Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 4, 2020

AINTUITIONISTIC FUZZY C-MEANS CLUSTERING BASED FEATURE SELECTION AND ENSEMBLE SUPPORT VECTOR MACHINE FRAMEWORK FOR BIG DATA CYBER-SECURITY

G.A. MYLAVATHI[1], Dr.B. SRINIVASAN[2]

[1]Assistant Professor, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam. mylavathiga@gmail.com
[2]Associate Professor of Computer Science, Gobi Arts &Science College, Gobichettipalayam.srinivasan-gasc@yahoo.com

Received: 05.01.2020 ***Revised: 12.02.2020*** ***Accepted: 22.03.2020***

ABSTRACT

In enterprises and industries, rapid progress in attained by information technology in recent days and which makes the term big data, a very popular one. Various sources like business record, digital videos, pictures in digital format, social media are producing data and which makes the expansion in data growth is a rapid one. Big data corresponds to management of this huge data and it is a challenging task. In apps and operating systems, there is a malicious software insertion, which makes big data hacking as a serious threat. So, enhanced security techniques are required for securing big data from cyber threats. Malicious software and unknown patterns are classified by proposing Machine Learning (ML) algorithms. For big data cyber security, an Ensemble Support Vector Machine (ESVM) framework is designed in previous system. Improved K-means clustering is used initially for selecting features. A bi-objective optimization problem is processed in this work by formulating ESVM configuration, where complexity of model and accuracy are considered as two conflicting objectives. Cuckoo Search (CS) optimization algorithm is used for performing bi-objective optimization. Due to local optimum results and slow convergence rate, CS is not used very often. So, it requires a better algorithm for selecting features and for enhancing accuracy. For avoiding this, for big data cyber security, a new framework is introduced in this work. In which, detection of malicious activity is done using an Ensemble Support Vector Machine (ESVM) framework. Intuitionistic fuzzy C-means clustering is used for selecting features initially in this work. ESVM technique is for performing malware and intrusion detection based on selected features. For solving multi objective function and optimizing parameters, a hyper metaheuristics algorithm framework is introduced in this work. Improved Artificial Bee Colony (IABC) optimization algorithm is used for parameter optimization.

Keywords: Feature Selection and Artificial Bee Colony, Cyber Security, Local Optimal Solution, Ensemble Support Vector Machine, Hacking Enhanced Security Methods, BigData.

DOI: xxxxxxx

INTRODUCTION

With a rapid increase in Internet's in-depth integration with social like, working and leaning of people are changed in recent days. But, there will be serious security threats. Different network identification, especially unknown attacks needs high concentration in recent days. Destruction, alteration, unauthorized access, data attacks, protection of networks, computers are provided by designing set of technologies in cyber security.

There are computer security system and network security system in network security system. Intrusion detection systems (IDS), antivirus software, firewalls are included in every systems in this. Unauthorized system behaviors like destruction, modification, copying and use are identified and determined using IDS [1,2, 3].

Internal intrusions and external intrusions are included in security breaches. For IDSs, there are three major network analysis. They are, hybrid, anomaly-based and signature-based. Attacks signatures are used in misuse-based detection techniques for detecting known attacks. Without generating huge false alarms attacks, they are used for known attacks types. But, it requires manual update of database rules and signatures by administrators.

According to misused technologies, there is no possibility to detect to zero-day attacks. System behavior and normal network studied using anomaly based techniques and form normal behavior, anomalies are identified. For every network, application and system, normal activity profiles are customized, which makes difficult for attackers for knowing undetected activities [4, 5].For misuse detectors, signatures can be defines using data on which anomaly-based techniques alert. For false alarm rates, potential of anomaly-based techniques is less, which is a major disadvantage of it, because previously unseen system behaviors can be categorized as anomalies. Anomaly and misuse detections are combined to from hybrid detection. Known intrusions, detection rate can be enhanced using this and for unknown attacks, false positive rate is reduced [6,7, 8].To handle, problems in big data security, Machine learning algorithms can be used. Malicious software and unknown patterns can be classified using Machine Learning (ML) algorithms. In unknown malware software identification and classification, promising results are shown by ML. In different real world problems, remarkable success is shown by a class of ML algorithm called Support Vector Machines (SVMs).

Scalability and strong performance makes SVM more popular. However, with these advantages, selected configurations, affects SVM performance in a huge manner, soft margin parameter (or penalty) selection and kernel type with its parameters are the typical SVM configuration [9, 10].So, enhanced security techniques are required for securing big data from cyber threats. Malicious software and unknown patterns are classified by proposing Machine Learning (ML) algorithms. For big data cyber security, an Ensemble Support Vector Machine (ESVM) framework is designed in previous system. Improved K-means clustering is used initially for selecting features. A bi-objective optimization problem is processed in this work by formulating ESVM configuration, where complexity of model and accuracy are considered as two conflicting objectives.Cuckoo Search (CS) optimization algorithm is used for performing bi-objective optimization. Due to local optimum results and slow convergence rate, CS is not used very often. So, it requires a better algorithm for selecting features and for enhancing accuracy. For avoiding this, for big data cyber security, a new framework is introduced in this work. In which, detection of malicious activity is done using an Ensemble Support Vector Machine (ESVM) framework. Intuitionistic fuzzy C-means clustering is used for selecting features initially in this work. ESVM technique is for performing malware and intrusion detection based on

selected features. For solving multi objective function and optimizing parameters, a hyper metaheuristics algorithm framework is introduced in this work. Improved Artificial Bee Colony (IABC) optimization algorithm is used for parameter optimization.

