Soft Comput (2019) 23:1321–1336
https://doi.org/10.1007/s00500-017-2856-4

A multi-objective evolutionary fuzzy system to obtain a broad and accurate set of solutions in intrusion detection systems

Salma Elhag[1] · Alberto Fernández[2] · Abdulrahman Altalhi[3] · Saleh Alshomrani[1] · Francisco Herrera[2,3]

Published online: 6 October 2017

Abstract Intrusion detection systems are devoted to monitor a network with aims at finding and avoiding anomalous events. In particular, we focus on misuse detection systems, which are trained to identify several known types of attacks. These can be unauthorized accesses, or denial of service attacks, among others. Whenever it scans a trace of a suspicious event, it is programmed to trigger an alert and/or to block this dangerous access to the system. Depending on the security policies of the network, the administrator may seek different requirements that will have a strong dependency on the behavior of the intrusion detection system. For a given application, the cost of raising false alarms could be higher than carrying out a preventive access lock. In other scenarios, there could be a necessity of correctly identifying the exact type of cyber attack to proceed in a given way. In this paper, we propose a multi-objective evolutionary fuzzy system for the development of a system that can be trained using different metrics. By increasing the search space during the optimization of the model, more accurate solutions are expected to be obtained. Additionally, this scheme allows the final user to decide, among a broad set of solutions, which one is better suited for the current network characteristics. Our experimental results, using the well-known KDDCup'99 problem, supports the quality of this novel approach in contrast to the state-of-the-art for evolutionary fuzzy systems in intrusion detection, as well as the C4.5 decision tree.

Keywords Intrusion detection systems · Computational intelligence · Evolutionary fuzzy systems · Multi-objective evolutionary algorithms · Misuse detection

1 Introduction

Current networks are exposed to attacks from malicious users everyday. Among others, we may stress flooding (that causes denial of service), port scanning (that searches for vulnerabilities of the system), password guessing (trying to make un unauthorized login), or buffer overflow attacks (a exploit that can allow to gain root privileges). Therefore, intrusion detection is a very important task for providing security and integrity in information systems. Analyzing the information gathered by security audit mechanisms, intrusion detection systems (IDS) apply several rules that discriminate between legitimate events or an undesirable use of the system (Vasilomanolakis et al. 2015).

When referring to IDS, we must stress two main categories (Debar et al. 1999): (1) *Misuse detection*, that are trained with an already established set of known attacks (Lee and Stolfo 2000); and (2) *Anomaly detection* that determines a profile for the "normal behavior", and sets up as "attack" any access

Communicated by V. Loia.

✉ Alberto Fernández
alberto@decsai.ugr.es

Salma Elhag
salma53ster@gmail.com

Abdulrahman Altalhi
ahaltalhi@kau.edu.sa

Saleh Alshomrani
sshomrani@kau.edu.sa

Francisco Herrera
herrera@decsai.ugr.es

[1] Faculty of Computing and Information Technology, University of Jeddah, Jeddah 21589, Saudi Arabia

[2] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain

[3] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

with different characteristics from the standard (Patcha and Park 2007). In this paper, we will focus on a static context, which implies the use of the first type of approaches.

Among different machine learning solutions, Computational Intelligent and Soft Computing techniques have shown a very robust behavior for detecting both known and unseen intrusion attacks and to recognize normal network traffic (Wu and Banzhaf 2010; Guo et al. 2014). In this area of research, fuzzy systems are a valuable tool as they are capable for modeling complex and dynamic systems. This is due to their good properties for representing uncertain knowledge, as well as for presenting a smooth adaptation to the context. We consider the use of fuzzy rule-based classification systems (FRBCSs) (Ishibuchi et al. 2004), an extension to classical rule-based systems, where its antecedents and consequent are composed of fuzzy logic statements. Furthermore, we focus on fuzzy sets with linguistic labels, allowing the output system to present a higher interpretability degree for the expert (Gacto et al. 2011).

In our last study on the topic (Elhag et al. 2015), we focused on the development of a methodology based on this paradigm for misuse detection. Specifically, we made use of the Fuzzy Association Rule-based Classification for High-Dimensional problems (FARC-HD) (Alcalá-Fdez et al. 2011), a linguistic Evolutionary Fuzzy System (EFS) (Fernandez et al. 2015) in synergy with the One-vs-One (OVO) class decomposition technique (Galar et al. 2011), in which binary subproblems are obtained by confronting all possible pair of classes. The high potential of this fuzzy rule learning approach was determined by the goodness in the correct identification for all types of attacks, including rare attack categories.

However, in the context of IDS, there are several metrics of performance to be optimized. Among others, we must stress the attack detection rate (ADR), which stands for the accuracy obtained for the attack classes managed as a whole, and the false alarm rate (FAR), i.e., the number of false positives. Additionally, we have emphasized the significance of identifying correctly every single type of attack, as it allows one being able to develop a different task in each case. For the aforementioned reasons, a Multi-Objective Evolutionary Algorithm (MOEA) (Coello-Coello et al. 2007) could be the best-suited approach for the optimization of these different measures in order to achieve several solutions which can fit different conditions (Fernandez et al. 2015; Mohammadi Shanghooshabad and Saniee Abadeh 2016).

In this work, we propose the integration of an MOEA approach within the learning procedure of the FARC-HD algorithm aiming for the achievement of a wide set of accurate solutions. The synergy between the multi-objective process and the fuzzy representation has already provided several advantages under different scenarios such as pixel classification from remote sensing imagery (Alok et al. 2016) or market clearing of joint energy and reserves auctions (Goroohi Sardou and Ameli 2016). Specifically, we must stress the following ones:

- Any desired metrics can be selected for the optimization.
- It allows a decision among several configurations (codified in each solution) depending on the final requirements of the system.
- The smoothness related to the linguistic fuzzy terms allows the model to provide a better separability among different types of alarms.
- Finally, a better understanding of the working procedure of the IDS is implicit to the use of linguistic fuzzy rules with a low number of antecedent values.

The goodness of our novel methodology will be tested using several datasets from the area of IDS. Specifically, we have selected the standard KDDCUP'99 dataset (Lee and Stolfo 2000), the NSL-KDD (Tavallaee et al. 2009), and the Gure-KDDCup (Perona et al. 2008). This way, the experimental results will be directly comparable with most of the Computational Intelligence approaches for intrusion detection. Specifically, for the evaluation of the experimental results, we will compare them versus the standard FARC-HD algorithm (Alcalá-Fdez et al. 2011), as well as our previous approach on the topic, i.e., FARC-HD-OVO (Elhag et al. 2015), which was shown to outperform the state-of-the-art EFS algorithms for misuse detection. Finally, we will complement our comparison with the classical C4.5 decision tree (Quinlan 1993).

The remainder of the manuscript is organized as follows. Section 2 introduces some preliminary concepts, including the context of IDS, some basic notions on FRBCSs and the details for the working procedure of FARC-HD. Next, Sect. 3 presents our proposal for the development of the methodology integrating the MOEA and the fuzzy rule learning in misuse detection. Next, in Sect. 4 we set up the experimental framework, including the features of the KDDCUP'99 dataset, metrics of performance, algorithms for comparison and their parameters. Then, the analysis of the performance of this novel approach with respect to the state-of-the-art is shown in Sect. 5. Finally, Sect. 6 summarizes and concludes the work.

2 Preliminaries: intrusion detection systems and fuzzy rule-based classification systems

In this section, we will introduce several concepts that will present a clearer definition of the work context, as well as the features of the type of classifier that will be used to build our final solution. In this way, a summary of the main concepts for IDS will be presented first in Sect. 2.1. Next, in Sect. 2.2, we will recall some brief details on FRBCSs. Finally, in Sect. 2.3, we will describe the features of the FARC-HD algorithm, which has been selected as baseline technique for this research.

2.1 Intrusion detection systems

In this data age, we are witnessing how computer systems are creating, processing, and sharing an overwhelming quantity of information. According to this fact, computer security must be regarded as a critical issue, so that the unauthorized access to this data from a computer and/or computer network, could imply a significant problem, as it compromises the integrity, confidentiality and availability of the resources (Chebrolu et al. 2005). This issue is of extreme importance in several application areas such as have medical (Mitchell and Chen 2015) or power systems (Pan et al. 2015). Therefore, a wide amount of computer security tools such as antiviruses, firewalls, data encryption, and so on. In addition to this, there are some complementary tools that monitor the activity of the network in order to detect and block intrusions.

