FEATURE SELECTION USING BINARY BAT ALGORITHM

Kayathri Devi D[1], Kanniga Devi P, Renuga Devi S, Thilagavathi R
[1]*Assistant Professor, Dept. of Information Technology, Kamaraj College of Engineering & Technology, K. Vellakulam, Madurai, India.*

[2,3,4]*UG Scholar, Dept. of Information Technology, Kamaraj College of Engineering & Technology, K. Vellakulam, Madurai, India.*

Abstract-Feature selection aims to find the most important information from a given set of features. Intrusion Detection System(IDS) have become an essential part of every security framework. It deal with large amount of data, one of the crucial tasks of IDS is to keep the best quality of feature that represent the whole data and remove the redundant and irrelevant features. The proposed model uses the BAT algorithm as a search strategy to obtain the optimal subset of features along with the decision tree classifier as a judgement on the selected features that are produced by the BAT algorithm. The wrapper approach combines the power of exploration of the bats together with speed of the decision tree classier to find the set of features that maximizes the accuracy in a validating set.

Keyword-Feature Selection, Bat Algorithm, Decision-tree classifier

I.INTRODUCION

Internet is a global public network. With the growth of the internet and its potential, there has been subsequent change in business models of an organization across the world. More and more people are getting connected to the internet every day to take the advantage of the new business model popularly known as E-Business. Inter network connectivity has therefore become very critical aspects of today's E-business. There are two sides of business on the internet. On one side, the internet brings in tremendous potential to business in terms of reaching the end users. At the same time it also brings in lot of risk to the business. There are both harmless and harmful users on the internet. While an organization makes its information system available to harmless internet users, at the same time the information is available to the malicious users as well. Malicious users or hackers can get access to an organization's internal system in various reasons.

» Software bugs called vulnerabilities

» Lapse in administration

» Leaving systems to default configuration

The malicious users use different techniques like Password cracking, sniffing unencrypted or clear text traffic etc to exploit the system vulnerabilities mentioned above and compromise critical systems. Therefore, there needs to be some kind of security to the organization's private resources from the internet as well as from inside users as survey says that eighty percent of the attacks happen from inside users for the very fact that they know the systems much more than an outsider knows and access to information is easier for an insider.

II. RELATED WORKS

Muhammad Saidu et al. [2016] proposed a Subset Feature Elimination Mechanism for Intrusion Detection System. In a large dataset, not all features contribute to represent the traffic, therefore reducing and selecting a number of adequate features may improve the speed and accuracy of the intrusion detection system. This approach is applied on NSL-KDD dataset which is an improved version of the previous KDD 1999 dataset. Using this, approach relevant features were identified inside the dataset and the accuracy was improved. The feature selection method proposed in this paper had achieved a high result in term of accuracy and features were identified based on information gain and ranking technique.

Dr. A. Malathi et al. [2013] proposed a Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. This paper focus on detailed study on NSL-KDD dataset that contains only selected record. This selected dataset provides a good

analysis on various machine learning techniques for intrusion detection. The statistical analysis showed that there are important issues in the dataset which highly affects the performance of the systems, and results in a very poor estimation of anomaly detection approaches. Using all the 41 features in the dataset to evaluate the intrusive patterns may leads to time consuming and it also reduce performance degradation of the system.

L. Dhanabal et al. [2015] gave a Study article on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms. In this paper the NSL-KDD dataset is analysed and used to study the effectiveness of the various classification algorithms in detecting the anomalies in the network traffic patterns. This paper uses the NSL-KDD dataset to reveal the most vulnerable protocol that is frequently used intruders to launch network based intrusions. In future, it is possibility of employing proposed to conduct an exploration on the optimizing *techniques to develop an intrusion detection model having a better accuracy rate.*

D. Asir Antony Gnana Singh et al. [2016] proposed a Feature Selection Methods for High-Dimensional Data. The feature selection methods that are categorized based on how the features are combined in the selection process namely feature subset-based and feature ranking-based and based on how the supervised learning algorithm used namely wrapper, embedded, hybrid, and filter. It is observed that the feature ranking-based methods are better than the subset-based methods in terms of memory space and computational complexity and the ranking-based methods do not reduce the redundancy.

