal processes. Then the pre-processed data is effectively clustered utilizing an Improved Fuzzy C-means clustering with Adaptive Particle swarm optimization (PSO) algorithm. Finally twitter data streaming using our proposed method is examined in apache spark engine. The flow diagram of this proposed twitter data streaming utilizing phase 1 methodology is given in Figure 1.

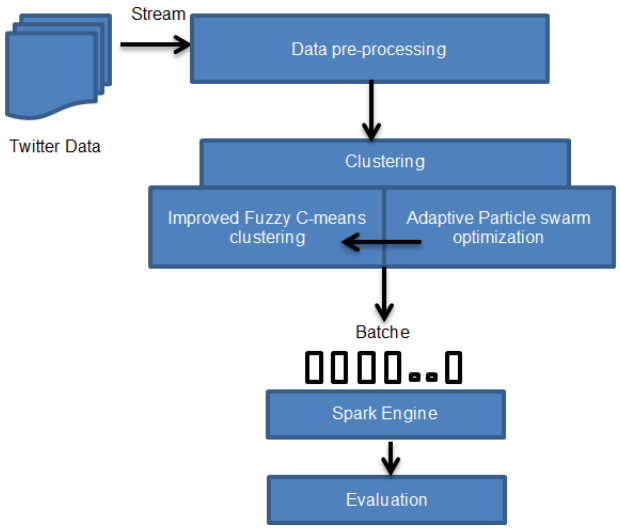

Figure 1. Flow diagram of phase 1 proposed methodology

3.1.1. Preprocessing

In the proposed twitter data streaming, at first the input data used for the proposed proficient information streaming is taken from the dataset [25-30]. Here, is the twitter input dataset. At that point the input twitter data is preprocessed utilizing tokenization and stop word removal processes which are utilized to expel the conflicting information or noisy information from dataset. Input data preprocessing incorporates the accompanying processes [31].

a. Symbolization

Symbolization is the task of splitting the input information up into pieces, called tokens, possibly in the meantime discarding certain characters, like punctuation. Basically, tokenization is the way toward separating the given text into units called tokens and it is utilized for further handling. The tokens might be words, number and punctuation sample [32-35]. The reason for symbolization is to expel all the punctuation marks like commas, full stop, hyphen and brackets. The input data after applying the tokenization is given in (1):

$$\bar{T}_i = \{D_1, D_2, D_3, \dots, D_n\} \quad (1)$$

where, $\bar{T}_i$ is the tokenized data and $i = 1,2,3, \dots, n$.

b. Stop word removal

After tokenizing, the tokenized information ($\bar{T}_i$) is given as the contribution for stop word removing and here some undesired words are rejected by utilizing stop word elimination. Stop words will be words that are by and large thought to be futile. The purpose for this procedure is

utilized to avoid conjunction, relational words, articles and other continuous words, like adverbs, action words and adjectives from textual information [36]. Some of the as often as possible utilized stop words are "a", "me", "of", 'the', 'he', 'she', 'you'. The tokenized information subsequent to applying the stop word elimination is given in (2):

$$S_w = \{T_1, T_2, T_3, \dots, T_n\} \quad (2)$$

here, S_w is the preprocessed set of data after eliminating stop words and $w = 1,2,3, \dots, n$.

3.1.2. Data Aggregation

Aggregation is the process of splitting a set of objects in the dataset into subsets or cluster. Each subset is a cluster, and attributes in a cluster are similar to each another. The proposed modified fuzzy clustering algorithm (MFCM) is used for effective clustering where the performance of the MFCM depends upon the updating the memberships function using sigmoid function. Additionally MFCM performance is improved by using support value based adaptive PSO algorithm. The preprocessed data is optimized using support value based adaptive PSO algorithm before modified fuzzy c-means clustering [37].

Clustering is the process of separating a set of items in the dataset into subsets or cluster. Every subset is a cluster, and traits in a group are like each another. The proposed modified fuzzy c-means clustering algorithm (MFCM) is utilized for viable clustering where the execution of the MFCM relies on the updating the membership functions utilizing sigmoid function. Also MFCM execution is improved by utilizing support value based adaptive PSO [38].

a. Support value based adaptive PSO

The PSO was developed as a heuristic population-based optimization method which was inspired by the flocking behaviour of birds. The PSO is presented as a collection of particles which individually represents a potential solution [39]. The particles pursue a basic behavior: copy the accomplishment of neighbouring particles and its own accomplished triumphs. The location of a particle is thusly affected by the best particle in a neighbourhood, $p'_{best\ j}$ just as the arrangement found g'_{best}. Particle position y_j is balanced utilizing the accompanying condition:

$$y_j(k' + 1) = y_j(k') + v_j(k' + 1) \quad (3)$$

where, the velocity component v_j signifies the step size. The velocity is updated via (4):

$$v_j(k' + 1) = w' v_j(k') + c_1 r_1 \{p'_{best\ j} - x_j(k')\} + c_2 r_2 \{g'_{best} - x_j(k')\} \quad (4)$$

where, w' is the inertia weight, c_1 and c_2 are the acceleration coefficients $r_1, r_2 \in [0,1]$, $p'_{best\ j}$ is the individual best position of particle j, and j'_{best} is the best position of the particles.

