Feature selection approach for intrusion detection system based on pollination algorithm

Shruti Dubb*, Yamini Sood**
*Department of Computer Science, Sri Sai University, Palampur
**Department of Computer Science, Sri Sai University, Palampur

ABSTRACT
An intrusion detection system refers to the systems which detect the security breaches in a network. The main goal of IDS is to detect the attacks and raise alarms if an attack is found. These IDS has limitation of raising false alarms and this limitation depends on improper classification of attacks due to high dimensionality of data on a network. Improper classification is due to improper feature selection. To overcome this limitation, optimization of feature selection technique is proposed in this paper. Pollination based optimization technique is proposed which eventually improves feature selection in high dimensional data of IDS and achieve reduction in false alarms.

Keywords - Classification, False alarms, High dimensionality, IDS, Pollination.

I. INTRODUCTION
Intrusion, in simple words, is an illegal act of entering in a network or computer system in an unauthorized way. And intrusion detection system is that system which is developed to detect unauthorized use of network or any malicious activities on computer system. These intrusion detection systems are suffering from the problem of false detection of attacks. False detection refers to classifying normal traffic as malicious (false positives) or classifying malicious traffic as normal (false negatives) [1]. Various approaches are being used to overcome the false alarm problem of IDS such as data mining, event correlation and classification of alerts. Data mining is a field which deals with hundreds of data which refer to as high dimensional data. And the main reason for false alarms in IDS is high dimensional data, thus, it can be solved by proper data mining technique. Data mining is seen as a necessary tool by stylish business to rework knowledge into business intelligence giving an informational advantage. It's presently used in an exceedingly smart range of identification practices, like promoting, surveillance, fraud detection, and scientific discovery. A primary reason for using data mining is to assist within the analysis of collections of observations of behaviour. One basic category of data mining is classification. Classification could be a task of categorizing one thing into predefined categories. Classification is done on the basis of features/ attributes/ behaviour. Feature selection is that the major issue upon which the classification depends, means, if features are properly selected, then the classification are going to be automatically correct and if features don't seem to be elect in correct approach, then classification can offer dangerous leads to data mining. This paper proposes feature selection based on pollination based optimization algorithm. KDD-NSL dataset is employed for implementing the planned approach.

II. LITERATURE SURVEY
1. Feature Selection
Feature selection has been a lively and fruitful field of analysis area in pattern recognition, machine learning, statistics and data mining communities. The main objective of feature selection is to decide on a set of input variables by eliminating options, that are unsuitable or of no prophetic data. A feature selection has verified in each theory and applies to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing quality of learned results. Feature selection in supervised learning incorporates a main goal of finding a feature set that produces higher classification accuracy [3]. Because the spatial property expands, the amount of options increases. Finding AN optimum feature set is tough. At this juncture, it's essential to explain ancient feature selection method, which consists of 4 basic steps, namely, set generation, stopping criterion, and validation. Set generation could be a search method that produces candidate feature subsets for analysis supported a particular search strategy. Every candidate set is evaluated and compared with the previous best one per a particular analysis. If the new set turns to be higher, it replaces best one. This method is continual till a given stopping condition is satisfied. Algorithms for feature selection fall under 2 broad classes specifically wrappers that use the training algorithmic rule itself assess the quality of options and filters that evaluate features.

www.ijaert.org
consistent with heuristics based on general characteristics of the data. Long Sheng Chen and Jhiih Siang Syu in [4] used Feature extraction based on local and global latent semantic indexing to improve classification performance. Work is done in two stages, first for feature extraction and second for optimization of dimension size by SVM, which was not suitable for multidimensional problems. In [5], Dr. Saurabh Mukherjee and Neelam Sharma, applied feature selection by using three different methods namely, correlation feature selection, information gain and gain ratio. Feature vitality based reduction method was used which deletes one feature at one time. The results were compared using NB classifier. Features got reduced to some extent, but it took very large time to select features.

In [6], Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi, Twae-kyung Park suggested a new feature selection method that uses the attributed average of total and each class data. The decision tree classifier evaluated with the NSL-KDD dataset to detect attacks. High accuracy was achieved but time taken was more to detect the attack type. Ayman I. Madbouly, Amr M. Gody, Tamer M. Barakat [7], applied four different learning classifiers separately and compare the results. They used best feature reduction method by gradually adding and deleting the feature and finally obtained the reduced set of 11 features out of 41 features. Ayman I. Madbouly, Tamer M. Barakat [8], again applied the same feature subset selection methodology with different approach of using CFS with seven search methods. They gained 70% reduction in feature subset which was similar to the previous results. Main limitation was the old KDD99 dataset used as it consists of lot of drawbacks.

