ISSN: 0067-2904

Modified Artificial immune system as Feature Selection

Jamal H. Assi*, Ahmed T. Sadiq
Computer Science Department, Technology University, Baghdad, Iraq.

Abstract:
Feature selection algorithms play a big role in machine learning applications. There are several feature selection strategies based on metaheuristic algorithms. In this paper a feature selection strategy based on Modified Artificial Immune System (MAIS) has been proposed. The proposed algorithm exploits the advantages of Artificial Immune System AIS to increase the performance and randomization of features. The experimental results based on NSL-KDD dataset, have showed increasing in performance of accuracy compared with other feature selection algorithms (best first search, correlation and information gain).

Keywords: AIS, feature selection, NSL-KDD.

تطوير خوارزمية نظام المناعة الاصطناعي لاستخدامها في اختيار الخصائص

جمال هلال*، احمد طارق
قسم علوم الحاسبات، الجامعة التكنولوجية، بغداد، العراق.

الخلاصة
خوارزميات اختيار الخصائص تلعب دورا كبيرا في تطبيقات تعليم الماكنة . هناك عدة إستراتيجيات في اختيار الخصائص ترتكز على خوارزميات (Metaheuristic). في هذا البحث تم اقتراح إستراتيجية اختيار الخصائص التي تعتمد على انظمة المناعة الاصطناعية المطورة. هذه الخوارزمية المقترحة توضح فوائد استخدام نظام المناعة الاصطناعي لزيادة الكفاءة والعشوائية في الخصائص. النتائج التجريبية التي أعتمدت على قاعدة بيانات (NSL–KDD) تظهر زيادة في دقة الاداء مقارنة مع خوارزميات اختيار الخصائص الاخرى مثل (Best First Search , Correlation and Information Gain).

I. Introduction

Artificial immune system is inspired by the living organism's immune system. This system has the ability to stand by the machine learning. AIS is defined as a data manipulation, classification, and representation [1].

AIS society is active for two decades that produces most productive research from the design of the biological system, to solve artificial or standard problems, to respond to the real-world applications, using the same immunization of immunological algorithms. Therefore, it is useful to reverse and invest in the subscription that this model offers the different areas that can be applied. AIS recommended many implementation, such as, "clustering and classification, anomaly and intrusion detection, optimization, control, bioinformatics, data recovery and web mining, user designing and personalized commendation and image processing" [2]. Feature selection methods choose attributes from original spaces set based on strategy like information gain, correlation, and decision table. etc. Hall and Smith [3] proposed a subset of attributes to be related if these attributes are high connection with class and are not connected with each other in terms of mutual information.
Feature selection offers many advantages some of them are illustrated below: [3]

*Email: jamalabomohammed@yahoo.com

- It diminishes feature dimensionality also support to get better performance of the algorithm.
- It discards unnecessary, extraneous or noisy data.
- It makes efficient data goodness which helps to get better the performance of learning technique.
- It enhances the precision of output model.
- It assists in data grasp to acquire knowledge about the operation that created the data.

The goal of this paper is to modify the AIS algorithm to work as a feature selection. This algorithm increases the power of the performance also it provides high diversity to the hidden features or the features that are not under control of the statistical measures because ways of choosing traditional features may sometimes eliminate important or potentially important features, so we sometimes need to increase searches in features to get the best of them. The modified algorithm helps to select the best features.

In addition to this section the paper is organized as follow. Section2 explain the related work of this paper. Section 3 gives an overview of the artificial immune system. Section 4 explains what is the NSL-KDD dataset and its features and the four types of attack. Section 5 showed the proposed algorithm which is modified artificial immune system for feature selection. Section 6 explain the results that gain from the proposed algorithm. Section 7 is the conclusion. Finally section 8 is the references.

II. Related Work

In the work of [2] five important classification methods with 3 features selection strategies were implemented to classify the network attack using NSL-KDD dataset. These methods are (J48 decision tree, support vector machine) decision table and Bayesian network). Several experiments were implemented to obtain good results using training and testing NSL-KDD within general attack (normal and anomaly) and within 4 attacks (U2R, R2L, Probe, DOS).