At that point, Map the location of each particle into solution space and evaluate its fitness esteem as indicated by the support value based fitness function. In the meantime, update $p'_{best\ j}$ and g'_{best} position if required. The support value is estimated by utilizing (5):

$$\bar{S}_v = \frac{y_1 * y_2 * \dots . y_n}{y_1 + y_2 + \dots . y_n} \quad (5)$$

Here, $\bar{S}_v$ denotes the support value, $y_1, y_2, \dots, y_n$ signifies the input population. This updating process is proceeds until a criterion is met, usually it used for finding optimum solution through number of iterations. The pseudo code of support value based adaptive PSO algorithm is given in Algorithm 1.

Algorithm 1: Support value base adaptive PSO algorithm

Step 1: Initialization
Set the initial size k' = 0
Set a population size of NP
Set velocities size v_j of the insect

Set 2:
While condition not reached
Do
For j = 1 to NP
Step 3: Calculate $p'_{best\ j}$ and g'_{best}
Evaluate the fitness of particles using (5)
Step 4: Update position and velocity
Calculate the positions and velocities of insect utilizing (3) and (4)
End For
Step 5: Increase the generation count
$k' = k' + 1$
End while

b. Modified fuzzy C-means (MFCM) clustering

Fuzzy c-means is a clustering method which permits the situation of one dataset belonging to more than one cluster at a time. The suggested MFCM clustering provides better clustering performance compared to the conventional FCM clustering methods. In modified fuzzy c means clustering, Let $p = \{p_1, p_2, p_3, \dots, p_I\}$ be the set of data points after adaptive particle swarm optimization and $q = \{q_1, q_2, q_3, \dots, q_J\}$ be the set of centers. The pseudo code of modified fuzzy c-means clustering algorithm is given in algorithm 2,

Algorithm 2: pseudo code of modified fuzzy c-means clustering

Input: input $p = \{p_1, p_2, p_3 \dots, p_I\}$ be the set of data points after adaptive particle swarm optimization and $q = \{q_1, q_2, q_3 \dots, q_J\}$ be the set of initialized centers.
Output: Clustered data
Begin
1. Initialize the centroids, $q_j, j = 1, \dots J$
2. Calculate the fuzzy membership $J_{\sigma m}$ by equation (6),
3. At J -step: calculate the fuzzy centers vectors v_{ij} using (8)
4. Compute the weighted mean distance σ_i using (7)
5. Update the cluster centroids z_j
6. If algorithm converges then STOP;
7. Otherwise return to step 2 until the algorithm converges;
8. return {Cluster}

End

The MFCM algorithm allots data to every class by utilizing fuzzy memberships. The modified objective function for partitioning the input dataset into clusters is defined as,

$$J_{\sigma n} = \sum_{i=1}^{I} \sum_{j=1}^{J} (v_{ij})^n \frac{\|p_i - q_j\|^2}{\sigma_i} \quad (6)$$

in (6), p_i represents the data, $q_j$is the j^{th} cluster center and n is the constant esteem. Where, sigmoid function σ_i denotes the weighted mean distance in cluster i, and it is adapted for the effective clustering in (6) given by:

$$\sigma_i = \left\{ \frac{\sum_{j=2}^{k} v_{ij}^n \|p_i - q_j\|^2}{\sum_{j=1}^{k} v_{ij}^n} \right\}^{1/2} \quad (7)$$

The function of being member signifies the likelihood of data flew which come from same cluster. The probability of data in FCM algorithm is based on the distance of individual

insect with other team in same cluster. The functions of membership and cluster center vectors are updated by the velocity and particle positions by (8) and (9).

$$v_{ij} = \frac{1}{\sum_{l=1}^{J}\left(\frac{\left\|p_i - {}^{q_j}/_{\sigma_i}\right\|}{\left\|p_i - {}^{q_l}/_{\sigma_i}\right\|}\right)^{\frac{2}{n-1}}} \quad (8)$$

the clusters centroid values are computed by utilizing (9)

$$z_j = \frac{\sum_{i=1}^{I} v_{ij}^{n}.x_i}{\sum_{i=1}^{I} v_{ij}^{n}} \quad (9)$$

algorithm will continue running till the change between two iterations reach the ξ, for the given sensitivity threshold.

$$\max_{ij}\left\|V_{ij}^{(k)} - V_{ij}^{(k+1)}\right\| < \psi \quad (10)$$

where, ψ = a termination condition lying in the range of 0 and 1, while δ = the iteration steps. Repeat the steps until efficient clustering reached.

3.2. Phase 2: Effective Classification for Higgs Data Streams in Apache Spark

In the second stage, the Higgs data streaming is viably performed by pre-processing the input information. Then the pre-processed information is classified utilizing the modified support vector machine (MSVM) classifier with grid search optimization. At long last the optimized information is assessed in spark engine then the assessed esteem is utilized to discover the confusion matrix is accomplished. The proposed stage 2 work utilizing Higgs datasets for the data streaming in Apache Spark. The flow diagram of phase 2 methodology for the effective classification of higgs data streams is given in Figure 2.