Anomalous activities are thus identified by IDSs, which comprise the process of monitoring and analyzing events occurring in a computer system or network in order to detect anomalous activity (Vasilomanolakis et al. 2015). In particular, IDS can be split into two categories according to the detection methods they employ, including (1) misuse detection and (2) anomaly detection.

The main difference between both types of systems is related to whether they use a signature detection or anomaly detection paradigm. Misuse detection systems take the majority of IDSs, and use an established set of known attack patterns, and then monitor the net trying to match incoming packets and/or command sequences to the signatures of known attacks (Lee and Stolfo 2000). Hence, decisions are made based on the prior knowledge acquired from the model. The main advantage of this type of IDS is that they provide high detection accuracy with few false positives, but with the disadvantage that they are not able to detect new attacks other than those previously stored in the database.

On the other hand, anomaly detection IDS have the ability to detect new attacks, but at the cost of increasing the number of false positives. In an initial phase, the anomaly-based IDS is trained in order to obtain a normal profile of activity in the system (Patcha and Park 2007). The learned profiles of normal activity are customized for every system, making it quite difficult for an attacker to know with certainty what activities it can carry out without getting detected. Then, incoming traffic is processed in order to detect variations in comparison with the normal activity, in which case it will be considered as a suspicious activity. In addition to the higher number of false alarms raised, another disadvantage of the development a system of these characteristics is the higher the complexity compared to the case of misuse detection.

2.2 Introduction to FRBCSs

Any classification problem consists of m training patterns $x_p = (x_{p1}, \ldots, x_{pn}, C_p)$, $p = 1, 2, \ldots, m$ from M classes where x_{pi} is the ith attribute value ($i = 1, 2, \ldots, n$) of the pth training pattern.

In this work, we use fuzzy rules of the following form for our FRBCSs:

$$\begin{aligned}&\text{Rule } R_j : \text{ If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn}\\ &\text{then Class} = C_j \text{ with RW}_j\end{aligned} \quad (1)$$

where R_j is the label of the jth rule, $x = (x_1, \ldots, x_n)$ is an n-dimensional pattern vector, A_{ji} is an antecedent fuzzy set, C_j is a class label, and RW_j is the rule weight (Ishibuchi and Yamamoto 2005). We use triangular MFs as antecedent fuzzy sets.

When a new pattern x_p is selected for classification, then the steps of the fuzzy reasoning method (Cordón et al. 1999) are as follows:

1. *Matching degree*, that is, the strength of activation of the if-part for all rules in the Rule Base with the pattern x_p. A conjunction operator (t-norm) T, is applied in order to carry out this computation:

$$\begin{aligned}&\mu_{A_j}(x_p) = T(\mu_{A_{j1}}(x_{p1}), \ldots, \mu_{A_{jn}}(x_{pn})),\\ &\qquad j = 1, \ldots, L\end{aligned} \quad (2)$$

2. *Association degree* To compute the association degree of the pattern x_p with the M classes according to each rule in the Rule Base. To this aim, a combination operator h, is applied to combine the matching degree with the rule weight (RW). In our case, this association degree only refers to the consequent class of the rule (i.e., $k = \text{Class}(R_j)$).

$$\begin{aligned}&b_j^k = h\left(\mu_{A_j}(x_p), \text{RW}_j^k\right),\\ &k = 1, \ldots, M; \qquad j = 1, \ldots, L\end{aligned} \quad (3)$$

3. *Pattern classification soundness degree for all classes* We use an aggregation function f, which combines the positive degrees of association calculated in the previous step.

$$Y_k = f\left(b_j^k, j = 1, \ldots, L \text{ and } b_j^k > 0\right), \qquad k = 1, \ldots, M \quad (4)$$

4. *Classification* We apply a decision function F over the soundness degree of the system for the pattern classification for all classes. This function will determine the class label l corresponding to the maximum value.

$$F(Y_1, \ldots, Y_M) = \arg\max(Y_k), \qquad [k = 1, \ldots, M] \tag{5}$$

where L denotes the number of rules in the Rule Base and M the number of classes of the problem.

2.3 Baseline fuzzy classifier: FARC-HD algorithm

The great challenge for IDS is to develop a system that can guarantee both a high attack detection rate and a low false alarm rate. In order to overcome this, Data Mining techniques, especially Soft Computing and Computational Intelligence approaches, have become essential pieces for addressing this problem (Wu and Banzhaf 2010).

Among these, FRBCSs (Ishibuchi et al. 2004) have emerged as a valuable solution in accordance with their inner properties. First, the use of fuzzy labels allows a smoother management for the user profile, i.e., decreasing false alarms. Additionally, security itself includes fuzziness in the definition between normal and abnormal behavior. Another positive feature for FRBCSs is their extension to EFSs (Fernandez et al. 2015), i.e., the hybridization between Fuzzy Systems and Evolutionary Algorithms (Eiben and Smith 2003), which leads to a leap of quality as the fuzzy system is adapted to the context of the problem.

Among different methods based in EFSs for their application in IDS, in our last work on the topic, we have shown the goodness of the FARC-HD approach (Alcalá-Fdez et al. 2011; Elhag et al. 2015) over the most representative algorithms of this paradigm. FARC-HD algorithm is based on association discovery (Zhang and Zhang 2002), whose integration with classification allows the achievement of precise and interpretable models. In addition, it has shown to be quite robust for high dimensional problems, a feature that is very significant in the case of IDS (Bostani and Sheikhan 2017).

In summary, the FARC-HD method is based on the following three stages (as depicted in Fig. 1):

Stage 1 *Fuzzy association rule extraction for classification* A search tree is employed to list all possible frequent fuzzy item sets and to generate fuzzy association rules for classification, limiting the depth of the branches in order to find a small number of short (i.e., simple) fuzzy rules.
Stage 2 *Candidate rule pre-screening* Afterward, the rule generation, the size of the rule set can be too large to be interpretable by the end user. Therefore, a preselection of the most interesting rules is carried out by means of a "subgroup discovery" mechanism based on an improved weighted relative accuracy measure (wWRAcc') (Kavsek and Lavrac 2006).
Stage 3 *Genetic rule selection and lateral tuning* Finally, in order to obtain a compact and accurate set of rules within the context of each problem, an evolutionary process will be carried out in a combination for the selection of the rules with a tuning of membership function, as its positive synergy has been shown in previous work on the topic (Alcala et al. 2007; Casillas et al. 2005).

3 Proposed methodology: a multi-objective evolutionary fuzzy system approach for IDS

The premise behind our research is to develop a misuse detection tool that may provide several advantages in this area. Specifically, we have focused on three different issues that are of high interest for the community in IDS. First, the output model must be interpretable by the final user, i.e., it should easily allow to explain the phenomena that has been detected. Second, it must have the ability to adapt to different requirements. Finally, it should be able not only to recognize an suspicious behavior, but also to identify the particular type of attack with a both good recall and precision.

To accomplish the first goal, we will use as baseline classifier the FARC-HD approach. As described in the previous section, this FRBCS is based on linguistic fuzzy labels, which enable rules' semantic to be closer to human natural language (Gacto et al. 2011). In addition to the former, during the learning stage of FARC-HD, the maximum total length of the final rules can be limited to few antecedents. In case of problems with a large number of variables, this is a clear advantage for the compactness of the output model.

The second and third issues are closely related. We must take into account that in the area of IDS there are many possible ways to determine the quality of the software system, which will be described with detail in Sect. 4.3). Therefore, for the user, it might be important to decide the desired main objectives of the IDS among all the possible metrics of performance (Kudlacik et al. 2016). However, most of these may be in conflict one with another. A clear main example is the achievement of a high attack detection rate and a low false alarm rate. Therefore, our proposed approach must be able not only to derive a different behavior according to the user requirements, but also to be robust enough even in case of conflicting objectives.