R.Y.M. Nakamura et al. [2013] proposed a Binary Bat Algorithm for Feature Selection. In this paper we discuss a new nature-inspired feature selection technique based on the bats behaviour, which has never been applied to this context so far. The wrapper approach combines the power of exploration of bats together with the speed of the Optimum -Path Forest classifier to find the set of features. This paper experiments against several meta-heuristic algorithms to show the robustness of the proposed technique and also the good generalization capability the bat-inspired technique.

generalization capability the bat-inspired technique.

Asri Ngadi et al. [2013] proposed a Adaptive Intrusion Detection Model based on Machine Learning Techniques. This paper focus on hybrid machine learning model based on combining the supervised and unsupervised classification techniques. Clustering approach based on combining the K-means,

obtain the normal patterns of a user's activity the technique is used as the first component for pre-classification to improve attack detection. The hybrid classifier approaches provide efficient techniques for anomaly based intrusion detection. This model is now at the infant stage of development.

Maryam Zaffar et al. [2017] proposed A Study of Feature Selection Algorithms for Predicting Students Academic Performance Feature Selection (FS) algorithms remove irrelevant data from the educational dataset and hence increases the performance of classifiers used in EDM techniques. This paper present an analysis of the performance of feature selection algorithms on student data set. The obtained results of the different FS algorithms and classifiers will also help the new researchers in finding the best combinations of FS algorithms and classifiers. Feature Selection is very dynamic and productive field and research area of machine learning and data mining. The main goal of feature selection is to choose a subset by eliminating non-productive data

Junfang Wu et al. [2017] proposed a Feature Selection Based on Features Unit. Feature Selection is to select certain quantity of important features from large number of original features. In this paper, a new feature selection algorithm based on features unit (FU) is presented. The algorithm uses entropy of information to determine whether a feature should integrate with other features on the basis of its relevance to the class. The method divides candidate features into features units according to the relevance of integration of candidate features to the class. Then sort all of the features units in order of their significances to classification.

Kalpana Jain et al. [2017] proposed a Feature Selection Algorithm for Improving the Performance of Classification. This paper gives overview of feature selection Algorithm which searches the feature space using the idea of evolutionary computation, in order to find the optimal feature subset. This paper summarizes various available Feature Selection Methods based. There is still lots of work is about to develop to handle the Multidimensional Data Set. Feature selection is considered one of the most crucial pre–processing steps of machine learning(ML). The feature selection is fairly significant because with the same training data it may perform better with different feature subsets.

Tahira Khorram Khorram et al. [2018] proposed Feature Selection in Network Intrusion Detection Using Metaheuristic Algorithms. In this paper, we aim to use these three metaheuristic algorithms for feature selection, KNN, and SVM as an evaluator for the selecting the right features by met heuristic algorithms. Firewalls, Intrusion Prevention System (IPS), and Intrusion Detection System (IDS) are the most widely used appliances. In this study, our focus is on IDS. Intrusions are a set of actions that try to overrule the security aspect of a system and violate the confidentiality, integrity, and availability of that computer network [3]. Intruders always try

to find a vulnerability in the system to launch an attack; it is intrusion detection system that monitors and analyses all events happening on the computer system, identifies intrusive activities and searches for a sign of security problems.

III.SYSTEM METHODOLOGY

A. Feature selection

Feature selection is a process of removing the irrelevant and redundant features from dataset in order to improve the performance of the Bat Algorithm in terms of accuracy and time to build the model. The process of feature selection is classified into two categories namely feature subset selection and feature ranking methods based on how features are combined for evaluation. The feature subset selection approach generates the possible number of combinations of the features subsets using any one of the searching strategies such as a greedy forward selection, greedy backward elimination , etc. to evaluate the individual feature subset with a feature selection metric such as correlation, consistency, etc. The process of feature selection is,