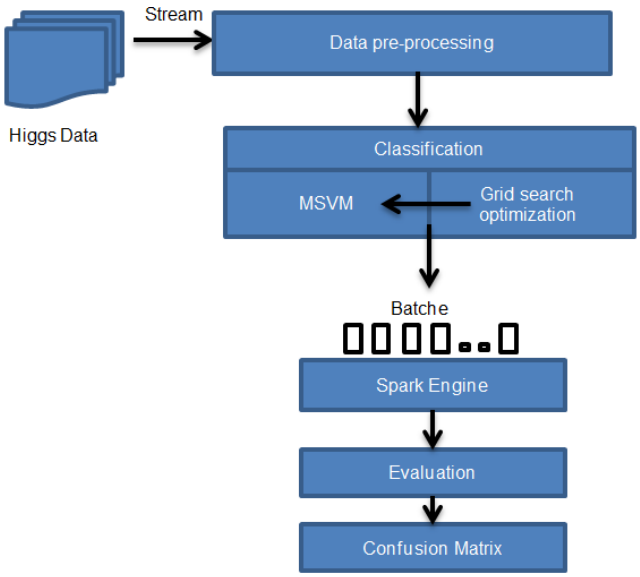

Figure 2. Flow diagram of phase 2 proposed methodology

3.2.1. Preprocessing

In the proposed Higgs data streaming, first the input information utilized for the proposed effective information streaming is taken from the dataset $D' = \{D_1', D_2', D_3', \ldots, D_n'\}$. Here, D' is the Higgs input dataset. Then the input Higgs information is preprocessed utilizing

tokenization and stop word removal processes which are utilized to expel conflicting information or noisy information from dataset. Here, the input data is first preprocessed by utilizing tokenization process given in (1) and subsequently tokenized data is processed by utilizing stop word removal process given in (2).

3.2.2. Data Streaming Classification Grid Search Based Modified Svm

The SVM as a binary classification method is reliant on the structural risk minimization approach. The SVM initiates by mapping the training data into a hyperplane which divides 2 classes of information in the feature space and maximize the edge of division among itself and those focuses lying closest to it. This decision surface would then be able to be utilized as a reason for categorizing unknown information [39]. SVM classification is improved by using network grid search optimization. The grid search improvement adequately tunes the SVM parameters for the better assortment.

In (11), Y = the input space, $y_i \in Y$= input vectors, $T = \{1, -1\}$ = target space, $t_i \in T$ = classes, and $S_t = \{(y_1, z_1), \ldots, (y_N, z_N)\}$ = training set. In the SVM, the most extreme edge hyperplane executes the partitioning of the 2 boundaries $T = \{1, -1\}$, i.e. the hyperplane which maximizes the closest distance to the data points and provides the optimum popularization on new models. Hence, a new point y_j can be categorized by first defining the assortment function $g(y_j)$:

$$g(y_j) = \operatorname{sgn}\left(\sum_{y_i \in sv} w_i t_i K(y_i, y_j) + b\right) \quad (11)$$

where, sv = the support vectors, $K(y_i, y_j)$= kernel function, w_i = weights, N= number of training samples, b = offset parameter. If $g(y_j) = +1$, y_j is in the positive class, if $g(y_j) = -1$, y_j is in the negative class. Training SVM requires the solution of the accompanying optimization issue expressed in (12) and (13) so as to attain the weight vector w and the offset b.

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C' \sum_{0}^{k} \xi_i \quad (12)$$

where (14) is subject to:

$$t_i(w^T \phi(y_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (13)$$

The reason for employing the Gaussian SVM which employs parameters C' and gamma (γ') is to transform the component vector space into the incensement of remoteness such that partition can be performed with higher accuracy. The diversion is accomplished using the kernel function $K(y_i, y_j) = \phi(y_i)\tilde{T}\phi(y_j)$, characterized for the Gaussian SVM is $K(y_i, y_j) = exp\left(-\gamma \|y_i - y_j\|^2\right), \gamma > 0$. The choice of proper learning parameters is a significant step in acquiring very much tuned support vector machines. For the most part, the settings of these parameters depend on a grid search. The pseudo code for the optimization of SVM parameter utilizing Grid search for better classification is given in algorithm 3.

The SVM initialize to main parameters C' and gamma (γ') and the procedure of optimization by isolating the hyper-plane to get identical way of work out the information and these are the parameter of SVM classifier for the regularization The parameter C' characterizes the mistake of data flew. When the value of C' increases the mistake rate also increases and brings down the number of permitted points in the error range. A smaller value of C'encourages a bigger error gap upon the isolation of the hyper-plane. For Gaussian SVM, the γ' parameter is determined as it affects its hyper-line adaptability. To reduce the values of γ', the hyper-plane line is almost linear, and for increasing the numbers, it works out to be progressively curved. Expanding the value of γ' to over-fitting on work out data. This grid search based modified SVM classification provides the effective process of data streaming.

Algorithm 3: Modified SVM with Grid search optimization

Set $C' = 10^k$
Set $\gamma' = 10^j$
Work out SVM with C' and γ' on Training Set
Evaluate SVM assortment on Validation Set
If rigor is better than Max rigor then
Save $MaxC' = C'$ and $Max\gamma' = \gamma'$
End if
End for
End for
Increment $C' = \max C'/m$
$C' = increment\ C'$
While $C' \leq \max C' * m$ do
Increment $\gamma' = \max \gamma'/m$
$\gamma' = increment\ \gamma'$
While $\gamma' \leq \max \gamma' * m$ do
Train SVM with C' and γ' on Training Set
Evaluate SVM classification on Validation Set
If precision is better than Max Precision then
Set $optimal\ C' = C'$ and $optimal\ \gamma' = \gamma'$
$\gamma' = \gamma' + increment\ \gamma'$ End while
$C' = C' + increment\ C'$
End while
Return $optimal\ C'$ and $optimal\ \gamma'$

4. Results and Discussion

The implementation of our proposed data streaming using adaptive clustering and classification is performed in the working stage of Java apache spark. The Twitter dataset and Higgs dataset is utilized to assess the proposed twitter data streaming. In order to investigate the performance of the proposed data streaming is distinguished with the existing artificial bee colony (ABC) optimization and Genetic algorithm (GA) techniques in regards of Recall, Precision, F-measure and Convergence.