In accordance with the former, we must tune FARC-HD toward the achievement of the highest results for a given set of IDS metrics of performance. A straightforward way for addressing the former task is by carrying out an optimization of the initial FARC-HD RB so that we may accomplish the user objectives with a high-quality fuzzy system.

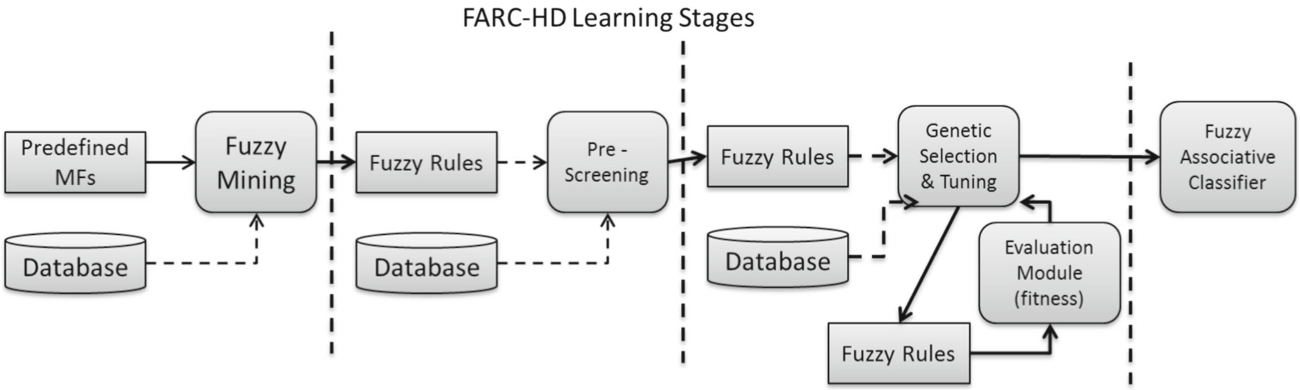


Fig. 1 Learning stages for the FARC-HD algorithm

Considering that we must manage a multi-objective problem, the most straightforward way to carry out this process is to build a single function that aggregates every objectives. However, this implies two main drawbacks: (1) to determine the optimal weight for each one of the objectives; and (2) it is only possible to obtain a single compromise solution.

For this reason, in these cases, it is more reliable to proceed via an MOEA, i.e., a Pareto-optimization (Coello-Coello et al. 2007). It allows the optimization of several objectives by evolving a set of non-dominated solutions. This "non-dominance" term refers to the case when a given solution s_1 is equal or better than another s_2 for all objectives but one. For example, in a two-objective minimization problem, s_1 and s_2 are considered to be a couple of non-dominated solutions when $s_1 = (0.0, 0.5)$ and $s_2 = (0.1, 0.3)$; on the contrary, s_1 dominates s_3 when $s_3 = (0.1, 0.6)$.

There are two main advantages related to this procedure. The first one is to make the search space broader, thus resulting in a better exploration ability. The second one is providing a set of solutions from which the user or expert can select the one that is better suited to the system's requirements. We must stress that both issues cover our initial premises for obtaining a robust IDS.

As stated previously, our proposal is acting on the optimization procedure of FARC-HD, namely the "Genetic rule selection and lateral tuning" that is carried out in the last stage of the FARC-HD algorithm (Alcalá-Fdez et al. 2011). Instead of carrying out a single evolutionary process, we substitute it with an MOEA approach that is able to generate different Knowledge Bases in accordance with the objectives selected.

For developing our model, we will make use of the NSGA-II algorithm (Deb et al. 2002), as it is widely known for being a high-performance MOEA. Its two main features are first the fitness evaluation of each solution based on both the Pareto ranking and a crowding measure, and the other is an elitist generation update procedure.

In order to codify the solutions, and following the same than in the original optimization process for FARC-HD, we will make use of a chromosome with two well differentiate parts: one (RS) for the rule selection and another one (DB) for the tuning of the Data Base:

- The first part will have a binary codification. The length of this part of the solution will be the number of rules selected in *Stage 2* (pre-screening) of the FARC-HD process. Therefore, a 0 value means that the rule will not take part for generating the classification model, whereas a 1 value stands for the opposite case.
- The second part will have a real codification in the range [0, 1]. There will be as many genes as different labels contained in the Data Base. Following the 2-tuples linguistic representation (Alcala et al. 2007; Fernández et al. 2010; Herrera and Martínez 2000), values in [0.0, 0.5) imply a lateral displacement to the left, whereas values in the (0.5, 1] are displacements to the right in the original universe of discourse. Accordingly, the initial state of the linguistic label corresponds to the value 0.5.

Chromosomes will be evaluated jointly with aims at obtaining the best synergy between both characteristics, instead of optimizing them separately. This issue is based on the fact that it is not clearly defined which the best order for carrying our both processes is. An initial chromosome will be built with all binary genes equal to '1' and real genes to '0.5' in order to implement the standard case study, i.e., the original Knowledge Base, whereas the remaining individuals will be generated at random.

In order to proceed with the evaluation of the chromosomes, the decoding implies the generation of a new Knowledge Base from the configuration parameters given in the solution. Then, this model will be applied to training set in order to obtain the values for the performance metrics selected as objectives. The list of most commonly used performance metrics in the context of IDS will be introduced in Sect. 4.3.

Finally, for the sake of allowing a better exploitation in the search, we have limited the number of objectives to two. In the experimental study (Sect. 5), we will present several case studies from which we can conclude the best synergy between the selected performance metrics.

4 Experimental framework

This section is devoted to establish the configuration for the experiments in order to carry out a thorough analysis. With this aim, Sect. 4.1 introduce the features of the selected benchmark problems. The learning algorithms and their configuration parameters will be presented in Sect. 4.2. Finally, the metrics of performance applied to compare the results obtained with the different classifiers are described in Sect. 4.3.

4.1 Benchmark data: IDS problems

Among different benchmark problems for IDS, the KDD-CUP'99 dataset is possibly the widest used one, being a standard until today (Benferhat et al. 2013; Khor et al. 2012; Chung and Wahid 2012). It was obtained by the Information System Technology (IST) group of Lincoln laboratories at MIT university under contract of DARPA and in collaboration with ARFL (Lee and Stolfo 2000). It consisted of an environment of a local area network (LAN) that simulates a typical U.S. Air Force LAN, including several weeks of raw TCP dump data with normal activities and various types of attacks.

It comprises 41 attributes in total, which are divided three main groups: intrinsic features (extracted from the headers' area of the network packets), content features (extracted from the contents area of the network packets), traffic features (extracted with information about previous connections).

Class labels are divided into normal and attack activities. This last class can be further divided into particular types of attack, which are basically grouped into four major categories, namely:

- Denial of Service (DOS): make some machine resources unavailable or too busy to answer to legitimate users requests (SYN flooding).
- Probing (PRB): Surveillance for information gathering or known vulnerabilities about a network or a system (port scanning).
- Remote To Local (R2L): use vulnerability in order to obtain unauthorized access from a remote machine (password guessing).
- User To Root (U2R): exploit vulnerabilities on a system to gain local super-user (root) privileges (buffer overflow attack).

From this classical and well-known dataset, several alternative problems have been generated trying to both overcome some of the limitations from the original data (such as the number of redundant records), and to update its inner characteristics for novel IDS requirements. Specifically, we have made use of the NSL-KDD (Tavallaee et al. 2009) and the Gure-KDDCUP (Perona et al. 2008) datasets.

The NSL-KDD dataset contains an informed subsample from the original KDDCUP'99 dataset. First, all redundant records in the train and test sets were removed, so that classifiers will not be biased toward more frequent records. Then, instances were selected in accordance to their difficulty level, whose value is computed regarding the number of correct hits obtained by 7 different learners, each of which executed three times. The number of attributes and classes remain the same with respect to the KDDCUP'99 dataset.

Gure-KDDCUP problem is a novel IDS dataset which followed an extraction process similar to that of KDDCUP'99. In this sense, authors processed "TCP dump" files with bro-ids[1] and stored each connection with its attributes. Finally, connections were labeled based on the connections-class files ("tcpdump.list") provided by MIT.