The work in [4] surveyed the newest in multi objective optimization with Artificial Immune System algorithms. This had focused on artificial immune systems and discovered that it has some essential characteristics that make it appropriate as a multi objective optimization algorithms. Next to this essential concept, diverse applications had been suggested in this survey. It targeted to supply a comprehensive review of the literature on multi objective optimization algorithms based on the simulation of the immune system.

The work in [5] suggested an enhanced Artificial Immune System based on a model that figures out multi thematic optimization issues. The major thematic of the solution of multi thematic optimization issue was to assist the decree making for finding the appropriate solution as the ultimate output. The inspiration of this algorithm was the biological immune system and clonal selection concept. From the current approach crossover technique is incorporated to classic artificial immune system algorithm be based on the theory of clonal selection. This Algorithm was suggested with actual variables not in binary code. Only non-control person and workable top antibodies were added to memory set. The algorithm was used to fix various real-life engineering multi-objective optimization issues. The fact that an adapted grouping of antibodies can produce 'intelligent' behavior is the main motivation for choosing artificial immune system to evolve algorithms. And hence, this power of computation could be ideal for multi objective optimization and the problems associated with it.

The main advantage of the design in [6] was that it improved universal optimization of the complete system and replaces the need for population control mechanism. To validate the system, it had been tested using different classification techniques based on seven benchmark data sets. The outcomes were found very competitive in comparison with other classifiers.

III. An overview for Artificial Immune System

The topic Artificial Immune System is a vital research field. It has been inspired by the natural immune system which is characterized in being self-adaptive and robust in protecting human body from harmful entities. The AIS approach adopts the biological immune system fundamentals to provide efficient solutions to real-world situations. The idea is to compose intelligent methodologies to create computational algorithms capable for solving complicated engineering problems [5]. The major task of AIS is to maintain healthy systems by detecting foreign patterns which may threaten the system and attack it holistically. The main basic elements that composes of AIS are:

- System component representation such as binary strings, vectors of real numbers, etc.).

- Interaction evaluation components. These sets of mechanisms are used to evaluate individual's interactions with each other and with the environment. Such an environment is normally simulated through an affinity function, which is based on the objective function(s) in the case of optimization problems.
- Adaptation procedure to indicate changes in system behavior over time. For example, they may be changes in the mutation operator's system over time.

The population-based metaheuristics Artificial Immune Systems has been widely used for various optimization and classification tasks [3]. AIS uses clone selection techniques and clonal selection theory to deal with the immune response to antigenic stimulus. The antibodies system can distinguish between antigens that are selected for further proliferate process from those that are not.

The rapidly evolving area of AIS has many applications in engineering sectors. The system has proven its efficiency and its interesting features through wide range of positive characteristics such as Strong mimicking, imitating, robustness, self-adaptability, memory, tolerance, recognition and so on [5]. These features have attracted researchers to adopt various immunological principles for the development of distinctive computational models.

IV. NSL-KDD Dataset

The NSL-KDD data set is an enhanced version of the KDD cup99 data set. The inherent deficiency in the KDD cup99 dataset has been uncovered by various statistical analyses has affected the detection accuracy of many Intrusion Detection System (IDS) modelled by researchers. NSL-KDD consists of fundamental records of the complete KDD dataset. NSL-KDD has the following features:

1- "unnecessary records are removed to permit the classifiers to produce fair results."

2- "Adequate number of records is accessible in train and test data set, which is sensible reasonable and enables to perform tests on the complete set."

3- "The number of selected records from each hard level group is oppositely proportional to the percentage of records in the original KDD data set."

Each one of the records, has 41 different features these features together decide that the label is normal or anomaly [2].

The records in the train data set is (125972) and the test data set is (22544), During a certain intrude, an intruder setup a linkage amidst the origin IP address to a destination IP, and forward data to intrude the base. There are four types of attacks illustrated as follows:

1. Denial of Service Attack (DoS): "is an attack class which exhaustion the victim's resources as a result of that making the victim unable to process the request this would shut down the intended device or to flood it in requests and therefore the authorized users cannot reach the device services. For example: ping to death and syn. flood.
2. User to Root Attack (U2R): "is a type of attack that takes the advantage of authorized users and tries to reach the root of the system from some vulnerability. For example, buffer overflow attack."
3. Remote to Local Attack (R2L): "occurs when an attacker who has the ability to send stream of bits to a device in a network but this attacker doesn't have an account on that device exploits some vulnerability to obtain a local access to that device. For example: password guessing."
4. Probing Attack: "is trying to collect a datum on a Net and detect the system vulnerabilities." These vulnerabilities will take the advantage to intrude the system. For example, Port scanning [2].