4.1. Performance Analysis of Proposed Clustering

The statistical metrics of F-score, precision, and recall can be expressed in the terms of TP, FP, FN, and TN Where, TP (true positive), FP (false positive), FN (false negative) and TN (true negative) esteems. The performance of our proposed work is analysed by utilizing the statistical measures mentioned in this section.

4.1.1. Precision

The fraction of data recognized which are appropriate to the original data is termed as precision:

$$precision = \frac{TP}{TP + FN} \tag{14}$$

the comparison graph of proposed data streaming using improved fuzzy c-means clustering with existing Fuzzy C-means clustering (FCM) and K-means clustering in terms of precision is appeared in Figure 3. It depicts the proposed data streaming using improved fuzzy c-means clustering resulting well in terms of precision than the existing Fuzzy C-means clustering (FCM) and K-means clustering.

4.1.2. Recall

Recall ascertains the fraction of data which are appropriate to the query data that are effectively recognized.

$$recall = \frac{TP}{TP + FN} \quad (15)$$

The comparison graph of proposed data streaming using improved fuzzy c-means clustering with existing Fuzzy C-means clustering (FCM) and K-means clustering in terms of recall is appeared in Figure 4. It depicts the proposed data streaming using improved fuzzy c-means clustering (IFCM) resulting well in terms of recall than the existing Fuzzy C-means clustering (FCM) and K-means clustering.

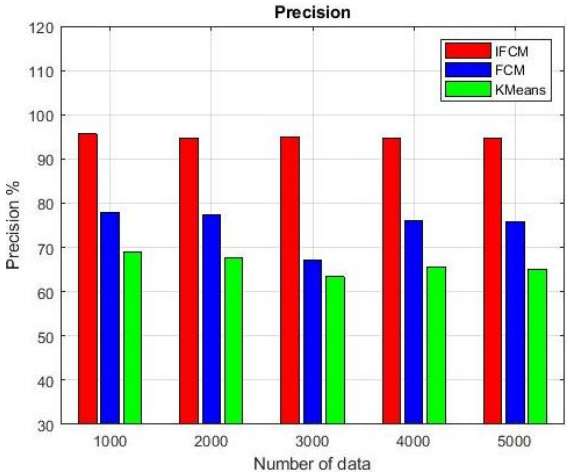

Figure 3. Comparison graph in terms of precision

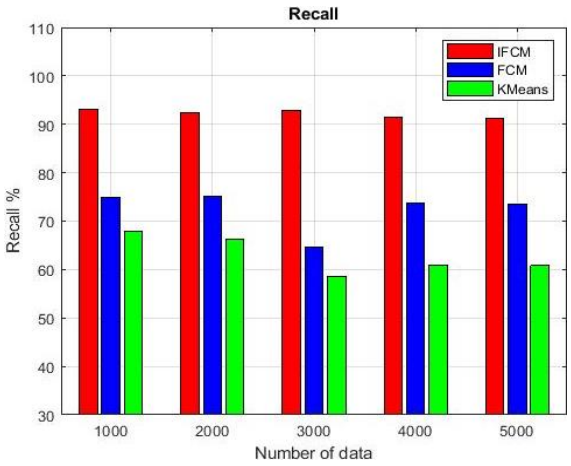

Figure 4. Comparison graph in terms of recall

4.1.3. F-Score

This value determines the accuracy of a test. The best F-measure value is 1 while the worst is 0. F-measure is computed using (16).

$$FMeasure = 2\frac{precision \times recall}{precision + recall} \quad (16)$$

The comparison graph of proposed data streaming using improved fuzzy c-means clustering with existing Fuzzy C-means clustering (FCM) and K-means clustering in terms of F-score is appeared in Figure 5. It depicts the proposed data streaming using improved fuzzy c-means clustering resulting well in terms of F-score than the existing Fuzzy C-means clustering (FCM) and K-means clustering.

4.1.4. Convergence Graph

The convergence graph of the suggested PSO using data streaming with ABC optimization and GA techniques is given in Figure 6. In the proposed PSO system, the convergence occurs between fitness and number of iterations is better than the existing ABC and GA convergence.

4.1.5. Computational Time

It is the quantity of time taken for the completion of proposed twitter data streaming. The computational time of data streaming in seconds can be obtained from the data stream size in bit and the bit rate in bit/sec as:

$$T_c = {}^{S_d}/_{b_r} \quad (17)$$

where, $\tilde{C}_T$ be the computational time of classification, D_s be the size of the data stream, B_r be the Bit rate. The performance result of our proposed IFCM with existing FCM and K-means

clustering in terms of computational time is given in Figure 7. It depicts the proposed data streaming using improved fuzzy c-means clustering (IFCM) achieved better computational time compared to FCM and K-means clustering. The comparison results regarding of various performance measures utilizing adaptive clustering is depicted in Table 1.