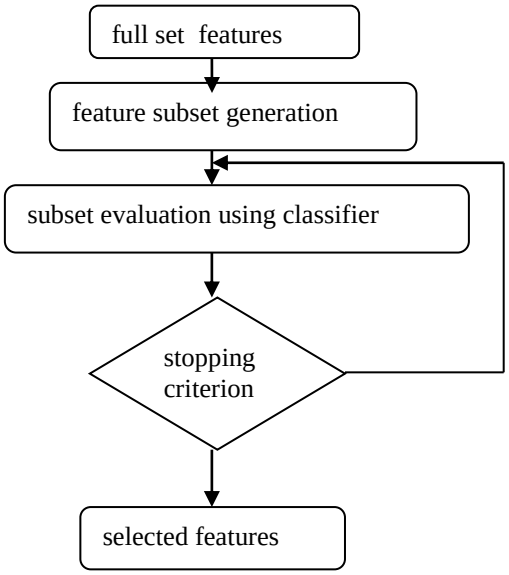


fig 1:flow diagram for feature selection

B. Bat Algorithm

Bats are fascinating animals and their advanced capabilities of echolocation have attracted attention of researchers from different fields. Echolocation works as a type of sonar; bats, mainly micro-bats, emit a loud and short pulse of sound, wait it hits into an object and, after a fraction of time, the echo returns back to their ears. Thus, bats can compute how far they are from an object. In addition, amazing orientation mechanism makes bats being able to distinguish the difference between an obstacle and a prey/food , allowing them to hunt even in complete darkness.

Based on the behavior of the bats, has developed a new and interesting meta-heuristic optimization technique called Bat Algorithm. Such technique has been developed to behave as a band of bats tracking prey/food using their capability of echolocation. In order to model this algorithm, has idealized some rules, as
Follows:

1) All bats are echolocation to sense distance, and they also “know” the
2) difference between food/prey and back-ground barriers in some magical way;
3) A bat βi fly randomly with velocity vi at position xi with a fixed frequency fmin, varying wavelength ƛ and loudness A0 to search for prey.
4) Although the loudness can vary in many ways, space assume that the loudness varies from a large A0 to a minimum constant value Amin.

Algorithm:

Objective function (x), $x = (x1, ..., xn)$.
Initialize the bat population xi and vi, $i = 1, 2, ...,$.
Define pulse frequency fi at xi, $\forall i = 1, 2, \ldots, m$.
Initialize pulse rates ri and the loudness Ai, $i = 1, 2, \ldots, m$
.
1. While $t < T$
2. For each bat bi, do
3. Generate new solutions through Equations (1),
4. (2) and (3).
5. If $rand > ri$, then
6. Select a solution among the best solutions.
7. Generate a local solution around the
8. best solution.
9. If $rand < Ai$ and f(xi) < (x^), then
10. Accept the new solutions.
11. Increase ri and reduce Ai.
12. Rank the bats and find the current best x^.

Firstly, the initial position xi, velocity vi, and frequency fi are initialized for each bat bi. For each time step t, being T the maximum number of iterations, the movement of the virtual bats is given by updating their velocity and position using equations 1,2 and 3, as follows

$$f_i = f_{min} + (f_{min} - f_{max})\,\beta, \quad (1)$$

$$v_i^j(t) = v_i^j(t-1) + [x^{\wedge j} - x_i^j(t-1)]f_i, \quad (2)$$

$$x_i^j(t) = x_i^j(t-1) + v_i^j(t), \quad (3)$$

Where β denotes a randomly generated number within the interval [0, 1]. Recall that x_i^j (t) denotes the value of decision variable j for bat I at time step t. The result of f_i (Equation 1) is used to control the pace and range of the movement of the

bats. The variable $x^{\wedge j}$ represents the current global best location (solution) for decision variable j, which is achieved comparing all the solutions provided by the m bats.

In order to improve the variability of the possible solutions, has proposed to employ random walks. Primarily, one solution is selected among the current best solutions, and then the walk is applied in order to generate a new solution for each bat that accepts the condition.

C. Bats for Feature selection

1. Frequency

Frequency in the proposed algorithm is represented as a real number as defined . The choice of minimum and maximum frequency depends on the application domain, where is a random number range between 0 and 1. Frequency also affects the velocity as,

$$fi = fmin + (fmin - fmax)\ \beta$$

2. Velocity

The velocity of each bat is represented as a positive integer number. Velocity suggests the number of bat attributes that should change at a certain moment of time. The bats communicate with each other through the global best solution and move towards the global best position (solution). The following equation shows the formula for velocity:

$$v_i^j(t) = v_i^j(t\text{-}1) + [x^{\wedge j} - x_i^j(t\text{-}1)|f_i$$

where $[x^{\wedge j} - x_i^j(t\text{-}1)]$ refers to the difference between the length of global best bat and the length of the ith bat. When the difference is positive, this means that the global best bat has more features than those of the ith bat. When the result is summed with the previous velocity, it will accelerate the ith bat towards the global best bat. If the difference is negative, this means that the ith bat has more features than those of the global best bat. Therefore, when the output is summed with the previous velocity, it will decrease the velocity of ith bat and help to attract it closer to global best bat.