LITRATURE REVIEW

Various cyber security systems are reviewed in this section using machine learning techniques. Zhang, et al [2017] learned big data features using double-projection deep computation model (DPDCM), where in hidden layers, raw input is projected as a two separate subspaces for learning big data's interacted features. In this conventional deep computation model (DCM)'s hidden layers are replaced by double-projection layers. Further, for training DPDCM, a learning algorithm is devised and its efficiency is enhanced using cloud computing via crowd sourcing the data on cloud.

Based on BGV encryption scheme, proposed a privacy-preserving DPDCM (PPDPDCM) for protecting private data. On Animal-20 and NUS-WIDE-14 dataset, carried out the experimentation for estimating PPDPDCM and DPDCM performance and it is compared with DCM. When compared with DCM, better results are achieved by DPDCM, as demonstrated by experimentation results. Especially, training parameter efficiency can be enhanced effectively using this PPDPDCM and proves its potential in big data feature learning.

Yin et al [2017] explored deep learning for modeling an intrusion detection system and recurrent neural networks (RNN-IDS) is used for proposing a deep learning approach for detecting intrusion. In multiclass classification and binary classification, model performance is studies and proposed model's performance is affected by neurons count and various learning rate.

Various machine learning techniques proposed by different researchers, support vector machine, random forest and artificial neural network are used fro making performance comparison. With high accuracy, a classification model can be modeled using RNN-IDS as demonstrated by experimental results. In multiclass as well as in binary classification, better performance is exhibited by this proposed method and it is superior to that of traditional machine learning classification techniques.

Jabbar et al [2017] detected intrusion by introducing Spark-Chi-SVM model. In this model, selection of features are done using ChiSq Selector and on Apache Spark Big Data platform, support vector machine (SVM) classifier is used for constructing intrusion detection model. Model is trained and tested using KDD99. In this experiment, Chi-Logistic Regression classifier and Chi-SVM classifier are compared. Better performance is exhibited by Spark-Chi-SVM model as shown in experimentation results and training time is also reduced by this.

Xie et al [2018] considered anomaly degree of whole provenance graph and both a single provenance path for proposing a hybrid technique called Pagoda. Intrusion can be identified quickly by this, if a serious compromise has been found on one path. In whole provenance graph, behavior representation is considered for enhancing detection rate.

For storing provenance, persistent memory database is used in Pagoda and for maximally reducing unnecessary I/O in detection analysis, it aggregates multiple similar items into one provenance record. In rule database, duplicate items are encoded in addition and noises without intrusion information are filtered. Its efficiency and performance are demonstrated on real-world applications.

Peng et al [2018] proposed a grouping technique for IDS dependent on Mini Batch K-implies joined with head part examination. Initial, a preprocessing strategy is proposed to digitize the strings and afterward the informational collection is standardized to improve the bunching proficiency. Second, the main part investigation technique is utilized to decrease the element of the prepared informational collection expecting to additionally improve the bunching proficiency, and afterward smaller than normal group K-implies strategy is utilized for information grouping.

All the more explicitly, we use K-means++ to instate the focuses of group so as to dodge the calculation getting into the neighborhood ideal, likewise, pick the Calsski Harabasz pointer with the goal that the bunching result is all the more handily decided. Contrasted and different strategies, the trial results and the time intricacy examination show that our proposed technique is compelling and proficient. Most importantly, our proposed grouping technique can be utilized for IDS over huge information condition.

Hassan, et al [2020] Proposed a crossover profound learning model to effectively recognize organize interruptions dependent on a convolutional neural system (CNN) and a weight-dropped, long momentary memory (WDLSTM) arrange. Utilize the profound CNN to extricate important highlights from IDS large information and WDLSTM to hold long haul conditions among removed highlights to forestall over fitting on repetitive associations. The proposed mixture strategy was contrasted and conventional methodologies as far as execution on a freely accessible dataset, exhibiting its agreeable execution.

Wheelus et al [2016]proposed and assessed a major information design that is established continuously organize traffic handling, dispersed informing and adaptable information stockpiling. The key advancement behind the proposed design is that it robotizes the investigation of heterogeneous system information, permitting the attention to stay on concocting compelling digital danger insight examination, instead of being prevented by information the executives, total, compromise and organizing. Experimental assessments researching the utilization of AI investigation by abusing the relics of the proposed engineering and by utilizing 100 GB of genuine system traffic, in fact exhibit the common sense, adequacy, and included estimation of the proposed design.

PROPOSED METHODOLOGY

Hyper meta-heuristic framework is applied in this research work and it applies Improved Artificial Bee Colony algorithm for producing configuration for devising ESVM in multi objective optimization function. With regard to system complexity and accuracy, it is highly required for enhancing system performance. So, for categorizing and recognizing malware detection and intrusion detection, applied ESVM. Intuitionistic fuzzy C-means clustering is employed for selecting features and for reducing feature dimensions, which tends to lessen classification complication. Fig.1 shows the proposed system's overview.