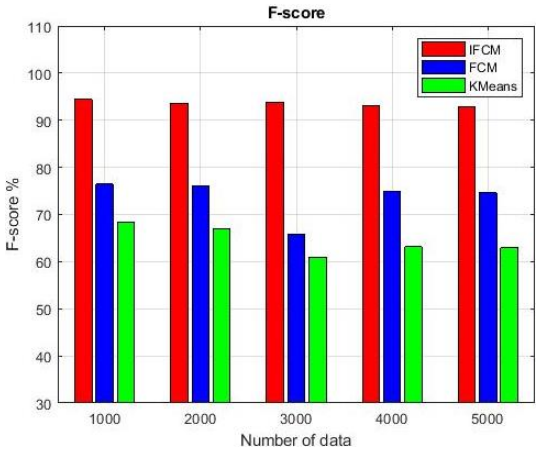

Figure 5. Comparison graph in terms of F-score

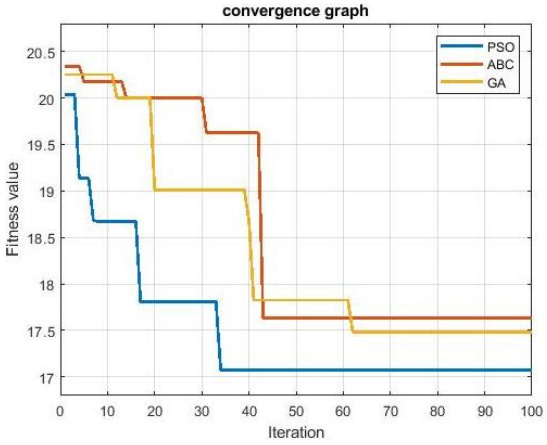

Figure 6. Convergence graph of proposed PSO utilized clustering with existing ABC and GA techniques

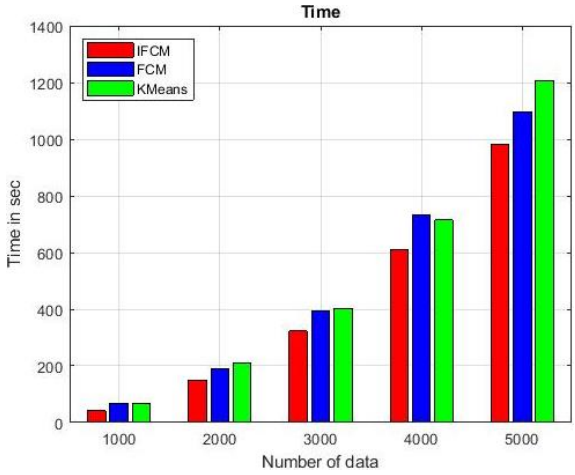

Figure 7. Comparison graph in terms of computational time

4.2. Average Classification Error Percentage

The comparison assessment of the classification error percentage is given in Table 2. The proposed modified support vector machine (MSVM) classification error percentage is significantly lesser than the existing SVM and Anti-Bayes Multi classification.

Table 1. Comparison of Proposed Clustering

Method	Precision	Recall	F-measure
Proposed IFCM	95.7	93.2	94.43
FCM	77.9	75	76.42
K-means	69.09	67.81	68.44

Table 2. Comparison of Proposed Classification in Terms Adaptive Clustering

Algorithm	Classification Error Percentage
SVM	26.99
Anti-Bayes Multi Classification	15.99
Proposed (MSVM)	13.13

4.2.1. Receiver Operating Characteristic (Roc) Curve

The ROC curve is a probability plot which expresses the fitness of a model in class recognition. The ROC curve is generated by plotting the TP values against the FP values. The assessment graph of proposed MSVM with existing Anti-Bayes Multi Classification and SVM in terms of ROC is displayed in Figure 8. The convergence graph of the proposed grid search optimized search optimization utilized classification with existing BAT and Cuckoo search optimization techniques is given in Figure 9. In the proposed system, the convergence occurs between fitness and number of iterations is better than the existing BAT and Cuckoo search optimization convergences.