In the NSL-KDD dataset, the number of records in the train and test sets are reasonable (1,074,992 records), making affordable to run the experiments on the complete set. However, in both KDDCUP'99 and Gure-KDDCUP problems, the total amount of data places them in the context of Big Data (Fernandez et al. 2014), i.e., affecting the scalability of current approaches. For this reason, usually a small portion of the whole data is randomly selected for its use with standard classifiers.

In accordance with the former, for the KDDCUP'99 dataset we will select just a 10% of the instances for our experiments. This implies a total of 494,021 connections. Then, we have also removed all duplicated instances, reducing the data to a total of 145,585 examples.

Analogously to the previous case, authors of the Gure-KDDCUP dataset made available a subset with a 6% of the records. This sample contains all no-flood attacks matched with tcpdump.list and a random subsample of normal connections matched with tcpdump.list, thus including 178,858 connections. It is important to point out, on the contrary to KDDCUP'99 and NSL-KDDCUP problem, this particular case study does not contain any record for the PRB attack type.

Finally, in order to carry out a validation procedure of the results, we have selected a hold-out methodology. Specifi-

[1] http://www.bro-ids.org/.

cally, we will employ a 10% of the datasets for training and the remaining 90% for test. However, in order to take into account the original distribution of classes, we will include a 50% of instances for U2R in both training and test. Table 1 shows the final distribution of examples for each partition/class.

4.2 Algorithms and parameters

In this paper, we have considered several algorithms for a fair analysis of the behavior of our proposal. The choice of FARC-HD (Alcalá-Fdez et al. 2011) as the baseline classifier makes mandatory to include its original version for this study. Additionally, we must make use of its multi-classifier extension, i.e., FARC-HD-OVO (Elhag et al. 2015), as it was shown to be one of the most accurate EFSs for the task of intrusion detection. Finally, we will include C4.5 (Quinlan 1993) in the experimental study as a state-of-the-art rule induction algorithm. In what follows, we detail the configuration of the parameters for each approach:

1. *FARC-HD* (Alcalá-Fdez et al. 2011) First, we have selected 5 labels per variables for the fuzzy sets, product t-norm as conjunction operator and additive combination for the inference procedure. As specific parameters of the learning stage, we have set up the minimum support to 0.05 and the minimum confidence to 0.8. Finally, we have fixed the maximum depth of the tree to a value of 3, and the k parameter for the pre-screening to 2. For more details about these parameters, please refer to Alcalá-Fdez et al. (2011).

   We must stress that this configuration will be shared for all three models based on FARC-HD, i.e., the standard approach, FARC-HD-OVO, and our proposed model FARC-HD-MOEA.
2. *FARC-HD-OVO* (Elhag et al. 2015) The learning procedure will be performed using all possible pairs of classes. In order to aggregate the outputs of each binary classifier into a single solution, we will make use of the preference relations solved by Non-Dominance Criterion (ND) (Fernández et al. 2010).
3. *FARC-HD-MOEA* The parameters of the NSGA-II MOEA have been set up as follows: 50 individuals as population size, with 20000 generations. The crossover and the mutation (per gen) probabilities are 0.9 and 0.025, respectively.
4. *C4.5* (Quinlan 1993) For C4.5 we have set a confidence level of 0.25, the minimum number of item-sets per leaf was set to 2 and the application of pruning was used to obtain the final tree. We must point out that, for the sake of allowing the output model to be compact and interpretable, we have carried out an extensive pruning. Specifically, we have limited the maximum depth of the tree to 3. Therefore, rules obtained from C4.5 will be of the same length than those learned by the FARC-HD algorithms, establishing a fair comparison between both techniques.

4.3 Performance metrics for IDS

In the specialized literature for IDS in general, and for misuse detection in particular, authors have made use of several metrics of performance for the evaluation of their results in comparison with the state-of-the-art. In this paper, we have selected different measures which will allow us to analyze the behavior of our approach under several perspectives:

1. *Accuracy* It stands for the global percentage of hits. In our case (IDS), its contribution is low as it does not take into account the individual accuracies of each class, but it has been selected as a classical measure.

   $$\text{Acc} = \frac{\sum_{i=1}^{C} TP_i}{N} \tag{6}$$

   where C stands for the number of classes, N stands for the number of examples and TP_i is the number of True Positives of the ith class.
2. *Mean F-measure* In the binary case, the standard f-measure computes a trade-off between precision and recall of both classes. In this case, we compute the average for the F-measure achieved for each class (taken as positive) and the remaining ones (taken as a whole as negative):

   $$\text{MFM} = \frac{\sum_{i=1}^{C} \text{FM}_i}{C} \tag{7}$$

   $$\text{FM}_i = \frac{2 \cdot \text{Recall}_i \cdot \text{Precision}_i}{\text{Recall}_i + \text{Precision}_i} \tag{8}$$

   $$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \tag{9}$$

   $$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{10}$$

   where TP_i, FP_i and FN_i are the number of true positives, false positives and false negatives of the ith class, respectively, percentage).
3. *Average accuracy* It is computed as the average of the individual hits for each class. For this reason, it is also known as the average recall:

   $$\text{AvgAcc} = \frac{1}{C}\sum_{i=1}^{C} \text{Recall}_i \tag{11}$$

Table 1 Number of examples per class in each dataset partition for KDDCUP'99, NSL-KDD and Gure-KDDCUP problems

Class	KDDCUP'99		NSL-KDDCUP		Gure-KDDCUP	
	#Ex. training	#Ex. test	#Ex. training	#Ex. test	#Ex. training	#Ex. test
Normal	8783	79,049	6674	60,669	17,488	157,385
DOS	5457	49,115	4648	41,279	112	1012
PRB	213	1917	103	892	–	–
R2L	100	899	1173	10,483	274	2465
U2R	26	26	26	26	37	37
Total	14,579	131,006				

Table 2 Case studies for different metrics of performance selected as objectives within the MOEA approach

Case Study	Metric 1	Metric 2
Case 1	Attack detection rate (ADR)	False alarm rate (FAR)
Case 2	Average accuracy (Avg-Acc)	False alarm rate (FAR)
Case 3	Attack accuracy (Att-Acc)	False alarm rate (FAR)
Case 4	Mean F-measure (MfM)	False alarm rate (FAR)
Case 5	Mean F-measure (MfM)	Average accuracy (Avg-Acc)

4. *Attack accuracy* In this case we omit the "Normal" instances and we focus in checking whether we guess correctly the different "Attack" types individually.

$$\text{AttAcc} = \frac{1}{C-1}\sum_{i=2}^{C}\text{Recall}_i \tag{12}$$

In this case, the first class ($i = 1$) is considered to be the "Normal" class.

5. *Attack detection rate* It stands for the accuracy rate for the attack classes. Therefore, it is computed as:

$$\text{ADR} = \frac{\sum_{i=2}^{C}\text{TP}_i}{\sum_{i=2}^{C}\text{TP}_i + \text{FN}_i} \tag{13}$$

Reader must take into account that also in this case, the first class ($i = 1$) is considered to be the "Normal" class.

6. *False alarm rate* In this case, we focus on the "Normal" examples, and we check which is the percentage of "false negatives" found, i.e., those instances identified as "alarms" but which are actually normal behavior.

$$\text{FAR} = \frac{\text{FP}_1}{\text{TP}_1 + \text{FP}_1} \tag{14}$$

As in the former metric (ADR), the "Normal" class has the first index ($i = 1$).

5 Experimental study

This section includes the experimental results and the analysis of the former to support the goodness of our proposed approach. In particular, we will first perform a selection of the best combination of objectives to carry out the search, i.e., those whose synergy allows a better exploration of the search space and thus show an good overall performance (Sect. 5.1). Then, once the best parameters have been found, we will contrast the behavior of this novel methodology with that of the state-of-the-art for rule learning and fuzzy rule learning, i.e., the C4.5 decision tree (Quinlan 1993), the original FARC-HD algorithm (Alcalá-Fdez et al. 2011), and FARC-HD-OVO (Elhag et al. 2015) (Sect. 5.2). Finally, we provide additional information for the experimental study by showing the whole Pareto fronts for all case studies in training and test, as well as the confusion matrices in the test partition (Sect. 5.3).