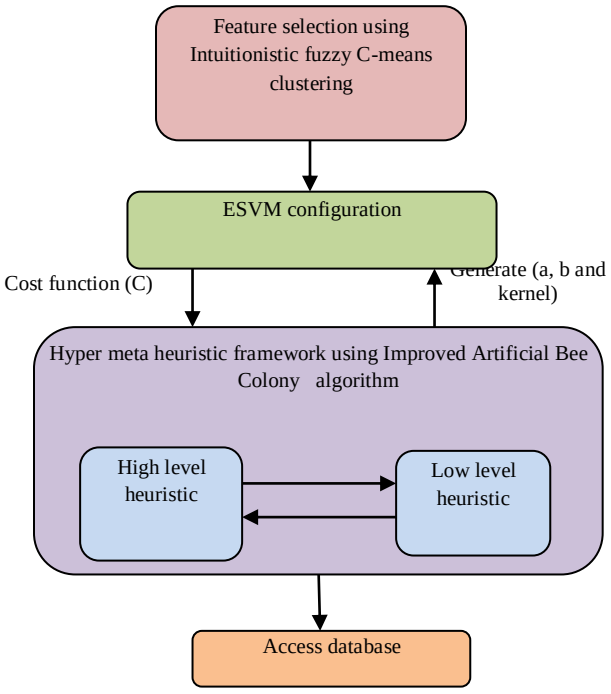

Fig. 1: Overview of Proposed Methodology

Feature Selection using Intuitionistic Fuzzy C-Means Clustering

From data sample, important features are selected by using fuzzy c-means in this research. This is done for enhancing the performance of classification and time complexity reduction. Fuzzy c-mean (FCM) clustering is a k-mean clustering fuzzy type. In FCM, fuzzy technique combination is used for allowing data inclusion in all clusters with various membership degree. Fuzzy clustering method is split into a set B in k clusters and every set is with N members as B ={b_1; b_2; b_3; . . .b_N}. There may be a uncertainty condition that data b_l is consigned to various clusters with different membership degree $u_{lm[18.19]}$. Data belongingness to a cluster is decided by paralleling its dissimilarity or distanced d_{lm}^2 from cluster centroid v_m. Euclidean formula is used for measuring distances.

$$FCM = \sum_{m=1}^{K} \sum_{l=1}^{N} u_{ij}^{p} d_{lm}^{2}, P \in (1,\infty), \sum_{m=1}^{k} u_{lm} = 1 \quad (1)$$

Over objective function, membership degree influence is mechanized using fuzzifier parameter (p). Expression (2) represents membership degree value and expression (3) represents centroid value.

$$u_{lm} = \frac{1}{\sum_{q=1}^{k}\left(\frac{d_{lm}^2}{d_{lq}^2}\right)^{\frac{1}{p-1}}} (2)$$

$$v_m = \frac{\sum_{l=1}^{N} u_{lm}^{P} x_l}{\sum_{l=1}^{N} u_{lm}^{p}} (3)$$

High sensitivity to noise and better performance is produced by FCM clustering results when compared with k-mean results. But, there is a limitation that, for each data, cluster-wise addition of all membership degrees to one produces abnormal points to be clusters members. These FCM drawbacks are resolved by integrating fuzzy c-mean with possibilisic approach and it is termed as possibilistic fuzzy clustering (PFCM). Major expressions of possibilistic fuzzy c-mean is expressed as,

$$\text{PFCM} = \sum_{m=1}^{K} \sum_{l=1}^{N} u_{lm}^{p} d_{lm}^{p} + \sum_{m=1}^{k} \lambda_m \left(\sum_{l=1}^{N} 1 - u_{lm}\right) (4)$$

Expression (5) represents membership degree function and expression (6) represents positive number values.

$$u_{lm} = \frac{1}{1+\left(\frac{d_{lm}^2}{\lambda_m}\right)^{\frac{1}{p-1}}} (5)$$

$$\lambda_m = w \frac{\sum_{l=1}^{N} u_{bm}^{p} d_{lm}^{2}}{\sum_{l=1}^{N} u_{lm}^{p}} (6)$$

Where, amendable weight is represented as W and it is set to 1 typically. Expression (3) in FCM is used for achieving an optimum solution for updating centroid. Expression (4) can be minimized really, if all clusters are coincident clusters. Gap between data and specific cluster defines membership function of expression (5), which does not consider any other clusters [20, 21].

Intuitionistic fuzzy clustering algorithm forms base for another improved fuzzy clustering. Intuitionistic fuzzy sets are used for modifying conventional fuzzy c-mean function. Fuzzy C-mean technique is integrated with intuitionistic properties by modifying cluster centers.

Proposed intuitionistic fuzzy sets with its hesitation degree is also explained. It is not true always, as membership and non-membership degree is being 1. This may produce hesitation degree and it is defined as 1 minus sum of membership and non-membership degrees. Following specifies hesitation degree,

$\pi_A =$ hesitation_ degree = 1 _ (membership_ degree +non_ membership _degree) (7)

Expression (8) is used for computing hesitation degree initially and intuitionistic fuzzy membership values are computed as,

$$u_{lm}^{*} = u_{lm} + \pi_{lm} (8)$$

Where, in l^{th} class, m^{th} data's intuitionistic fuzzy membership is represented as $u_{lm}^{*}(u_{lm})$. In expression (3), expression (9) is replaced to find a modified cluster center is computed as,

$$v_m^{*} = \frac{\sum_{l=1}^{N} u_{lm}^{*P} x_l}{\sum_{l=1}^{N} u_{lm}^{*p}} (9)$$

Update the cluster center using expression (9) and membership matrix is also updated simultaneously.

In every iteration, center of cluster and membership matrix are updated and if previous matrix and updated matrix are same, algorithm is stopped. Intuitionistic fuzzy sets is used for modifying conventional FCM's criterion function.