3. Position Adjustment

In the proposed algorithm, each bat position is formulated as a binary string of length N, where N is the total number of features. Each feature is represented by bit, where "1" means that the corresponding feature is selected and the "0" means that it is not selected. The positions are categorized into two groups according to the bit difference between the th bat and the global best bat in order to align exploitation and exploration during searching.

The bat's position is adjusted depending on one of the following conditions. In the case where the velocity of ith bat is lower or equal to the number of different bits, ith bat will copy some features from global best bat, thus moving towards global best bat, while still exploring new search space. In the case where the velocities of ith bat are higher than the velocity of global best bat, then the ith bat will import all features from the global best bat to be the same as the global best bat with a few different bits to facilitate further exploitation. The following equation shows the position adjustment, where x is bat position, and v is the velocity of the ith bat at time t:

$$x_i^j(t) = x_i^j(t\text{-}1) + v_i^j(t)$$

4. Loudness

Loudness A_i in the proposed algorithm is represented as the change in number of features at certain time during local search around the global best bat, as well as local search around the ith bat. The formula for loudness, where A^t_i the average loudness of all the bats at certain iteration . The value for sound loudness (A) ranges between the maximum loudness and minimum loudness. Consider the following:

$$X_{new} = x_{old} + \epsilon A^t$$

Generally, the loudness value will decrease when the bat starts approaching the best solution. The following equation shows that the amount of decrease is determined by α:

$$A_i^{t+1} = \alpha A_i^t$$

The value for sound loudness also plays an important role in obtaining good quality solutions within a reasonable amount of time. The choice of the maximum and minimum loudness depends on the domain of application and also the size of the dataset.

5. Pulse Rate

Pulse rate r_i has the role to decide whether a local search around the global best bat solution should be skipped or otherwise. Higher pulse rate will reduce the probability of conducting a local search around the global best and vice versa. Therefore, when the bat approaches the best solution, pulse rate value will increase and subsequently reduce the chances to conduct a local search around the global best. The amount of increase is determined by γ as defined in the following:

$$r_i^{t+1} = r_i^0[1 - exp(-\gamma t)]$$

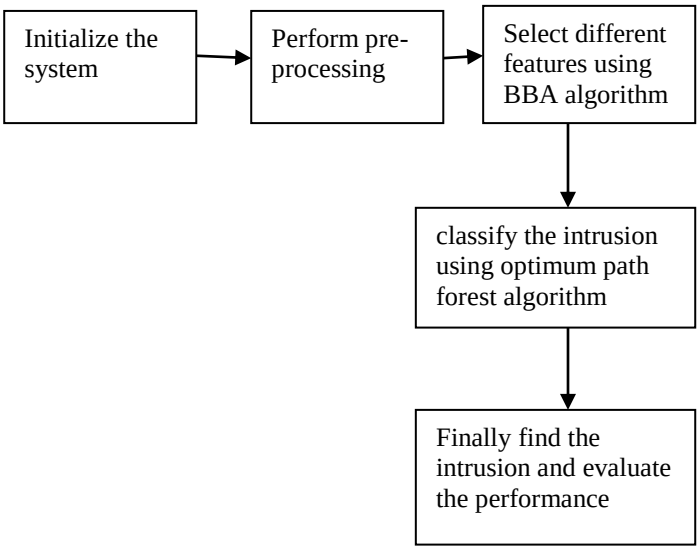
6.Fitness Function

Each candidate solution is using a fitness function defined the below formula, where $p(y_j|x)$ is the classification accuracy, TF is the total number of all features, and SF is the number of selected features.

δ and ϕ are two parameters corresponding to the weight of classification accuracy and subset length, where $\delta \epsilon [0,1]$ and ϕ=1- δ. we can see that the importance of classification accuracy and subset size is weighted differently. Generally, classification accuracy is given more weight than the size of the subset. In this experiment, the two parameters have been set as follows δ=0.9, ϕ=0.1 Consider the following:

$$Sol_{A=} \delta.P(Y_j|X)+ \phi.TF-SF/TF.$$

IV. PROPOSED SYSTEM



In the proposed system, the features are selected by implementing bat algorithm. To detect unknown attacks and to improve the classification accuracy particular selected features will be imported to the SVM classification algorithms. Even in high dimensional noisy dataset SVM give the good generalization and local minima avoidance and give the good accuracy.