V. Proposed Algorithm

The main idea of the proposed algorithm is to modify the AIS so that it tackles the problem of feature selection in such a way that there are features selected by mathematical operations those are not the best. So, another strategy has been chosen to enhance the performance of the algorithm by dividing the features into three parts, the good features, the bad features and in between the good one is the pattern. We exclude the bad and generates a new population from the features that between the good and bad features then apply to those features all AIS operations (clonal selection, affinity, mutation, hypermutation). then add the best features to the good features and test the system again until reaching stop criteria. The algorithm is as follow:

Modified AIS Algorithm for Features Selection

Input: F represent the features for a Dataset D with T as a target;

Output: the selected features X;

Begin

Rank the features F of D depending on scores measures related with T (Mutual

Information or PCA…);
Assign the highest feature scores as a patterns P;
Drop the lowest feature scores;
Generate initial population randomly as a subset from the reminder features A;
Repeat
For each pattern P Do
Compute affinity to A;
Select n highest affinity from P as an affinity maturation;
Clone and mutate to affinity;
Add new mutants to P;
End For
Select highest affinity from P;
Replace M number of random new ones;
Add the best features to the pattern P features;
Evaluate the current features (depends on accuracy performance using Decision Tree learning algorithm);
Until stopping criteria;
End.

The proposed algorithm exploits the advantages of AIS to increase the performance and randomization of features. The idea is that there are some features sometimes play a big role but not chosen in other feature selection strategy, this problem is resolved by the proposed algorithm. The proposed algorithm gives importance to these features by examining whether they are really important or not, by adopting them as important elements of the AIS algorithm.

VI. Experimental Results

By using modified AIS mentioned above for feature selection, the algorithm chooses 10 important features and drop other features (31 features) this reduce the search space and time. The power of available diversity through clonal selection has been used in the proposed algorithm also through the random generation for the clonal selection and mutation to create a diverse population to enhance the access or recognize the target. The table below shows these results for four types of attacks. The table shows the result of the proposed algorithm which selected 10 important features with j48 classifier the accuracy is 84.1509% and time is 0.01 seconds which is better than the best result (Info. Gain feature selection with J48 classifier 80.9693 % and time is 8.78 seconds).

Table1-Experimental Results of NSL-KDD Training and Testing Dataset for 4 Types (Dos, Probe, R2L, and U2R) of Attacks by Using Three Feature Selection Strategy (Best First Search, Correlation and Info. Gain) and Comparing the Accuracy and Time with Proposed Algorithm

Classification Method	Feature Selection Method	No. of Selected Att.	NSL-KDD Accuracy	Time
J48	Best first search	8	72.96 %	7.09 seconds
		11	72.8803 %	5.34 seconds
		18	74.2536 %	10.68 seconds
	Correlation	8	68.3973 %	7.33 seconds
		16	73.6733 %	10.95 seconds
		20	78.1829 %	16.63 seconds
	Info Gain	8	78.0588 %	4.17 seconds
		13	80.9693 %	8.78 seconds
		17	74.1605 %	12.1 seconds
	Modified AIS	10	84.1509%	0.01 seconds
	Best first search	8	72.424 %	228.31 seconds
		11	74.4042 %	181.78 seconds