5.1 Analysis of the best objectives' combination

The field of IDS has several metrics of performance depending on the user's requirements. Section 4.3 introduced the main objectives that are commonly used to measure the quality of the designed models to address this task. In this work, we do not seek a particular goal, but an "multi-purpose" method that will be able to adapt well under several scenarios, trying to maximize the precision and recall among all classes, regardless of their type.

According to the previous fact, our first step is to find the better combination of objectives that leads to a robust average performance. Among all available metrics, we consider that the "False Alarm Rate" (FAR) is the mandatory one to

Table 3 Experimental results in CASE STUDY 1 (AttDet vs. FAR)

Metric	KDDCUP'99						NSL-KDDCUP						Gure-KDDCUP					
	Best-AttDet		Best-FAR		Knee		Best-AttDet		Best-FAR		Knee		Best-AttDet		Best-FAR		Knee	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	78.74	78.25	97.24	**97.23**	95.09	95.16	96.29	96.11	94.79	94.96	97.37	**97.33**	93.47	93.49	98.94	**99.03**	98.86	98.71
MfM	68.22	58.95	69.01	66.75	73.33	**74.65**	84.49	73.21	56.68	56.76	80.93	**75.69**	76.30	62.33	73.39	67.17	86.55	**71.49**
AvgAcc	87.09	85.64	63.54	61.92	80.90	**90.25**	87.91	**84.78**	55.73	55.82	73.65	70.52	91.27	**82.54**	65.67	61.51	88.03	77.86
ADR	99.74	**99.71**	93.05	93.07	98.77	98.59	97.51	**97.37**	88.97	89.27	94.57	94.55	97.16	**97.98**	55.32	56.29	85.34	81.27
FAR	35.124	35.856	0.0000	**0.0266**	7.332	7.099	4.794	4.978	0.0300	**0.1071**	0.1349	0.2489	6.621	6.606	0.0000	**0.0165**	0.8120	0.9003
AttAcc	92.64	**91.02**	54.43	52.41	77.95	89.59	86.08	**82.22**	44.66	44.80	67.10	63.21	90.57	**78.92**	54.23	48.69	84.31	70.78

Bold values correspond to the best result in test for each metric and dataset

Table 4 Experimental results in CASE STUDY 2 (AvgAcc vs. FAR)

Metric	KDDCUP'99						NSL-KDDCUP						Gure-KDDCUP					
	Best-AvgAcc		Best-FAR		Knee		Best-AvgAcc		Best-FAR		Knee		Best-AvgAcc		Best-FAR		Knee	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	98.59	**98.58**	97.59	97.41	98.59	**98.58**	96.68	**96.60**	93.23	93.26	95.58	95.53	95.02	94.89	98.90	98.91	99.24	**99.15**
MfM	90.38	80.83	87.16	**81.42**	90.38	80.83	85.64	74.50	81.00	75.94	86.79	**78.51**	68.67	65.50	75.62	65.34	87.69	**73.24**
AvgAcc	92.78	**91.60**	79.93	80.76	92.78	**91.60**	88.54	**85.29**	73.47	71.58	81.89	77.09	93.61	**86.72**	67.28	60.29	81.25	71.51
ADR	97.14	**97.32**	93.94	93.59	97.14	**97.32**	96.20	**96.36**	85.66	85.64	90.84	90.80	88.42	**82.61**	53.43	51.02	69.98	65.51
FAR	0.4440	0.6009	0.0000	**0.0873**	0.4440	0.6009	2.8918	3.1795	0.0300	**0.1286**	0.1948	0.3626	4.8204	4.8353	0.0000	**0.0178**	0.0515	0.0940
AttAcc	91.08	**89.65**	74.91	75.97	91.08	**89.65**	86.40	**82.40**	66.85	64.51	77.42	71.46	93.08	**83.90**	56.38	47.06	75.02	62.04

Bold values correspond to the best result in test for each metric and dataset

Table 5 Experimental results in CASE STUDY 3 (AttAcc vs. FAR)

Metric	KDDCUP'99						NSL-KDDCUP						Gure-KDDCUP					
	Best-AttAcc		Best-FAR		Knee		Best-AttAcc		Best-FAR		Knee		Best-AttAcc		Best-FAR		Knee	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	91.20	90.92	93.87	**93.40**	93.75	93.18	89.23	88.89	85.24	85.49	96.29	**96.23**	95.30	95.11	99.11	**99.06**	99.02	98.79
MfM	70.12	60.50	75.07	**65.00**	74.30	64.65	67.22	59.28	73.23	68.82	86.19	**74.58**	68.41	64.98	84.02	**70.61**	85.87	69.68
AvgAcc	91.53	90.00	91.11	89.55	91.33	**90.17**	87.60	84.78	64.80	63.08	88.42	**84.92**	94.53	**88.01**	75.79	65.81	91.26	82.50
ADR	96.55	**96.62**	95.76	95.69	95.60	95.29	97.19	**97.23**	68.71	68.91	94.17	94.32	89.13	**83.38**	62.17	57.97	87.23	82.10
FAR	12.3306	12.8288	7.3779	**8.1102**	7.4690	8.2050	17.8604	18.3636	0.0150	**0.1038**	1.8130	2.1032	4.5460	4.6243	0.0000	**0.0248**	0.6919	0.8336
AttAcc	92.49	**90.70**	90.74	88.97	91.03	89.76	88.96	**85.56**	56.00	53.88	85.98	81.68	94.23	**85.56**	67.72	54.42	88.58	76.95

Bold values correspond to the best result in test for each metric and dataset

Table 6 Experimental results in CASE STUDY 4 (MfM vs. FAR)

Metric	KDDCUP'99						NSL-KDDCUP						Gure-KDDCUP					
	Best-MfM		Best-FAR		Knee		Best-MfM		Best-FAR		Knee		Best-MfM		Best-FAR		Knee	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	98.11	**97.89**	97.35	97.19	98.09	97.85	97.80	**97.64**	92.93	93.08	95.37	95.22	99.38	99.20	99.07	99.01	99.36	**99.21**
MfM	91.99	86.06	81.82	77.94	91.97	**86.39**	89.26	**78.66**	80.81	74.69	87.79	77.27	90.86	75.60	83.51	69.77	90.55	**75.64**
AvgAcc	89.57	**89.30**	73.68	75.64	89.46	89.15	85.39	**80.28**	72.83	71.12	83.36	78.39	87.76	**78.80**	75.24	64.74	87.04	78.13
ADR	95.84	**95.53**	93.32	93.00	95.70	95.36	96.17	**96.07**	85.04	85.26	90.49	90.36	82.74	**78.12**	60.76	55.66	80.85	76.15
FAR	0.3871	0.5528	0.0000	**0.0620**	0.3302	0.5123	0.7492	0.9956	0.0300	**0.1286**	0.2847	0.5571	0.2173	0.3253	0.0000	**0.0241**	0.1944	0.2770
AttAcc	87.06	**86.77**	67.10	69.57	86.91	86.56	81.93	**75.60**	66.04	63.93	79.27	73.13	83.75	**71.85**	66.99	53.00	82.78	70.93

Bold values correspond to the best result in test for each metric and dataset

Table 7 Experimental results in CASE STUDY 5 (MfM vs. AvgAcc)

Metric	KDDCUP'99						NSL-KDDCUP						Gure-KDDCUP					
	Best-MfM		Best-AvgAcc		Knee		Best-MfM		Best-AvgAcc		Knee		Best-MfM		Best-AvgAcc		Knee	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	98.18	97.89	97.95	97.82	98.11	**97.89**	97.47	97.42	97.41	97.20	97.47	**97.42**	99.34	**99.17**	95.72	95.56	97.47	97.37
MfM	92.50	**81.95**	86.64	76.62	89.77	79.21	89.91	77.79	88.14	75.25	89.91	**77.94**	89.97	**73.87**	70.33	66.29	82.05	66.80
AvgAcc	91.93	89.12	92.91	90.51	92.88	**91.20**	88.05	82.84	88.81	**83.59**	88.05	82.84	87.21	77.44	93.56	**83.57**	90.91	82.21
ADR	96.33	95.84	96.60	**96.35**	96.62	96.30	96.18	96.16	96.59	**96.52**	96.18	96.16	82.51	78.54	86.29	79.71	91.96	**91.63**
FAR	0.5921	**0.7641**	1.1613	1.2132	0.8995	1.0690	1.3935	1.4835	1.8580	2.2137	1.3935	**1.4769**	0.2573	**0.3704**	4.0485	4.0811	2.3959	2.4990
AttAcc	90.06	86.59	91.43	88.44	91.32	**89.27**	85.42	78.92	86.48	**80.04**	85.42	78.92	83.03	70.05	92.77	**79.46**	88.67	77.12

Bold values correspond to the best result in test for each metric and dataset

control the bias toward the majority class, i.e., *Normal activity*. Therefore, we will carry out the combination for all the remaining metrics together with the former one. Additionally, we will also combine the "mean F-measure" with the "average accuracy" as both try to balance the correct recognition among all classes. Specifically, all combinations are summarized in Table 2. We must point out that the standard accuracy was not considered as it does not take into account the individual rates for the different concepts of the problem.