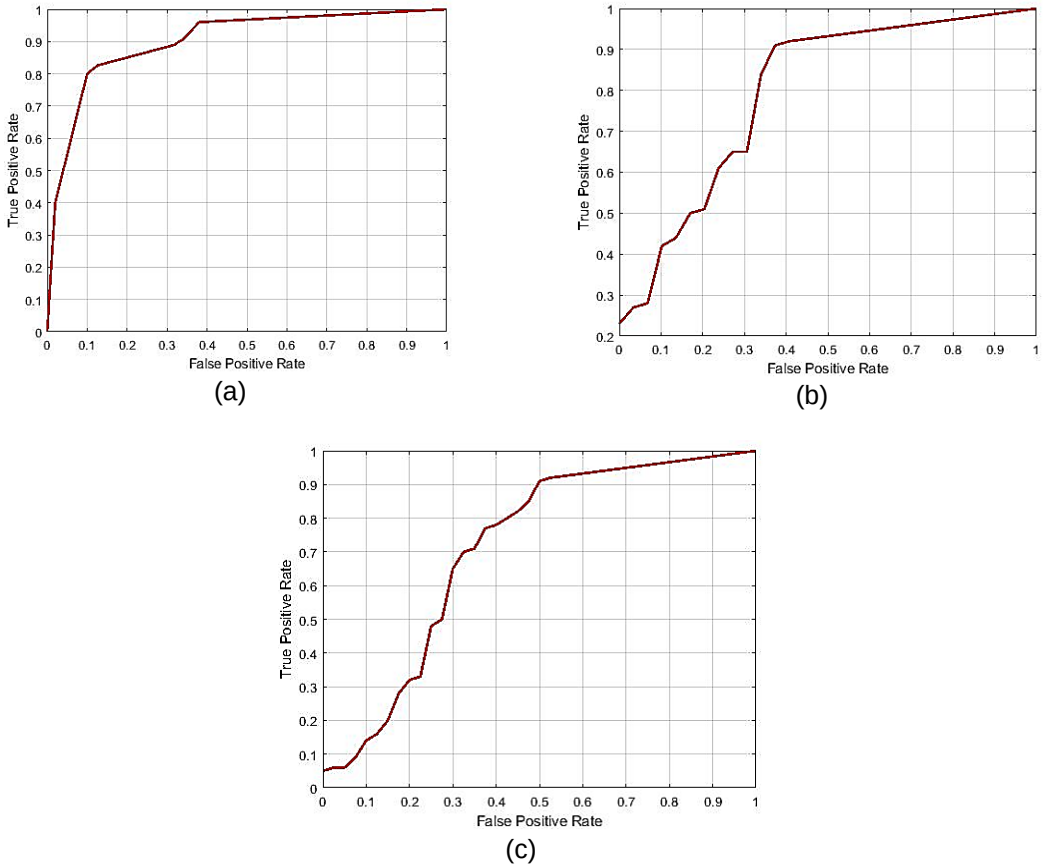

Figure 8. Comparison between ROC curves obtained for
the (a) proposed MSVM with existing (b) anti-Bayes multi classification and (c) SVM

4.2.2. Simulation Time

It is the quantity of time taken for the completion of proposed twitter data streaming. The computational time of data streaming in seconds can be obtained from the data stream size in bit and the bit rate in bit/sec as:

$$T_c = {}^{S_d}/_{b_r} \tag{18}$$

where, T_c is the simulation time of classification, S_d is the size of the data stream, b_r is the Bit rate. The performance result of our proposed MSVM with existing SVM and Anti-Bayes Multi Classification in terms of computational time is given in Figure 10. Figure 10 depicts the proposed data streaming using MSVM provides better results in terms of computational time than the existing SVM and Anti-Bayes Multi Classification.

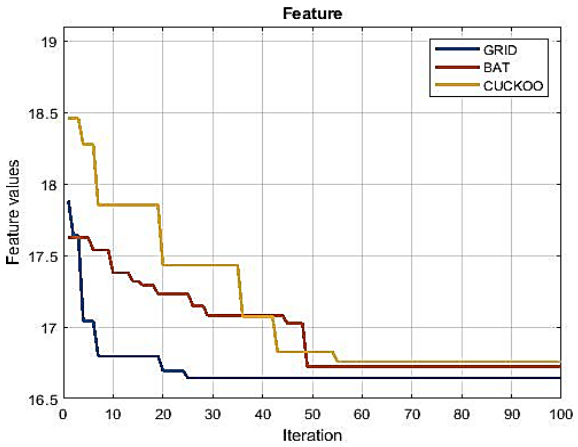

Figure 9. Convergence graph of proposed grid search optimization utilized classification with existing BAT and Cuckoo search techniques

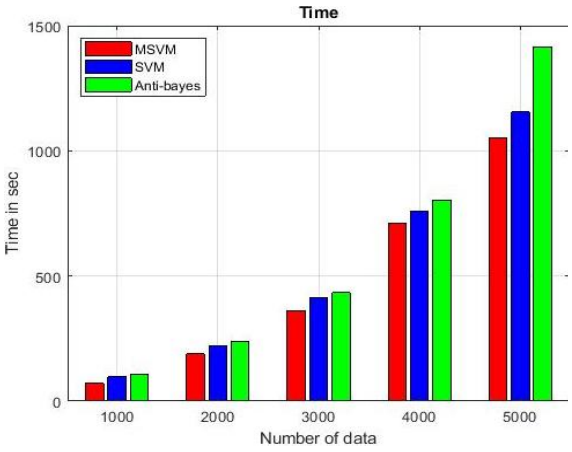

Figure 10. Comparison graph in terms of computational time

4.2.3. Accuracy

Accuracy is percentage of real outcome irrespective of the existence of TP or TN in a given population. It estimates the level of accuracy of a data classification process. Accuracy is computed using (19).

$$\text{Accuracy}=\frac{(\text{TN}+\text{TP})}{(\text{TN}+\text{TP}+\text{FN}+\text{FP})} \tag{19}$$

The comparison results regarding of accuracy using proposed MSVM classification with existing SVM and Anti-Bayes Multi Classification is depicted in Table 3. The comparison graph of proposed MSVM classification with existing SVM and Anti-Bayes Multi Classifications in terms of accuracy is shown in Figure 11. It illustrates the proposed MSVM classification provides better classification results than the existing SVM and Anti-Bayes Multi Classifications.