IPFCM Algorithm 1

Step 1. Initialization: Parameters of proposed algorithm and B; k; d; u; v are initialized.

Step 2. Compute

$$\text{PFCM} = \sum_{m=1}^{K} \sum_{l=1}^{N} u_{lm}^{p} d_{lm}^{p} + \sum_{m=1}^{k} \lambda_m \left(\sum_{l=1}^{N} 1 - u_{lm}\right)$$

Sub-Step 2 (a). Find $u_{lm} = \frac{1}{1+\left(\frac{d_{lm}^2}{\lambda_m}\right)^{\frac{1}{p-1}}}$

Sub-Step 2(b). Find $u_{lm} = w \frac{\sum_{l=1}^{N} u_{lm}^{P} d_{lm}^{2}}{\sum_{l=1}^{N} u_{lm}^{p}}$

Step 3. Expression (7) is used for computing hesitation degree initially.

Step 4. Intuitionistic fuzzy membership value is computed as: $u_{lm}^{*} = u_{lm} + \pi_{lm}$, where, $u_{lm}^{*}(u_{lm})$ represents intuitionistic fuzzy membership of m^{th} data in l^{th} class.

Step 5. Replace Sub-Section Eq. 2(b), modified cluster center is: $\lambda_m = w \frac{\sum_{l=1}^{N} u_{bm}^{p} d_{lm}^{2}}{\sum_{l=1}^{N} u_{lm}^{p}}$ clustercenter is updated and simultaneously membership matrix is updated.

Step 6. Iteration termination is computed. Convergence criterion is checked. Iteration is stopped, if convergence is reached, otherwise, go to Step 2.

Ensemble SVM

Attributable to the assortment of the areas perceived from the adherent of tweets, it is captivate to examine if an assortment of AI calculations is sufficiently fit to use the different choice limits delivered from the individual classifiers to purposely blend the consequences of order and in this manner a prevalent working is acknowledged than is possible with a solitary classifier. In classifying target audience from followers list, this research work concentrated on SVM assemblies potential [20]. In the following section, introduced SVM in addition to bootstrapping technique and algorithms of ensemble [22, 23].

- ***SVM Configuration***

In various applications, SVM , which is a supervised learning technique, used for two-or multi-class classification and categorization of text can also be done effectively using this. Through a hyperlane, it pulls out a known given set of {+1, -1} labelled training data, which are highly distant from negative and positive samples. In input space, this ideally parts hyperplance in feature space related to non-linear decision boundary.

A N distinct samples (xi,yi) with xi∈R^D and yi∈R^d is considered for modelling SVM as,

$$\sum_i a_i K(x, x_i) + b, i \in [1, N] \quad (10)$$

Where, kernel function is represnted as K(x, xi), SVM parameter is represented as a and its threshold is represented as b. Two major goals to be enhanced is expressed as,

$$\min_{s.t} F(x) = |f_1(x), f_2(x)|, \; f_1(x) = error\,, f_2(x) = NSV \quad (11)$$

Where, cost function C is represented as f(x), misclassified datasets count is represented as err and support vectors quantity is signified by NSV.

- ***Bootstrapping Using a Single SVM Model***

Bootstrapping is a common a technique used for tackling imbalance data issue via minority class resampling by means of replacement. As we are concerning about temporal effect, there will be restriction on tweets count which are shared by various owners within 6 months duration.

So, it is impossible for gathering enough samples for avoiding difficulties of either initiating unfairness in minority class preference or running risk of information loss in majority class.

An alternative to conventional distributional assumptions, Bootstrap sampling bring a computation style, non-parametric technique is accepted by it for statistical inference with the aim of better sample

distribution estimation rather than sample replication. It also offers, boots trapping's pseudocode description.

- ***Ensembles Using Multiple SVM Models***

This research work majorly concentrates on finding when to utilize tweets of account owner for computing audience in list of followers. It is necessary for analyzing, if training datasets of various domains are used for constructing classifier ensemble. When compared with pre-defined general bootstrapping method, superior performance can be achieved using this.

Irrespective of everything, success of ensemble system is mostly relies on types of classifiers used for constituting the ensemble. In this study, applied ensemble learning algorithms set has, majority vote, stacking and bagging. Fig. 2 shows, ensemble classifier's common architecture, where various SVM models are exploited. In every ensemble learning algorithms, aggregation techniques differs a lot.

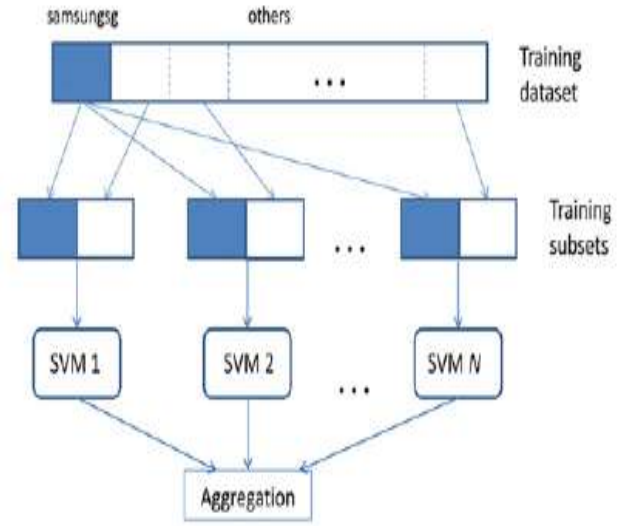

Fig. 2: A General Architecture of Ensemble System Using Multiple SVM Models

In various domains, this architecture is accepted as a multiplicity manipulation function due its configurations, various training datasets and different algorithms.