SVM		18	77.1286 %	72.4 seconds
	Correlation	8	68.003 %	105.79 seconds
		16	71.6134 %	9440.29 seconds
		20	73.4828 %	107.7 seconds
	Info Gain	8	74.6345 %	94.84 seconds
		13	71.7197 %	389.18 seconds
		17	71.5868 %	232.75 seconds
	Modified AIS	10	77.7632%	27.13 seconds
Decision Table	Best first search	8	72.96 %	7.09 seconds
		11	68.5833 %	20.64 seconds
		18	71.4627 %	32.5 seconds
	Correlation	8	64.406 %	8 seconds
		16	68.4903 %	36.15 seconds
		20	71.0729 %	47.21 seconds
	Info Gain	8	71.352 %	17.3 seconds
		13	67.0373 %	25.62 seconds
		17	66.674 %	51.16 seconds
	Modified AIS	10	75.529%	0.11 seconds
Back Propagation	Best first search	8	68.0296 %	183.19 seconds
		11	72.681 %	3497.34 seconds
		18	73.8859 %	8224 seconds
	Correlation	8	68.8624 %	402.68 seconds
		16	69.0883 %	543.78 seconds
		20	72.2646 %	5518.78 seconds
	Info Gain	8	71.0508 %	1.26 seconds
		13	74.6833 %	0.66 seconds
		17	73.7397 %	0.71 seconds
	Modified AIS	10	77.8944%	1.19 seconds
Bayes Net	Best first search	8	72.9911 %	0.86 seconds
		11	72.6411 %	1.78 seconds
		18	71.4627 %	32.5 seconds
	Correlation	8	65.5765 %	1.89 seconds
		16	70.497 %	2.76 seconds
		20	75.6135 %	0.42 seconds
	Info Gain	8	71.0508 %	1.26 seconds
		13	72.6322 %	0.13 seconds
		17	72.1051 %	0.11 seconds
	Modified AIS	10	70.4429%	0.28 seconds

the selected features using MAIS algorithm is as follow:

- Protocol.
- Service.
- Flag
- Src_bytes.
- Same_srv_rate.
- Dst_host_same_srv_rate.
- Dst_host_diff_srv_rate.
- Dst_host_srv_serror_rate.
- Dst_host_rerror_rate.

- Dst_host_serror_rate.

In the table above, we use three methods for feature selection which are Best First Search, correlation, and Info. Gain, the best result was Info Gain feature selection reduce the features from 41 features to 13 and with J48 classifier the accuracy was 80.9693% and time required is 8.78 sec., this result is good compared to other methods of feature selection as shown in the table above. When we used the proposed algorithm as a feature selection and applied to the same classifier (J48), it gives the following results, first reduced the number of features from 13 to 10 features second increased the accuracy to 84.1509% and third greatly reduced the time to (0.01) seconds as shown in the table above. The specification hardware that apply the proposed algorithm is (Laptop HP Core I7, Ram 8GB, H.D.D 1TB) within Windows 10 and the programming language is Java Eclipse Neon.

VII. Conclusion

In this paper, a strategy was proposed for the feature selection based on developed artificial immune systems. This proposed algorithm illustrates the benefits of using the artificial immune system to increase efficiency and randomness in the characteristics of NSL-KDD dataset, and reduce these dataset attributes from 41 to 10 important features that are related to the class (normal or 4 types of attacks). The results were based on the NSL-KDD dataset and showed an increase in the accuracy of the performance compared with the other feature selection methods. The proposed algorithm used the power of clonal selection and affinity to generate a random diversity population to find the best features in the dataset that are related together to conclude the class.

References

1. Dionisios, N., Sotiropoulos & George A., Tsihrin Tzis, **2002.** *Machine Learning Paradigms Artificial Immune Systems and their applications in software personalization*. Poland Warsaw.Janvszkacprzyk. polish academy of sciences, volume 113
2. Ahmed, T., Sadiq & Jamal H. Assi. **2017**. NSL-KDD dataset classification using Five classification methods and Three feature selection strategies. *Iraq. JACSTR*, **7**(1).
3. Dhruba Kumar Bhattacharyya & Jugal Kumar Kalita **2014**. *Network Anomaly Detection a Machine Learning Perspective*. London. A Chapman & Hall Inc.
4. Carlos A. Coello & Nareli Cruz Cort'es. **June 2005.** Solving Multiobjective Optimization Problems Using an Artificial Immune System. *Springer science + Business media*, **6**(2).
5. MS Garima Singh & Sunita Bansal. **2013.** Artificial immune System Approach for multi-objective optimization. India. *Babu Banarasi Das university, Lucknow*. **4**(13).
6. Kevin Leung, France Cheong & Christopher Cheong. **2007**. Generating compact classifier system using a simple artificial immune system. *IEEE Transaction on system, man, and CyberNetics-Part B: cybernetics*, **37**(5).