Tables 3, 4, 5, 6 and 7 show the experimental results for the training and test partitions of the three selected problems, namely KDDCUP'99, NSL-KDDCUP and Gure-KDDCUP. All performance measures are considered, as introduced in Sect. 4.3: Accuracy (Acc), Mean F-measure (MFM), Average Accuracy (AvgAcc), Attack average accuracy (AttAcc), Attack Detection Rate (ADR), and False Alarm Rate (FAR). Since we must extract a single KB from a single solution of the final set, we have chosen three different points from the Pareto. Specifically, we have considered the maximal values for each objective, as well as the "knee point" (Branke et al. 2004), since this solution is likely to be the most relevant to the decision maker, as it represents the compromise between both objectives.

We acknowledge that if we focus on individual measures, the case study that has selected it for the learning process will achieve unequivocally the highest results. It is interesting to point out that, when the "False alarm rate" metric is selected in combination with different metrics, the obtained results vary. This fact implies the benefits of the use of the MOEA to seek for a wide variety of IDS depending on the desired behavior.

Among all case studies that have been analyzed, the one that probably shows the best results is the combination between "mean F-measure" and "False alarm rate." Specifically, if we focus on the values obtained using the "best MfM," we must highlight a superior behavior in most of the metrics, or at least a similar performance, thus stressing the advantage of this configuration for the search procedure.

5.2 Comparison versus the state-of-the-art

Once we have selected the best combination of objectives for the genetic learning, we will analyze the goodness of this approach versus the methods from the state-of-the-art selected in the experimental framework, i.e., the original FARC-HD (Alcalá-Fdez et al. 2011) and FARC-HD-OVO (Elhag et al. 2015), and the C4.5 decision tree (Quinlan 1993). Experimental results are divided with respect to the benchmark IDS problem. Performance values for the KDDCUP'99 dataset are shown in Table 8, for the NSL-KDD dataset in Table 9, and for the Gure-KDDCUP in Table 10.

For all selected problems, FARC-HD-MOEA improves the results of FARC-HD is most of the considered metrics

Table 8 Complete experimental results for our proposed approach (FARC-HD-MOEA (MfM+FAR)) versus the state-of-the-art (FARC-HD, FARC-HD-OVO, and C4.5) over the reduced KDDCUP'99 dataset for different metrics of performance: accuracy (Acc), mean F-measure (MFM), average accuracy (AvgAcc), attack average accuracy (AttAcc), attack detection rate (ADR), and false alarm rate (FAR)

Metric	FARC-HD-MOEA		FARC-HD		FARC-HD-OVO		C4.5	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	98.11	97.89	98.42	98.30	99.18	99.00	99.49	**99.44**
MfM	91.99	**86.06**	90.69	84.26	97.72	84.12	92.96	80.85
AvgAcc	89.57	89.30	88.31	87.76	96.50	**89.32**	91.20	86.84
AttAcc	87.06	**86.77**	85.44	84.77	95.64	86.70	89.04	83.61
ADR	95.84	95.53	96.27	96.17	98.07	97.77	98.96	**98.93**
FAR	0.3871	0.5528	0.1708	0.2948	0.0797	**0.1910**	0.1594	0.2277

Bold values correspond to the best result in test for each metric and dataset

Table 9 Complete experimental results for our proposed approach (FARC-HD-MOEA (MfM + FAR)) versus the state-of-the-art (FARC-HD, FARC-HD-OVO, and C4.5) over the NSL-KDD dataset for different metrics of performance: accuracy (Acc), mean F-measure (MFM), average accuracy (AvgAcc), attack average accuracy (AttAcc), attack detection rate (ADR), and false alarm rate (FAR)

Metric	FARC-HD-MOEA		FARC-HD		FARC-HD-OVO		C4.5	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	97.80	97.64	97.77	97.86	98.10	**98.10**	97.36	97.12
MfM	89.26	78.66	72.40	71.49	90.70	**79.24**	86.08	78.80
AvgAcc	85.39	80.28	69.38	68.58	87.52	**82.91**	84.91	81.88
AttAcc	81.93	75.60	61.92	60.95	84.59	**78.86**	81.90	78.20
ADR	96.17	96.07	96.13	96.44	96.79	96.91	97.85	**97.72**
FAR	0.7492	0.9956	0.7642	0.9000	0.7342	**0.8588**	3.0716	3.3955

Bold values correspond to the best result in test for each metric and dataset

Table 10 Complete experimental results for our proposed approach (FARC-HD-MOEA (MfM + FAR)) versus the state-of-the-art (FARC-HD, FARC-HD-OVO, and C4.5) over the Gure-KDDCUP dataset for different metrics of performance: accuracy (Acc), mean F-measure (MFM), average accuracy (AvgAcc), attack average accuracy (AttAcc), attack detection rate (ADR), and false alarm rate (FAR)

Metric	FARC-HD-MOEA		FARC-HD		FARC-HD-OVO		C4.5	
	Tr	Tst	Tr	Tst	Tr	Tst	Tr	Tst
Acc	99.38	99.20	99.37	99.24	99.40	99.24	99.35	**99.51**
MfM	90.86	75.60	88.17	74.15	91.27	**75.98**	70.43	70.30
AvgAcc	87.76	**78.80**	84.02	73.87	88.75	78.71	68.06	67.94
AttAcc	83.75	**71.85**	78.77	65.26	85.07	71.72	57.42	57.27
ADR	82.74	78.12	81.80	78.29	83.45	78.69	73.29	**79.17**
FAR	0.2173	0.3253	0.2059	0.2916	0.2116	0.3050	0.0229	**0.0318**

Bold values correspond to the best result in test for each metric and dataset

Table 11 Comparison of number of rules (#Rules) and average number of antecedents (#Avg. Ant.) for the algorithms selected in the experimental study

Dataset	FARC-HD-MOEA		FARC-HD		FARC-HD-OVO		C4.5	
	#Rules	#Avg. Ant.	#Rules	#Avg. Ant.	#Rules	#Avg. Ant.	#Rules	#Avg. Ant.
KDDCUP'99	44	2.6590	25	2.3600	84	2.2238	150	2.1385
NSL-KDD	72	2.7777	46	2.6956	111	2.3249	168	2.5128
GURE-KDD	31	2.1935	17	2.1176	44	1.6408	41	2.8133