- ***Random Sampling with Majority Vote***

Splitting majority dataset class into various subsets via random sampling prior to joining with minority class for developing a balanced training dataset for classification purpose is a one of a plainest solutions to imbalance data issue. A straightforward majority vote of individual classifier decisions are gathered for combining or aggregating it.

On majority class, random sampling is performed in place of minority class, which is a conflict to bootstrapping algorithm. Major objective here is to form a small size minority class subset by partitioning majority dataset. Until reaching required size, randomly select a majority class in existing record and it will be placed in a subset.

Until generating all subsets, recurred this random selection procedure. Records are distinctive within same subset but, these records duplications are placed in various subsets.

- ***The Majority Vote Algorithm***

Input:

D represents training dataset having labels representing C classes
L represents learning algorithm
W represents training dataset labels
N represents employed L's count
Do n=1 to N

1. Call L with Dn and obtain classifier Ln.
2. Compare Cn and Wn which are created from Ln, vote is updated.
3. Aggregate vote to ensemble.

End

- ***Framework of Hyper Meta Heuristics Algorithm Using Improved Artificial Bee Colony Algorithm***

The SVM parameters are optimized in this work by using Improved artificial bee colony algorithm.

Basic ABC Algorithm

In ABC, artificial bees colony has three classes of bees: They are employed, onlooker and scout bees. Specific food sources are associated with employed bees, employed bees dance are being watched by onlooker bees within hive for selecting food sources and food sources are searched randomly by scout bees. In initial phase, scout bees discovers, positions of all food sources.

Then, onlooker and employed bees exploits, food sources nectar and continuous exploitation will make food sources to become exhausted.

Then, scout bees are formed from employed bees. Employed bees exploiting exhausted food sources will become scout bee for searching food sources again [24].

In ABC algorithms, possible solution of a problem is represented by a food source position and associated solutions quality is represented by food sources nectar amount. Every employed bee is associated with only one food source, because of this, employed bees count is equal to food source count. Following describes the details of this algorithm [25].

$$\min f = f(x)\,,\ \mathrm{x}(x_1 x_2,\ldots x_m \in s, s = [x_{iL}, x_{iH}] \quad (12)$$

Where, objective function is represented as f, i th-dimensional variable's upper and lower bounds are indicated using a m -dimensional variable $[x_{iL}, x_{iH}]$ called x.

With N number of onlooker and employed bees, major steps of this algorithm is described as follows.

a) Initialization

In a random manner 2N locations are produced and they are evaluated. Employed bees are moved into N food sources having high nectar values.

b) Using expression (13), new food source around employed bees are explored by them.

$$V_{ij} = x_{ij} + R_{ij}(x_{ij} - x_{kj}) \quad (13)$$

Where, new location is represented as V_{ij}, random number is represented as R_{ij}, which lies between -1 to 1, $k \in \{1,2,3,\ldots, N\}$ and $k \neq i$.

c)From candidate food sources between a) and b), N food sources having high nectar amount are marked.

d) New food sources are explored by onlooker bees.

In roulette wheel selection technique, on selected food sources, onlookers are placed. Then, food source x_i's neighborhood are explored by every onlooker bee using (13),

Probability P_i is computed as:

$$P_i = \frac{fit_i}{\sum_{i=1}^{N} fit_i} \quad (14)$$

Where, fit_i is computed as:

$$fit_i = \begin{cases} \frac{1}{1+f_i}, f_i \geq 0 \\ 1 + abs(f_i), f_i < 0 \end{cases} \quad (15)$$

Where, solution's fitness value is represented as f_i.

e) A solution will be abandoned, if it is not enhanced in "limit" trials. This solutions are replaced by new solution which are produced by scout bees.

f) Between candidate solutions generated in step c)and step d), N better solutions are selected

g) Till now, obtained best solution is recorded and until maxiterations, step b) to f) are repeated.

- ***Improved ABC Algorithm***

A simple, robust as well as easily controlled algorithm is a basic ABC algorithm. Convergence characteristics of ABC algorithm is slow due its random optimization nature and it gets stuck with local solutions easily. In order to get a better value of optimization, basic ABC algorithm is modified in this work.

- ***Gaussian Mutation***

In an individual vector, every element is added with a random value from a Gaussian distribution in Gaussian mutation for creating new offspring. Following expression is employed in this method.

$$\text{mutation}(x) = \xi*(1+ N(0,1)) \quad (16)$$

Where, numerical value of every object is represented as x, random value which is extracted from normal or Gaussian distribution is represented as N(0,1), for an individual, new value after Gaussian mutation is represented as mutation(x).

Based on pre-determined probability, selected the individuals and under Gaussian distribution, at probability, computed their positions. At local search stage, there is a possibility for wide range search. In final stage, improved ABC algorithm with Gaussian mutation is used for enhancing searching efficiency.

- ***Flow of Improved ABC Algorithm***

Following summarizes the Improved ABC algorithm with Gaussian mutation:

Step1. Initial population, which is a food source position is generated.

Step2. Expression (13) defines employed bees for exploring new food sources termed as offspring.

Step3. Every food sources fitness value is compared with its respective offspring and better performance location are marked.