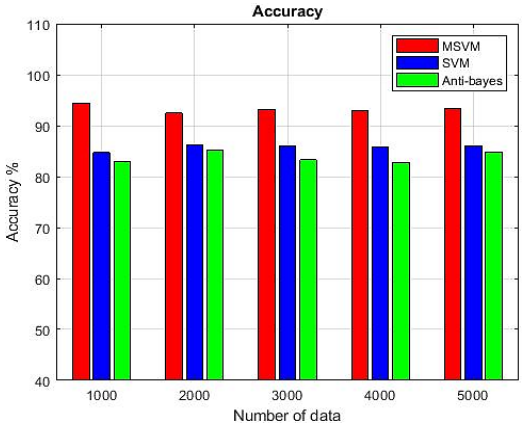

Figure 11. Comparison graph-based classification accuracy

Table 3. Comparison of Proposed Clustering based Accuracy

Dataset size	MSVM	SVM	Anti-Bayes Multi Classification
1000	94.44	84.70	83.01795
2000	92.45	86.30	85.2773
3000	93.21	85.99	83.2773
4000	92.90	85.79	82.8422
5000	93.33	86.08	84.8422

5. Conclusion

In this paper we have presented an effective twitter data streaming using adaptive clustering and classification algorithm. Here, the pre-processed data utilizing Improved Fuzzy C-means clustering effectively clusters the twitter information with improved by utilizing an Adaptive Particle swarm optimization (PSO) algorithm. Furthermore, the modified support vector machine (MSVM) classifier with grid search optimization effectively performs the twitter data streaming. The experimental outcomes exhibits that our proposed data streaming outperforms the existing ABC, GA optimized clustering and also existing SVM and Anti-Bayes Multi classifications regarding performance measures such as, accuracy, precision, recall, convergence, ROC curve and F-score. This results proves that the proposed clustering technique effectively process the twitter data streaming than the existing techniques and also the proposed classification technique effectively process the Higgs data streaming than the existing techniques.

References

[1] Danthala MK. Tweet analysis: twitter data processing using Apache Hadoop. *International Journal of Core Engineering & Management (IJCEM)*. 2015; 1: 94-102.

[2] Elzayady H, Badran KM, Salama GI. *Sentiment Analysis on Twitter Data using Apache Spark Framework.* 2018 13th International Conference on Computer Engineering and Systems (ICCES). 2018: 171-176.

[3] Tasoulis SK, Vrahatis AG, Georgakopoulos SV, Plagianakos VP. Real Time Sentiment Change Detection of Twitter Data Streams. *arXiv preprint arXiv: 1804.00482*. 2018.

[4] Wang, Yunli, and Cyril Goutte. *Detecting Changes in Twitter Streams using Temporal Clusters of Hashtags.* Proceedings of the Events and Stories in the News Workshop: 10-14. 2017.

[5] Panagiotou N, Katakis I, Gunopulos D. Detecting events in online social networks: Definitions, trends and challenges. In: Michaelis S, Piatkowski N, Stolpe M. Editors. Solving Large Scale Learning Tasks Challenges and Algorithms. *Springer, Cham*. 2016: 42-84.

[6] Wu Y, Cao N, Gotz D, Tan YP, Keim DA. Keim. A survey on visual analytics of social media data. *IEEE Transactions on Multimedia*. 2016; 18(11): 2135-2148.

[7] Cao J, Guo J, Li X, Jin Z, Guo H, Li J. Automatic rumor detection on microblogs: A survey. *arXiv preprint arXiv:*1807.03505. 2018.

[8] Anantharam P, Thirunarayan K, Sheth A. *Topical anomaly detection from twitter stream*. Proceedings of the 4th Annual ACM Web Science Conference. 2012: 11-14.

[9] Sakshi SU. Twitter Streaming API Using Apache Spark in Big Data Analytics. *International Journal of Scientific & Engineering Research*. 2016; 7(12): 354- 359.

[10] Ali AH, Abdullah MZ. *Recent trends in distributed online stream processing platform for big data: Survey*. 2018 1st Annual International Conference on Information and Sciences (AiCIS). 2018.

[11] Yadranjiaghdam B, Yasrobi S, Tabrizi N. *Developing a real-time data analytics framework for twitter streaming data*. 2017 IEEE International Congress on Big Data (BigData Congress). 2017: 329-336.

[12] Liu SM, Chen JH. A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*. 2015; 42(3): 1083-1093.

[13] Ahmed MA, Hasan RA, Ali AH, Mohammed MA. The classification of the modern Arabic poetry using machine learning. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2019; 17(5): 2667-2674.

[14] Miller Z, Dickinson B, Deitrick W, Hu W, Wang AH. Twitter spammer detection using data stream clustering. Information Sciences. 2014; 260: 64-73.

[15] Alrubaian M, Al-Qurishi M, Hassan MM, Alamri A. A credibility analysis system for assessing information on twitter. *IEEE Transactions on Dependable and Secure Computing*. 2018; 15(4): 661-674.

[16] Shi L, Wu Y, Liu L, Sun X, Jiang L. Event detection and identification of influential spreaders in social media data streams. *Big Data Mining and Analytics*. 2018; 1(1): 34-46.

[17] Liang S, Yilmaz E, Kanoulas E. Collaboratively tracking interests for user clustering in streams of short texts. *IEEE Transactions on Knowledge and Data Engineering*. 2019; 31(2): 257-272.