2. NSL-KDD Dataset Description
A knowledge set with a large quantity of quality data that mimics the real time will only help to train and check an intrusion detection system. The NSL-KDD knowledge set may be a refined version of its forerunner KDD*99 knowledge set. NSL dataset consists of downloadable files that are used as normal files by researchers to figure on IDS enhancements. Redundant records from KDD cup ninety nine dataset were removed that developed NSL dataset. In every record there are a unit forty one attributes describing totally different options and labels. Labels indicate that network is normal or an attack. forty second attribute contains data regarding 5 categories, one traditional category and 4 attack categories. The Four attack categories are: Denial of Service attack, Probe attack, User to Remote attack and Remote to Local attack. [9]

3. Nature Inspired Algorithms
Algorithms having source of inspiration from nature are called as nature inspired. Nature inspired algorithms can be bio-inspired, chemical based, physics based or even they can be music based also. Largest numbers of nature inspired algorithms are biology inspired which are most popular. Some of them are swarm based namely, Ant colony optimization, PSO and Cuckoo search [10]. Seyed Mojtaba Hosseini Bamakan, Behnam Amiric, Mahboubeh Mirzabagherib and Yong Shia in [11] proposed a new intrusion detection approach using PSO based multiple criteria linear programming which results in robust and simple implementation to improve performance of classifier. KDD cup99 dataset was used to evaluate performance. High performance was achieved but PSO with multiclass is better than simple PSO technique. Good algorithmic rules have each combination and variety management so the algorithm can explore the large search area with efficiency, whereas converges relatively quickly when necessary. Sensible algorithms like particle swarm optimisation, differential evolution, cuckoo search and firefly algorithms all have each global search and intensive native search capabilities, which can be partly why they're so efficient. Adel Sabry Eesa, Zeynep Orman and Adnan Mohsin Abdulazeez Brifcani [12], applied Bat algorithm for feature selection which is one of nature inspired algorithms. They used ID3 classifier to check the classification performance. Three performance measures were used to evaluate the proposed approach namely, detection rate, false alarm rate and accuracy rate. BA+ID3 were a good combination in terms of accuracy. BA had a drawback in selecting its neighborhood search.

4. Pollination based optimization (PBO)
The objective of the PBO is that the survival of the fittest and therefore the best copy of plants in terms of numbers in addition as fittest. This can be in reality an optimisation process of plant species. It's estimated that...

www.ijaert.org
there are over 1 / 4 of 1,000,000 varieties of flowering plants in Nature and that regarding eightieth of all plant species are flowering species. Flowering plant has been evolving for over a hundred twenty five million years and flowers became thus important in evolution, we cannot image how the plant world would be without flowers. The most purpose of a flower is ultimately reproduction via fertilization. Fertilization is often related to the transfer of pollen, and such transfer is commonly joined with pollinators like insects, birds, batty and different animals. In fact, some flowers and insects have co-evolved into a really specialised flower-pollinator partnership. As an example, some flowers will solely attract and may solely rely upon a particular species of insects for flourishing fertilization [13].

Pollination will take 2 major forms: abiotic and organic phenomenon. Regarding ninetieth of flowering plants belong to organic phenomenon fertilisation, that is, pollen is transferred by a insect like insects and animals. Regarding 100% of fertilisation takes abiotic kind that doesn't need any pollinators. Wind and diffusion in water facilitate fertilisation of such flowering plants and grass could be a smart Example. fertilisation will be achieved by pollination or cross-pollination. Cross-pollination means that fertilisation will occur from pollen of a flower of a distinct plant, while self-pollination is that the fertilization of 1 flower, like peach flowers, from pollen of the same flower or totally different flowers of a similar plant, which frequently happens once there's no reliable insect out there. Biotic, cross-pollination might occur at long distance, and therefore the pollinators like bees, bats, birds and flies will fly an extended distance, so they'll considered as the global pollina-tion. Additionally, bees and birds might behave as L’evy flight behaviour with jump or fly distance steps adjust an L’evy distribution. Moreover, flower constancy will be used an increment step using the similarity or distinction of 2 flowers. First, the convergence rate is actually exponential as we’ve seen from the convergence comparison within the previous section. Second, the reasons that FPA is efficient are often twofold: they'll escape any native landscape and later explore larger search area.

5. Naive Bayes Classifier
This classifier is based on frequency table and supported the principle of Bayes Theorem. Bayes theorem states that one attribute value doesn't show alternative attribute value. It states that the error could be a result of 3 factors: training data noise, bias, and variance. Training data noise will solely be minimised by selecting sensible training data. The training data should be divided into numerous groups by the machine learning algorithmic rule. Bias is that the error because of groupings within the training data being terribly massive. Variance is that the error due to those groupings being too little [5]. This model provides answer to the queries like if few determined events are given then what's the chance of a specific kind of attack. It can be done by using formula for probability. When the data set is very massive it'll be efficient to use NB models [14].