Feature selection is a process of removing the irrelevant and redundant features from a dataset in order to improve the performance of the accuracy. The process of feature selection is classified into two categories namely feature subset selection and feature ranking methods based on how the features can combined for evaluation. The feature subset selection approach generates the possible number of the combinations of the feature subsets using any one of the searching strategies such as greedy forward selection and backward selection elimination. evaluate the individual feature subset with a feature selection metric such as correlation, consistency.

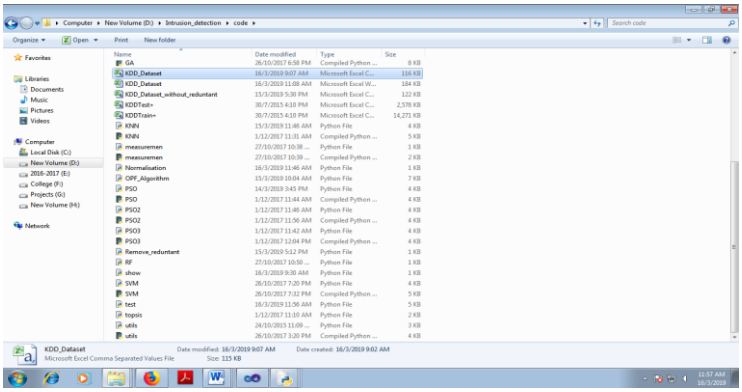
In feature selection can be developed for high –dimensional data using the filter approach with ranking method for selecting the significant features from the high-dimensional space.

As the feature selection can be seen as optimization problem, we have proposed a bat-inspired algorithm for this task. A binary version of the well-known continuous-valued Bat Algorithm was derived in order to position the bats in binary coordinates along the corners of the search space, which represents a string of bits that encodes whether a feature will be selected or not.

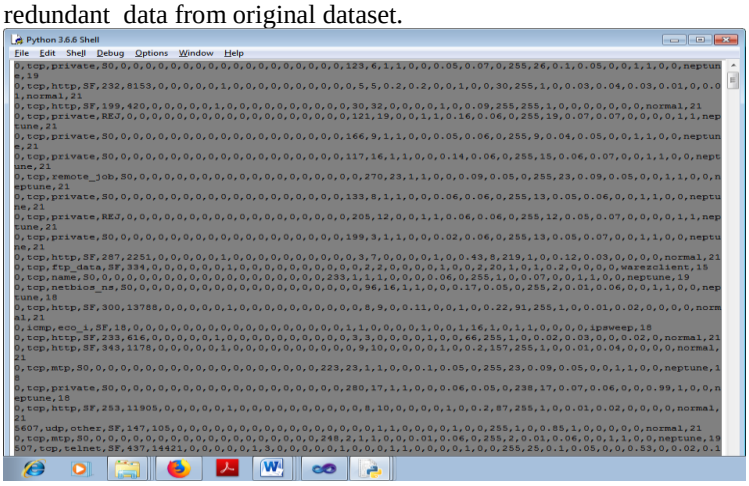
We conducted experiments against several meta-heuristic algorithms to show the robustness of the proposed technique and also the good generalization capability of the bat-inspired technique. we have employed five public datasets to accomplish this task. The proposed algorithm has outperformed the compared techniques in 3 out of 5 dataset, being the second best in the remaining two datasets.