Step4. Roulette wheel selection technique defines onlookers for selecting food sources computed from step 3 and new position around it are explored, so called offspring population.

Step5. Between off spring population and marked locations, food sources having better performance are selected as true food sources and to these locations, employed bees are moved.

Step6. All food sources mean fitness value are computed. For every bee, carried out Gaussian mutation, if mean fitness value is less than its fitness value, else, corresponding bee is abandoned.

Step7. Satisfaction of terminating condition is judged. If it is satisfied, optimum solution is produced as output and end, else, move to step 2.

RESULT AND DISCUSSION

For ensuring accuracy and model complexity in cyber security protection methods, algorithm comparison is done with benchmarked techniques. Sher Lock data collection agent is influenced by Google Funf framework. In Funf Open Sensing framework, for data processing, a framework was developed by MIP Media Lab, specifically for mobile devices. Intensified frequent feature monitoring like motion sensors statistics computing cannot used Funf.

So, for modifying robustness and stability steadiness, there is a need of pipeline processing framework. On every running application, information are gathered by incorporating probes with that framework. Sensors are used for obtaining virtual or physical data sources like memory intake and exterior temperature. PULL and PUSH type sensors are generally used.

Arrival of SMS or screen gets on are sensed by Event-based PUSH sensors. CPU sample or device are accelerated using gathered PULL sensors. With JSON format, in a temporary basis, data gathered by Sher Lock's are stored on volunteer's device in text format file. If file size exceeds 500MB, it will be zipped to ~50MB. Next, when end user bond to Wi-Fi, and in sequence, temporarily stored zip files on device are get into server. App Profiling & Malware Detection of cyber security research uses Sher Lock dataset. Applications implicitly operation is used for performing App's malware profiling and detection.

There are different contextual features like battery utilization, device location and movement in dataset, which are used to improve detection of malicious threat. With respect to f-measure, precision, recall and accuracy, proposed ESVM with IPFCM approach's performance is compared available ESVM-KMC , Online Support Vector Machines (O-SVM)[26] , Hyper-Heuristic Support Vector Machines (HH-SVM) [27] and XGBoost (AE).

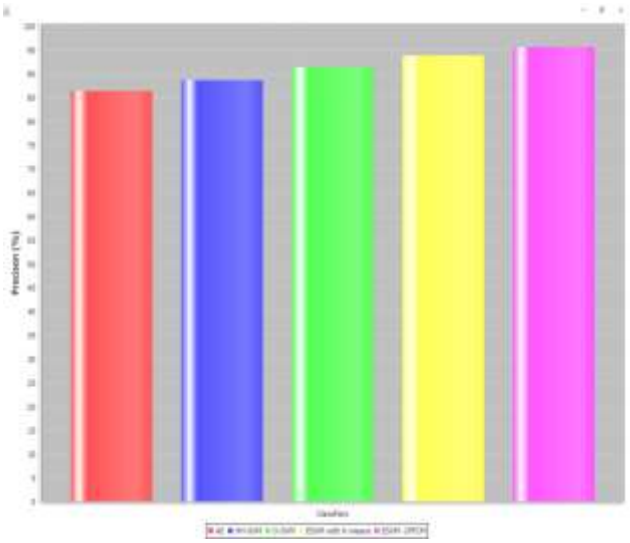

Fig. 3: Precision Results of Classifiers Vs Comparison

Precision metric comparison of available EVSM-KMC, O-SVM, HH-SVM, AE and proposed ESVM – IPFCM techniques are shown in fig.3. In this graph, various techniques are represented in x-axis and in y-axis, precision metric is represented. Improved fuzzy c-means clustering is used for selecting features in proposed research work. ESVM technique is used to perform malware detection based on intrusion detection. True positive rate is enhanced by this. As shown in that figure, around 96% of precision value is produced by proposed ESVM – IPFCM technique, which is a greater one, while ESVM -KMC producing 94% of precision value, O-SVM producing 91% of precision value, HH-SVM producing 89% of precision value and AE producing 86.4% of precision value.

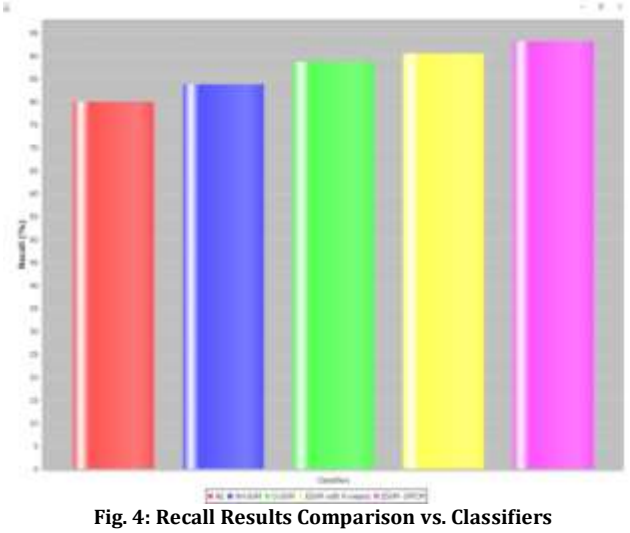

Fig. 4: Recall Results Comparison vs. Classifiers

Recall metric comparison of available EVSM-KMC, O-SVM, HH-SVM, AE and proposed ESVM – IPFCM techniques are shown in fig. 4. In this graph, various techniques are represented in x-axis and in y-axis, recall metric is represented. As shown in that figure, around 92% of recall value is produced by proposed ESVM – IPFCM technique, which is a greater one, while ESVM -KMC producing 90% of recall value, O-SVM producing 89% of recall value, HH-SVM producing 84% of recall value and AE producing 80% of recall value.