6. J48 Classifier
J48 is an extension of ID3. The extra features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In alternative algorithms the classification is performed recursively until each single leaf is pure, that's the classification of the data should be as excellent as possible. This algorithmic rule it generates the principles from which specific identity of that data is generated. The target is increasingly generalization of a decision tree till it gains equilibrium of flexibility and accuracy [15].

This classifier has the advantage that it will handle each continuous as well as separate value of attributes or attributes with missing values. It's extremely efficient when handling large datasets.

III. PROPOSED METHODOLOGY
The whole planning of proposed methodology is shown in Fig.1. The paper shows a new optimization technique for feature selection which improves the classification in Intrusion detection systems. In this paper, we propose a new algorithm, named as, pollination based optimization algorithm which is inspired by the pollenation process of flowers. NSL-KDD dataset is used in this implementation. Here is the description of all steps used in methodology.

Step1: Input data
KDD-NSL Dataset is used for the experiment. It is the standardized dataset developed in MIT Lincoln Labs.
This data is standard data which is used by researchers to survey and evaluate IDS. 22,544 instances of NSL dataset are used for testing.

**Step 2: Feature selection through PBO**

Feature selection is optimized using pollination based optimization technique. Dataset is considered as plants. Population is number of instances and Features of dataset are considered as pollinators.

The basic POLINATION algorithm is as follows:
1. Produce an initial population (usually at random generated string)
2. Judge all of the people (apply some function or formula to the individuals)
3. Choose a replacement population from the previous population based on the fitness of the people as given by the evaluation function.
4. Apply some POLINATION operators (mutation & crossover) to members of the population to make new solutions.
5. Measure these recently created people.
6. Repeat steps 3-6 (one generation) till the termination criteria has been satisfied (usually perform for a definite fixed range of generations)

**Step 3: Classification**

After applying PBO based optimization technique on feature selection, the selected features are given to two different classifiers namely, NB classifier and J48 classifier. These classifiers are used because these are different and have their own advantages and better performance than other classifiers as discussed above in Section 5 and Section 6. The results are obtained in terms of Accuracy, time to build model and reduced features.

**IV. RESULTS**

NSL-KDD Dataset is used as an input. The input is given to proposed algorithm, that is, pollination based optimization. Features are selected using optimization technique. Then the selected features are given to two different classifiers namely, NB classifier and J48 classifier. On the basis of features selected and classified by classifiers, precision and recall are obtained which gives the accuracy results. Time taken is also obtained.

Precision: It tells that the features or items which we have selected are correct or not. In other words, it shows that how much our prediction is correct. Precision obtained in case of 20 pollinators is 99%. Fig. 2 below indicates the comparison of precision obtained for the same factors in previous approach that is genetic algorithm and proposed approach which is pollination based optimization. It is the percentage of selected features which are correct. Fig. 2 clearly shows that genetic algorithm has less precision than PBO.

Recall: It tells that all the features which are correct and which are selected also. We obtained 94% of correct features among 42 features in NSL dataset which are selected by pollination based optimization algorithm. Figure 3 below shows recall obtained for PBO as well as previous approach which is GA.

Time to Build: It is time required to build the model. It is calculated in milliseconds. It is observed that feature selection technique with pollination optimization takes very less time to build model. It took 0.25 milliseconds
which is very less as compare to previous approaches which was 0.22 seconds.

![Fig.3 Recall in percentage](image)

**Fig.3 Recall in percentage**

F-ratio: It is a combined measure which gives precision recall trade-off, below figure 4 shows the F-ratio graph obtained for both previous as well as proposed approach. PBO has 89% f-ratio for 20 pollinators which is greater than previous approach.

![Fig.4 F-ratio in % comparison of PBO and GA](image)

**Fig.4 F-ratio in % comparison of PBO and GA**

V. CONCLUSION AND FUTURE SCOPE

In this paper, a new optimization technique is used which is known as pollination based optimization. This technique is used for optimization of feature selection process in NSL-KDD dataset. After features selection, the selected features are given to Naive bayes and j48 classifier one by one and classification results are obtained. From the results it can be concluded that proposed method gives better results than previous approach. After optimizing feature selection it helps classifiers to classify the intrusions accurately.

In this algorithm, size of pollinators is limited, further enhancement in the algorithm can be made by applying more number of pollinators.

REFERENCES


[10] Iztok Fister Jr.1, Xin-She Yang2, Iztok Fister1, Janez Brest1, Du’san Fister, A Brief Review of Nature-Inspired Algorithms for Optimization.


[13] Xin-She Yang, Flower Pollination Algorithm for Global Optimization.