V. EXPERIMENTAL RESULTS AND DISCUSSION

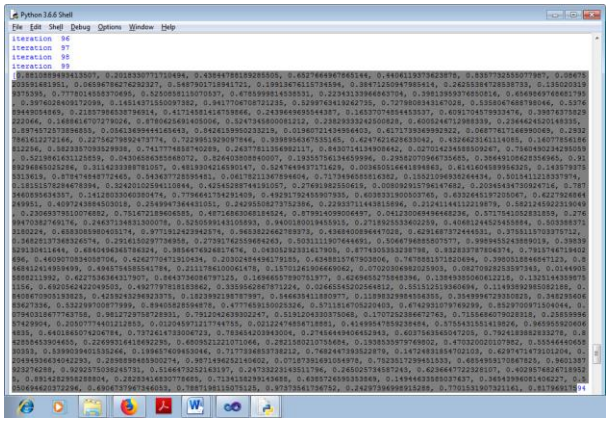
we collect the dataset "NSL KDD_Dataset"

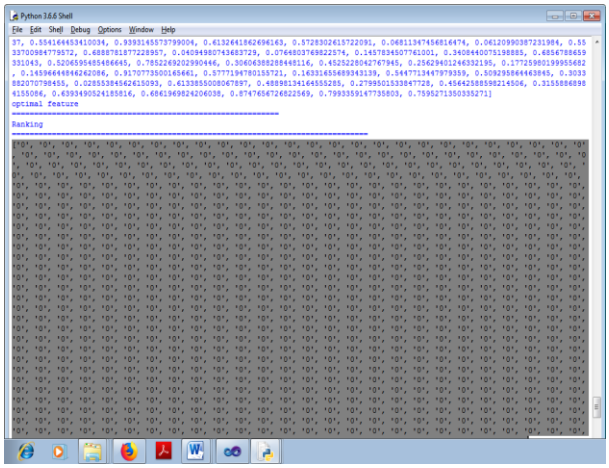


At first in pre-processing we have to normalize the input dataset as well as extract the redundant data from original dataset.



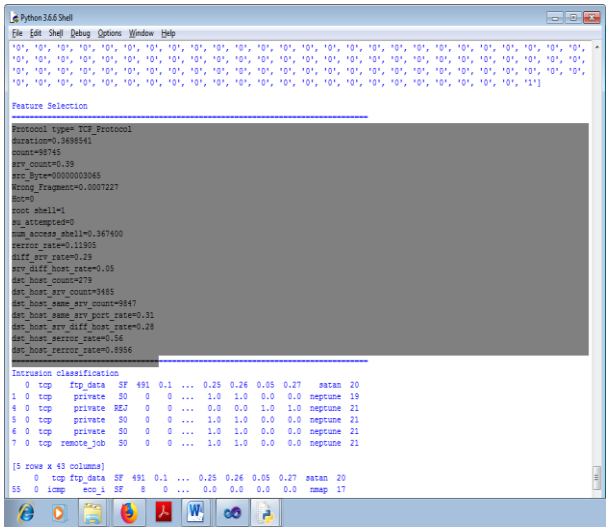
implement feature extraction and ranking each and individual features

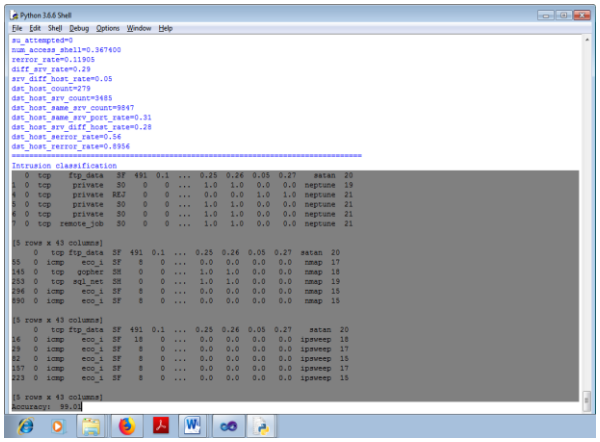




After that implement feature selection using Binary Bat Algorithm ,we select only 20 features

.





VI .DISCUIONS CONCLUSION

Likely many metaheuristic algorithms, bat algorithm has the advantage of simplicity and flexibility. BA is easy to implement, and such a simple algorithm can be very flexible to solve a wide range of problems as we have seen in the above review.

Why Bat Algorithm is Efficient

A natural question is: why bat algorithm is so efficient? There are many reasons for the success of bat-based algorithms. By analysing the key features and updating equations, we can summarize the following three key points/features:

Frequency tuning: BA uses echolocation and frequency tuning to solve problems. Though echolocation is not directly used to mimic the true function in reality, frequency variations are used. This capability can provide some functionality that may be similar to the key feature used in particle swarm optimization and harmony search. Therefore, BA possess the advantages of other swarm-intelligence-based algorithms.