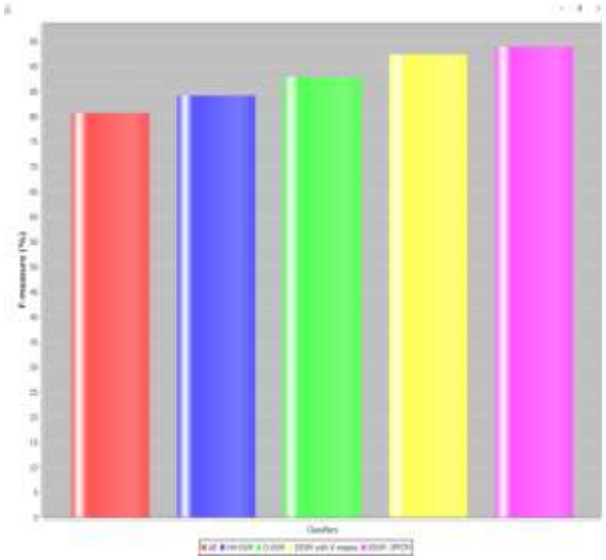

Fig. 5: F-Measure Results Comparison vs. Classifiers

F-measure metric comparison of available EVSM-KMC, O-SVM, HH-SVM, AE and proposed ESVM – IPFCM techniques are shown in fig. 5. In this graph, various techniques are represented in x-axis and in y-axis, F-measure metric is represented. As shown in that figure, around 94% of F-measure value is produced by proposed ESVM – IPFCM technique, which is a greater one, while ESVM -KMC producing 93% of F-measure value, O-SVM producing 88% of F-measure value, HH-SVM producing 84% of F-measure value and AE producing 81.5% of F-measure value.

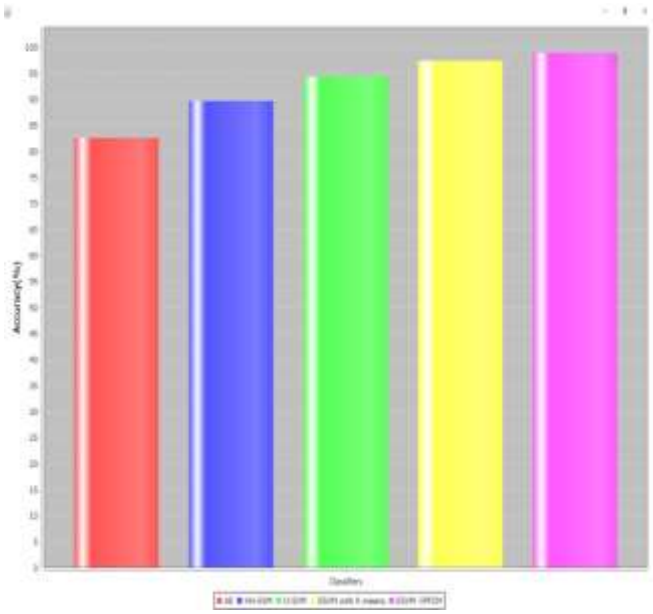

Fig. 6: Accuracy Results Comparison vs. Classifiers

Accuracymetric comparison of available EVSM-KMC, O-SVM, HH-SVM, AE and proposed ESVM – IPFCM techniques are shown in fig. 6. In this graph, various techniques are represented in x-axis and in y-axis, Accuracy metric is represented. For enhancing accuracy and reducing complexity, a multi objective optimization problem is designed as ESVM configuration process. For multi-objective optimization, implemented an improved artificial bee colony optimization algorithm. As shown in that figure, around 98.2% of accuracy value is produced by proposed ESVM – IPFCM technique, which is a greater one, while ESVM -KMC producing 97.3% of accuracy value, O-SVM producing 94.5% of accuracy value, HH-SVM producing 90% of accuracy value and AE producing 83% of accuracy value.

CONCLUSION AND FUTURE WORK

There is rapid change in cyber-attacks due to internet development and the situation of cyber security is not optimistic. For big data cyber security, a new framework is introduced in this work. In which, detection of malicious activity is done using an Ensemble Support Vector Machine (ESVM) framework. Intuitionistic fuzzy C-means clustering is used for selecting features initially in this work. ESVM technique is for performing malware and intrusion detection based on selected features.

For solving multi objective function and optimizing parameters, ahyper metaheuristics algorithm framework is introduced in this work. Improved Artificial Bee Colony (IABC) optimization algorithm is used for parameter optimization. With respect to f-measure, precision, accuracy, recall, better performance is exhibited by proposed work as shown in results of experimentation. Update of network information is happening very fast, which brings training of ML and DL model and there is a need to retrain the model quickly as well as in long-term manner. So, in future, lifelong learning and incremental learning can be focused.