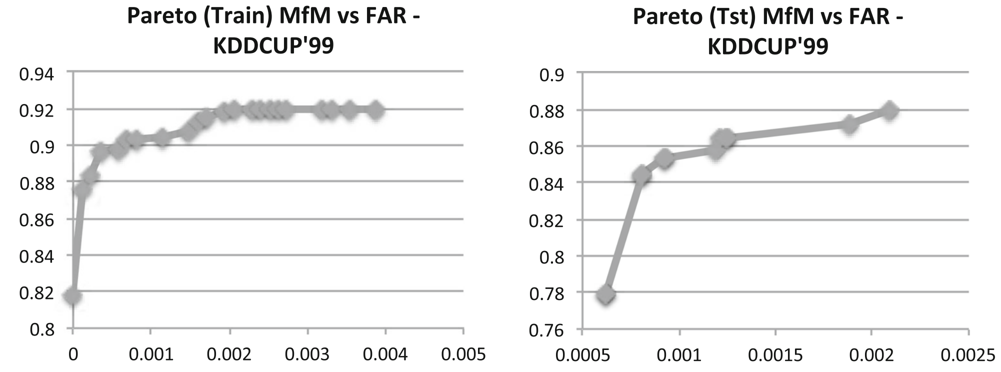
Table 12 Confusion Matrix in the test partition for the FARC-HD-MOEA approach (MfM vs. FAR, best solution for MfM) in the KDDCUP’99 dataset

	Normal	DOS	PRB	R2L	U2R	Recall
Normal	78,612	214	75	130	18	99.45
DOS	2056	47027	32	0	0	95.75
PRB	90	32	1789	4	2	93.32
R2L	91	2	5	798	3	88.77
U2R	4	0	0	4	18	69.23
Precision	97.23	99.48	94.11	85.26	43.90	–

Table 13 Confusion Matrix in the test partition for the FARC-HD-MOEA approach (MfM vs. FAR, best solution for MfM) in the NSL-KDD dataset

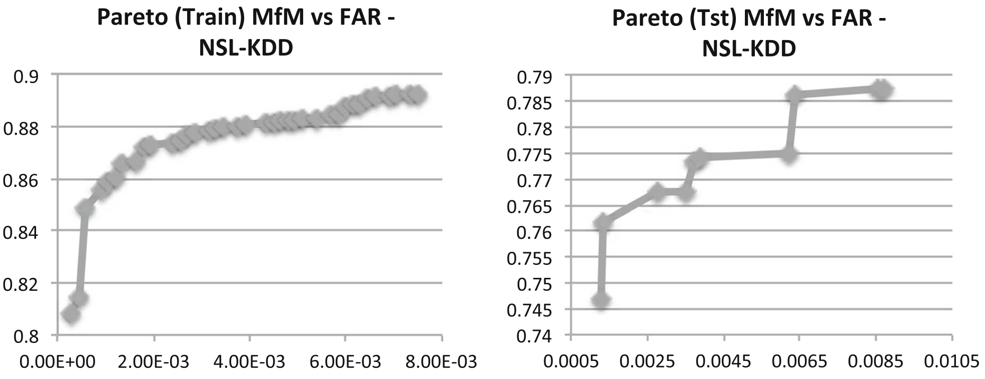
	Normal	DOS	PRB	R2L	U2R	Recall
Normal	60,331	37	143	119	39	99.44
DOS	3796	37267	0	216	0	90.28
PRB	290	0	599	0	3	67.15
R2L	622	136	0	9725	0	92.77
UR2	11	0	4	0	11	42.31
Precision	92.75	99.54	80.29	96.67	20.75	–

of performance. We must recall that the same configuration is shared by both approaches. In this sense, the initial KB obtained afterward stages 1 and 2 (refer to Sect. 2.3) will be the same for both models. This fact suggests the goodness in the design and capabilities of the proposed MOEA optimization procedure versus the standard Genetic Algorithm when dealing with IDS problems. In particular, we must stress the differences with respect to the values of the mean f-measure, average accuracy, and attack accuracy are especially remarkable, improving up to 10–15 points in some cases.



(a) (b)

Fig. 2 Pareto front obtained in the test stage with FARC-HD-MOEA approach. Objectives selected during the search were the mean F-measure (MfM) and the false alarm rate (FAR). **a** Pareto front in KDDCUP’99 dataset with FARC-HD-MOEA approach. **b** Pareto front in KDDCUP’99 dataset with FARC-HD-MOEA approach



(a) (b)

Fig. 3 Pareto front obtained in the test stage with FARC-HD-MOEA approach. Objectives selected during the search were the mean F-measure (MfM) and the false alarm rate (FAR). **a** Pareto front in KDDCUP’99 dataset with FARC-HD-MOEA approach. **b** Pareto front in NSL-KDD dataset with FARC-HD-MOEA approach

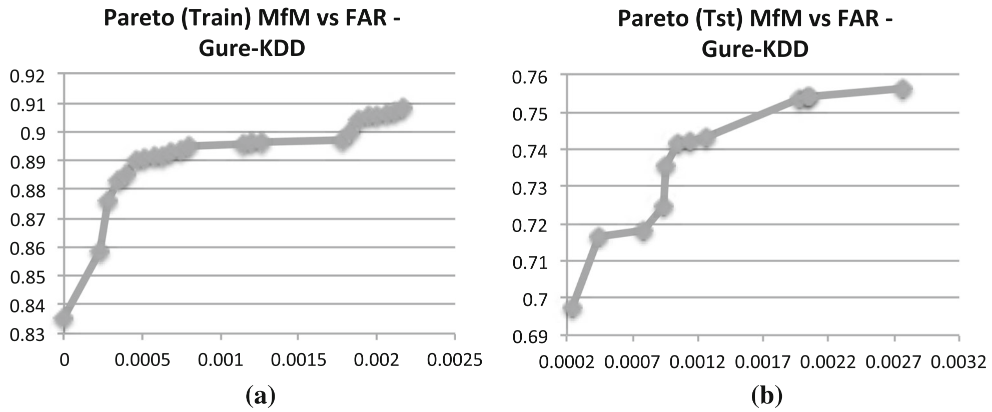


Fig. 4 Pareto front obtained in the test stage with FARC-HD-MOEA approach. Objectives selected during the search were the mean F-measure (MfM) and the false alarm rate (FAR). **a** Pareto front in KDDCUP'99 dataset with FARC-HD-MOEA approach. **b** Pareto front in Gure-KDD dataset with FARC-HD-MOEA approach

Regarding the comparison versus FARC-HD-OVO, we must stress that the performance values are more similar in this case. However, there is a clear advantage of our novel proposed technique, which relies in the use of a single classifier, instead of a whole ensemble. This fact avoids a combination among different outputs during the inference, which can degrade the system response.

When our proposed FARC-HD-MOEA is contrasted versus the C4.5 decision tree, we observe an interesting behavior. Whereas global metrics of performance such as accuracy and/or attack detection rate are usually higher for C4.5, the goodness of FARC-HD-MOEA lies in the ability of providing a good average recognition, if we focus on the mean f-measure and the average accuracy. Furthermore, the false alarm rate obtained with our approach is always below the 1%, whereas C4.5 raises this value for the NSL-KDD dataset up to the 4% (Tables 11, 12, 13).

In accordance with the whole analysis that has been carried out, we must stress that our FARC-HD-MOEA proposal is a robust choice for the IDS problem. Its main advantage in contrast to the remaining methods is its ability to achieve a good trade-off between recall (average accuracy) and precision (mean F-measure). This issue implies that our approach reaches a high average performance for all concepts/classes of the problem. Additionally, it enhances the false alarm rate from the baseline FARC-HD approach, also maintaining a very close value compared to that of the remaining algorithms. Finally, we have observed that a low number of simple (compact) linguistic rules are enough to cover the whole problem space accurately.

5.3 Complementary results: Pareto front of solutions and confusion matrices

For the sake of complementing this study, we show in Figs. 2, 3 and 4 the complete Pareto front obtained from the genetic optimization for the KDDCUP'99, NSL-KDD and GURE-KDD datasets. In all cases of study, we may observe a wide amount of non-dominated solutions from both the training and test sets, all of which are homogeneously distributed in the solution space. This issue reflects the good properties of the search procedure, as it covers a wide amount of different cases from which the expert can select the most appropriate one for a desired profile of behavior.

Table 14 Confusion Matrix in the test partition for the FARC-HD-MOEA approach (MfM vs. FAR, best solution for MfM) in the GURE-KDD dataset

	Normal	DOS	R2L	U2R	Recall
Normal	15,6873	11	443	58	99.67
DOS	46	966	0	0	95.45
R2L	696	3	1761	5	71.44
UR2	14	1	4	18	48.65
Precision	99.52	98.47	79.76	22.22	–

Finally, we include in Tables 12, 13, 14 the confusion matrices obtained in the test partition for FARC-HD-MOEA algorithm for all three IDS problems. This is done with aims at showing additional information for the experimental results, so that any interested research could reproduce and extend the current study for additional future work.