[18] Al-Khateeb T, Masud MM, Al-Naami KM, Seker SE, Mustafa AM, Khan L, Trabelsi Z, Aggarwal C, Han J. Recurring and novel class detection using class-based ensemble for evolving data stream. *IEEE Transactions on Knowledge and Data Engineering*. 2016; 28(10): 2752-2764.

[19] Nair LR, Shetty SD, Shetty SD. Applying spark based machine learning model on streaming big data for health status prediction. *Computers & Electrical Engineering*. 2018; 65: 393-399.

[20] Nagarajan SM, Gandhi UD. Classifying streaming of Twitter data based on sentiment analysis using hybridization. *Neural Computing and Applications*. 2018; 31(5): 1-9.

[21] Mohammed, NQ, Ahmed MS, Mohammed MA, Hammood OA, Alshara HAN, Kamil AA. *Comparative Analysis between Solar and Wind Turbine Energy Sources in IoT Based on Economical and Efficiency Considerations.* In 2019 22nd International Conference on Control Systems and Computer Science (CSCS). IEEE. 2019; 448-452.
[22] Hasan RA, Mohammed MA, Salih ZH, Bin Ameedeen MA, Ţăpuş N, Mohammed MN. HSO: A Hybrid Swarm Optimization Algorithm for Reducing Energy Consumption in the Cloudlets. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2018; 16(5): 2144-2154.
[23] Hasan RA, Mohammed MA, Ţăpuş N, Hammood OA. *A comprehensive study: Ant Colony Optimization (ACO) for facility layout problem*. 2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet). 2017: 1-8.
[24] Hasan RA, Mohammed MN. A krill herd behaviour inspired load balancing of tasks in cloud computing. *Studies in Informatics and Control*. 2017; 26(4): 413-424.
[25] Mohammed MA, Salih ZH, Ţăpuş N, Hasan RAK. *Security and accountability for sharing the data stored in the cloud*. 2016 15th RoEduNet Conference: Networking in Education and Research. 2016: 1-5.
[26] Mohammed MA, Ţăpuş N. A Novel Approach of Reducing Energy Consumption by Utilizing Enthalpy in Mobile Cloud Computing. *Studies in Informatics and Control*. 2017; 26: 425-434.
[27] Mohammed MA, Hasan RA. *Particle swarm optimization for facility layout problems FLP—A comprehensive study*. 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). 2017: 93-99.
[28] Salih ZH, Hasan GT, Mohammed MA. *Investigate and analyze the levels of electromagnetic radiations emitted from underground power cables extended in modern cities*. 2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). 2017: 1-4.
[29] Mohammed MA, Hasan RA, Ahmed MA, Tapus N, Shanan MA, Khaleel MK, Ali AH. *A Focal load balancer based algorithm for task assignment in cloud environment*. 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). 2018: 1-4.
[30] Z. Salih ZH, Hasan GT, Mohammed MA, Klib MAS, Ali AH, Ibrahim RA, editors. Study the Effect of Integrating the Solar Energy Source on Stability of Electrical Distribution System. 2019 22nd International Conference on Control Systems and Computer Science (CSCS); 2019 28-30 May 2019.
[31] Ahmed MA, Hasan RA, Ali AH, Mohammed MA. Using machine learning for the classification of the modern arabic poetry. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2019; 17(5): 2667-2674.
[32] Hasan RA, Mohammed MN, Ameedeen MAB, Khalaf ET. Dynamic Load Balancing Model Based on Server Status (DLBS) for Green Computing. *Advanced Science Letters*. 2018; 24(10): 7777-7782.
[33] Hammood OA, Nizam N, Nafaa M, Hammood WA. RESP: Relay Suitability-based Routing Protocol for Video Streaming in Vehicular Ad Hoc Networks. *International Journal of Computers, Communications & Control*. 2019; 14(1): 21-38.
[34] Hammood OA, Kahar MNM, Mohammed MN.. *Enhancement the video quality forwarding Using Receiver-Based Approach (URBA) in Vehicular Ad-Hoc Network*. 2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET). 2017: 64-67.
[35] Ayoob AA, Su G, Wang D, Mohammed MN, Hammood OA. *Hybrid LTE-VANETs based optimal radio access selection*. International Conference of Reliable Information and Communication Technology. 2017: 189-200.
[36] Salh A, Audah L, Shah NSM, Hamzah SA. Pilot reuse sequences for TDD in downlink multi-cells to improve data rates. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2019; 17(5): 2161-2168.
[37] Al-janabi HDK, Al-janabi HDK, Al-Bukamrh RAH. Impact of Light Pulses Generator in Communication System Application by Utilizing Gaussian Optical Pulse. In 2019 22nd International Conference on Control Systems and Computer Science (CSCS). IEEE. 2019; 459-464.
[38] Hammood, OA Kahar, MNM Mohammed, MN Hammood WA, Sulaiman J. The VANET-Solution Approach for Data Packet Forwarding Improvement. *Advanced Science Letters*. 2018; 24(10): 7423-7427.
[39] Mohammed, MN Nahar, AK Abdalla, AN Hammood, OA. *Peak-to-average power ratio reduction based on optimized phase shift technique.* In 2017 17th International Symposium on Communications and Information Technologies (ISCIT). IEEE. 2017; 1-6.