REFERENCES

1. Buczak, A.L. and Guven, E., 2015. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials,* 18(2), pp.1153-1176.
2. Janusz, A., Kałuza, D., Chądzyńska-Krasowska, A., Konarski, B., Holland, J. and Ślęzak, D., 2019, December. IEEE BigData 2019 Cup: Suspicious network event recognition. *In 2019 IEEE International Conference on Big Data (Big Data)* (pp. 5881-5887). IEEE.
3. Wu, J., Ota, K., Dong, M., Li, J. and Wang, H., 2016. Big data analysis-based security situational awareness for smart grid. *IEEE Transactions on Big Data*, 4(3), pp.408-417.
4. Yavanoglu, O. and Aydos, M., 2017, December. A review on cyber security datasets for machine learning algorithms. *In 2017 IEEE International Conference on Big Data (Big Data)* (pp. 2186-2193). IEEE.
5. Mahmood, T. and Afzal, U., 2013, December. Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. *In 2013 2nd national conference on Information assurance (ncia)* (pp. 129-134). IEEE.
6. Terzi, D.S., Terzi, R. and Sagiroglu, S., 2017, October. Big data analytics for network anomaly detection from netflow data. *In 2017 International Conference on Computer Science and Engineering (UBMK*) (pp. 592-597). IEEE.
7. Xu, L., Jiang, C., Wang, J., Yuan, J. and Ren, Y., 2014. Information security in big data: privacy and data mining. *Ieee Access,* 2, pp.1149-1176.
8. Watkins, L., Beck, S., Zook, J., Buczak, A., Chavis, J., Robinson, W.H., Morales, J.A. and Mishra, S., 2017, January. Using semi-supervised machine learning to address the Big Data problem in DNS networks. *In 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 1-6). IEEE.

9. Sabar, N.R., Yi, X. and Song, A., 2018. A bi-objective hyper-heuristic support vector machines for big data cyber-security. *IEEE Access,* 6, pp.10421-10431.
10. Arora, D., Li, K.F. and Loffler, A., 2016, March. Big data analytics for classification of network enabled devices. *In 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)* (pp. 708-713). IEEE.
11. Zhang, Q., Yang, L.T., Chen, Z., Li, P. and Deen, M.J., 2017. Privacy-preserving double-projection deep computation model with crowdsourcing on cloud for big data feature learning. *IEEE Internet of Things Journal, 5*(4), pp.2896-2903.
12. Yin, C., Zhu, Y., Fei, J. and He, X., 2017. A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access, 5*, pp.21954-21961.
13. Jabbar, M.A., Aluvalu, R. and Reddy, S.S.S., 2017, December. Intrusion Detection System Using Bayesian Network and Feature Subset Selection. *In 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-5). IEEE.
14. Xie, Y., Feng, D., Hu, Y., Li, Y., Sample, S. and Long, D., 2018. Pagoda: A hybrid approach to enable efficient real-time provenance based intrusion detection in big data environments. *IEEE Transactions on Dependable and Secure Computing.*
15. Peng, K., Leung, V.C. and Huang, Q., 2018. Clustering approach based on mini batch kmeans for intrusion detection system over big data. *IEEE Access,* 6, pp.11897-11906.
16. Hassan, M.M., Gumaei, A., Alsanad, A., Alrubaian, M. and Fortino, G., 2020. A hybrid deep learning model for efficient intrusion detection in big data environment. *Information Sciences*, 513, pp.386-396.
17. Wheelus, C., Bou-Harb, E. and Zhu, X., 2016, November. Towards a big data architecture for facilitating cyber threat intelligence. *In 2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)* (pp. 1-5). IEEE.
18. Lei, T., Jia, X., Zhang, Y., He, L., Meng, H. and Nandi, A.K., 2018. Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering. *IEEE Transactions on Fuzzy Systems*, 26(5), pp.3027-3041.
19. Wang, C., Pedrycz, W., Yang, J., Zhou, M. and Li, Z., 2019. Wavelet Frame-Based Fuzzy C-Means Clustering for Segmenting Images on Graphs. *IEEE transactions on cybernetics.*
20. Lei, T., Jia, X., Zhang, Y., Liu, S., Meng, H. and Nandi, A.K., 2018. Superpixel-based fast fuzzy C-means clustering for color image segmentation. *IEEE Transactions on Fuzzy Systems,* 27(9), pp.1753-1766.
21. Jiang, Z., Li, T., Min, W., Qi, Z. and Rao, Y., 2017. Fuzzy c-means clustering based on weights and gene expression programming. *Pattern Recognition Letters,* 90, pp.1-7.
22. Liu, P., Wang, Z., Song, Q., Xu, Y. and Cheng, M., 2019. Optimized SVM and Remedial Control Strategy for Cascaded Current-Source-Converters Based Dual Three-phase PMSM Drives System. *IEEE Transactions on Power Electronics.*
23. Yu, H., Chen, B., Yao, W. and Lu, Z., 2017. Hybrid seven-level converter based on T-type converter and H-bridge cascaded under SPWM and SVM. *IEEE Transactions on Power Electronics,* 33(1), pp.689-702.
24. Mukhopadhyay, S.C. and Lay-Ekuakille, A., 2019. Advances in Biomedical Sensing, Measurements, Instrumentation and Systems ABC.
25. Khardenvis, M.D., Tembhare, S.B. and Pande, V.N., 2018. Artificial Bee Colony (ABC) Algorithm based Transmission Expansion Planning with Security Constraints. *Power Research,* 14(1), pp.27-36.
26. Mylavathi and Srinivasan, "A Meta-Heuristic Online Support Vector Machines for Big Data Cyber-Security", *Jour of Adv Research in Dynamical & Control Systems,* Vol. 11, 01-Special Issue, 2019
27. Mylavathi, G.A., A Hyper Meta-Heuristic Cascaded Support Vector Machines for Big Data Cyber-Security.