6 Concluding remarks

In the context of IDS, profiles may change depending on the users' requirements. In this sense, we must usually face confronting objectives for the working procedure of this type of system, mainly between achieving a low number of false alarms, and a good average recognition for different types of attacks.

With this aim, in this research, we have proposed the integration of a MOEA within a linguistic fuzzy association rule mining, the FARC-HD algorithm. Specifically, the genetic optimization has been focused on the last stage to carry out the rule selection and data base tuning. The goodness for the use of the MOEA is related to the simultaneous optimization of different metrics of performance in the scenario of IDS. The aim for this procedure is being able to both extending the search space and obtaining a wide amount of accurate solutions. By doing so, the final user may select the most suitable classification system for the current work context.

On the first part of our analysis, we have considered several case studies depending on the combination of metrics for the learning process. Among them, we have found out that the synergy between the mean F-measure and the false alarm rate was the most relevant for achieving good average results under the different IDS benchmark problems. Nevertheless, we must recall any other configuration could be also valuable according to the final requirements of the application.

Finally, the comparison of this approach versus the state-of-the-art, which included both the original FARC-HD classifier, FARC-HD with OVO, and the C4.5 decision tree, supported the high quality of our novel methodology. In particular, we remarked the good trade-off obtained between precision and interpretability for all cases of study, in accordance with the length of the RB and average antecedents per rule.

Acknowledgements This paper was funded by King Abdulaziz University, under Grant HiCi. The authors therefore, acknowledge technical and financial support of KAU.

Compliance with ethical standards

Conflict of interest None

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Alcala R, Alcalá-Fdez J, Herrera F (2007) A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection. IEEE Trans Fuzzy Syst 15(4):616–635

Alcalá-Fdez J, Alcalá R, Herrera F (2011) A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans Fuzzy Syst 19(5):857–872

Alok AK, Saha S, Ekbal A (2016) Multi-objective semi-supervised clustering for automatic pixel classification from remote sensing imagery. Soft Comput 20(12):4733–4751

Benferhat S, Boudjelida A, Tabia K, Drias H (2013) An intrusion detection and alert correlation approach based on revising probabilistic classifiers using expert knowledge. Appl Intell 38(4):520–540

Bostani H, Sheikhan M (2017) Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems. Soft Comput 21(9):2307–2324

Branke J, Deb K, Dierolf H, Osswald M (2004) Finding knees in multi-objective optimization. In: Yao X, Burke EK, Lozano JA, Smith J, Guervós JJM, Bullinaria JA, Rowe JE, Tiño P, Kabán A, Schwefel HP (eds) PPSN, Lecture Notes in Computer Science, vol 3242. Springer, New York, pp 722–731

Casillas J, Cordón O, del Jesús MJ, Herrera F (2005) Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set reduction. IEEE Trans Fuzzy Syst 13(1):13–29

Chebrolu S, Abraham A, Thomas JP (2005) Feature deduction and ensemble design of intrusion detection systems. Comput Secur 24(4):295–307

Chung YY, Wahid N (2012) A hybrid network intrusion detection system using simplified swarm optimization (SSO). Appl Soft Comput 12(9):3014–3022

Coello-Coello CA, Lamont G, van Veldhuizen D (2007) Evolutionary algorithms for solving multi-objective problems, genetic and evolutionary computation, 2nd edn. Springer, Berlin

Cordón O, del Jesus MJ, Herrera F (1999) A proposal on reasoning methods in fuzzy rule-based classification systems. Int J Approx Reason 20(1):21–45

Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197

Debar H, Dacier M, Wespi A (1999) Towards a taxonomy of intrusion-detection systems. Comput Netw 31(8):805–822

Eiben AE, Smith JE (2003) Introduction to evolutionary computation. Springer, Berlin

Elhag S, Fernández A, Bawakid A, Alshomrani S, Herrera F (2015) On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems. Expert Syst Appl 42(1):193–202

Fernández A, Calderón M, Barrenechea E, Bustince H, Herrera F (2010) Solving multi-class problems with linguistic fuzzy rule based classification systems based on pairwise learning and preference relations. Fuzzy Sets Syst 161(23):3064–3080

Fernández A, del Jesus MJ, Herrera F (2010) On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets. Inf Sci 180(8):1268–1291

Fernandez A, del Rio S, Lopez V, Bawakid A, del Jesus MJ, Benitez JM, Herrera F (2014) Big data with cloud computing: an insight on the computing environment, mapreduce and programming frameworks. Wiley Interdisc Rev Data Min Knowl Discov 4(5):380–409

Fernandez A, Lopez V, del Jesus MJ, Herrera F (2015) Revisiting evolutionary fuzzy systems: taxonomy, applications, new trends and challenges. Knowl Based Syst 80:109–121

Gacto M, Alcalá R, Herrera F (2011) Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. Inf Sci 181(20):4340–4360

Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F (2011) An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recogn 44(8):1761–1776

Goroohi Sardou I, Ameli MT (2016) A fuzzy-based non-dominated sorting genetic algorithm-II for joint energy and reserves market clearing. Soft Comput 20(3):1161–1177

Guo C, Zhou Y, Ping Y, Zhang Z, Liu G, Yang Y (2014) A distance sum-based hybrid method for intrusion detection. Appl Intell 40(1):178–188

Herrera F, Martínez L (2000) A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 8(6):746–752

Ishibuchi H, Yamamoto T (2005) Rule weight specification in fuzzy rule-based classification systems. IEEE Trans Fuzzy Syst 13:428–435

Ishibuchi H, Nakashima T, Nii M (2004) Classification and modeling with linguistic information granules: advanced approaches to linguistic data mining. Springer, Berlin

Kavsek B, Lavrac N (2006) Apriori-sd: Adapting association rule learning to subgroup discovery. Appl Artif Intell 20(7):543–583

Khor KC, Ting CY, Phon-Amnuaisuk S (2012) A cascaded classifier approach for improving detection rates on rare attack categories in network intrusion detection. Appl Intell 36(2):320–329

Kudlacik P, Porwik P, Wesołowski T (2016) Fuzzy approach for intrusion detection based on user's commands. Soft Comput 20(7):2705–2719

Lee W, Stolfo S (2000) A framework for constructing features and models for intrusion detection systems. ACM Trans Inf Syst Secur 3(4):227–261

Mitchell R, Chen I (2015) Behavior rule specification-based intrusion detection for safety critical medical cyber physical systems. IEEE Trans Dependable Secure Comput 12(1):16–30

Mohammadi Shanghooshabad A, Saniee Abadeh M (2016) Sifter: an approach for robust fuzzy rule set discovery. Soft Comput 20(8):3303–3319

Pan S, Morris T, Adhikari U (2015) Developing a hybrid intrusion detection system using data mining for power systems. IEEE Trans Smart Grid 6(6):3104–3113

Patcha A, Park JM (2007) An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput Netw 51(12):3448–3470

Perona I, Gurrutxaga I, Arbelaitz O, Martín JI, Muguerza J, Pérez JM (2008) Service-independent payload analysis to improve intrusion detection in network traffic. In: Proceedings of the 7th Australasian Data Mining Conference (AusDM08), pp 171–178

Quinlan J (1993) C4.5: programs for machine learning. Morgan Kauffman, San Mateo

Tavallaee M, Bagheri E, Lu W, Ghorbani A (2009) A detailed analysis of the KDD cup 99 data set. In: Second IEEE symposium on computational intelligence for security and defense applications (CISDA09), pp 53–58

Vasilomanolakis E, Karuppayah S, Muhlhauser M (2015) Taxonomy and survey of collaborative intrusion detection. ACM Comput Surv 47(4):55:1–55:33

Wu SX, Banzhaf W (2010) The use of computational intelligence in intrusion detection systems: a review. Appl Soft Comput 10(1):1–35

Zhang C, Zhang S (2002) Association rule mining, models and algorithms, Lecture Notes in Computer Science, vol 2307. Springer, Berlin