Automatic zooming: BA has a distinct advantage over other metaheuristic algorithms. That is, BA has a capability of automatically zooming into a region where promising solutions have been found. This zooming is accompanied by the automatic switch from explorative moves to local intensive exploitation. As a result, BA has a quick convergence rate, at least at early stages of the iterations, compared with other algorithms.

Parameter control: Many metaheuristic algorithms used fixed parameters by using some, pre-tuned algorithm-dependent parameters. In contrast, BA uses parameter control, which can vary the values of

parameters (A and r) as the iterations proceed. This provides a way to automatically switch from exploration to exploitation when the optimal solution is approaching. This gives another advantages of BA over other metaheuristic algorithms. In addition, preliminary theoretical analysis by Huang et al.(2013) suggested that BA has guaranteed global convergence properties under the right condition, and BA can also solve large-scale problems effectively.

VII. REFERENCES

1. Herve Nkiama, Syed Zainudeen Mohd Said, and Muhammad Saidu, "A Subset Feature Elimination Mechanism for Intrusion Detection System" International Journal of Advanced Computer Science and Applications, vol. 7, No. 4, 2016
2. S. Revathi, and Dr. A. Malathi "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection" International Journal of Engineering Research and Technology (IJERT) vol.2 Issue 12 Dec-2013
3. L. Dhanabal, and Dr. S.P. Shantharajah, "A study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms" International Journal of Advanced research in Computer and Communication Engineering vol. 4, Issue 6, June 2015
4. D. Asir Antony Gnana Singh, S. Appavu alias Balamurugan, E. Jebamalar Leavline, "Literature Review on Feature Selection Methods for High-Dimensional Data" International Journal of Computer Applications vol. 136- No.1, February 2016.
5. R.Y.M. Nakamura, L.A.M Pereira, K.A. costa, D. Rodrigues, J.P. Papa and X.S. Yang "A Binary Bat Algorithm for Feature Selection" SIBGRAPHI Conference on Graphics, Patterns and Images, 2012
6. Salima Omar, Asri Ngadi, Hamid H. Jebur "An Adaptive Intrusion Detection Model based on Machine Learning techniques" International Journal of Computer Applications vol.70-No.7, May 2013
7. Maryam Zaffar, K. S Savita, Manzoor Ahmed Hashmani, Syed Sajjad Hussain Rizvi "A Study of Feature Selection Algorithms for Predicting Students Academic Performance" International Journal of Advanced Computer Science and Applications, Vol. 9, No. 5, 2018
8. Mary Walowe Mwadulo "A Review on Feature Selection Methods For Classification" International Journal of Computer Applications Technology and Research Vol. 5, 2016
9. Ms. Shweta Srivastava , Ms. Nikita Joshi ,and Ms. Madhvi Gaur "A Review Paper on Feature Selection Methodologies and Their Applications" International Journal of Engineering Research and Development" Vol. 7, 2013
10. Kalpana Jain "A Survey on Feature Selection Techniques" International Journal of Innovations in Engineering Research and Technology Vol.4, May-2017
11. E. Emary, M. Hossam Zawbaa, and Aboul Ella Hassanien "Binary grey wolf optimization approaches for feature selection" Neurocomputing 172 (2016) 371–381
12. Tahira Khorram Khorram and Nurdan Akhan Baykan "Feature Selection in Network Intrusion Detection Using Metaheuristic Algorithms" International Journal of Advance Research, Ideas and Innovations in Technology, 2018
13. Jun Wang, Jin-Mao Wei, Zhenglu Yang, and Shu-Qin Wang "Feature Selection by maximizing independent classification information" IEEE Transactions on Knowledge and Data Engineering, 2016
14. Junfang Wu and Chao Li " Feature Selection Based on Features Unit" International Conference on Information Science and Control Engineering, 2017
15. Kajal Naidu, Aparna Dhenge and Kapil Wankhade "Feature Selection Algorithm for Improving the Performance of Classification" International Conference on Communication Systems and Network Technologies, 2014
16. Kapil Wankhade, Dhiraj Rane and Ravindra Thool "A New Feature Selection Algorithm for Stream Data Classification" International Conference on Advances in Computing, Communications and Informatics, 2013
17. Anukrishna P.R and Dr. Vince Paul "A Review on feature selection for high dimensional data" International Conference on Inventive Systems and Control,2017